EB_2020_5_Hartley

Ebola Data Platform - Data Access Application Form

Please review the Data Access Guidelines and the Data Transfer Agreement[1] before completing this form. Note that the details of all approved applications will be made publicly available on the Ebola Data Platform website.
Please complete all sections of this form *fully* and return to ebolaDAC@iddo.org with the following attachments:

- Academic CV of the Lead Requestor (any format)
- Conflict of Interest Forms completed by the Lead Requestor and each of the Co applicants listed

SECTION A: RESEARCHER / RESEARCH TEAM INFORMATION	
Lead Requestor Details *(please attach an academic CV)*	
Title	MD, Ph.D, MPH
First name (given name)	Mary-Anne
Surname (family name)	Hartley
Gender	Female
Position at employing organisation/ institution	EPFL: École Polytechnique Fédérale de Lausanne, Switzerland • iGH: Intelligent Global Health **Project lead** Unisanté CHUV Hospital • Tropical medicine department: **Resident clinician** • Digital Global Health group: **Senior scientist**
ORCID ID https://orcid.org/	0000-0002-8826-3870
Email	[redacted]
Telephone/Skype/WhatsApp	[redacted]
Employing Organisation/Institution *Institution with a remit including health, research or academic pursuit, and with legal status which includes the scope to sign the **Data Transfer Agreement***[1]	
Institution name	**EPFL: École Polytechnique Fédérale de Lausanne** • Ranked 1st in Europe, 6th in the world (QS) • The IDDO also has the possibility of requesting that the data is hosted at my other affiliation: a purely medical institution: Unisanté/CHUV which is the leading academic hospital in Switzerland and ranked 9th in the world (Newsweek, independent international annual review)
Address	MLO, EPFL, Lausanne , Switzerland 1015
Department (if applicable)	MLO: Machine Learning and Optimization Laboratory (intelligent global health project)
Please acknowledge that your institution agrees to execute the Data Transfer Agreement	YES

[1] The ***Data Transfer Agreement*** *is a contract between the University of Oxford (on behalf of IDDO) and the recipient institution that governs the legal obligations and restrictions, as well as compliance with applicable laws and regulations, related to the* ***transfer*** *of such* ***data*** *between the parties. The named Institution will be required to sign the data transfer agreement before the release of any data by IDDO.*

Co-applicants (ALL individuals accessing the data must be listed. Any additions must be notified to the Ebola DAC) *Add rows as necessary.* *Please attach copies of the Conflict of Interest Form, completed by each of the individuals above.*	
1. Name	Andres Colubri
1. Title	MFA Ph.D
1. Organisation/Institution	Broad Institute of MIT and Harvard
2. Name	Martin Jaggi
2. Title	Professor
2. Organisation/Institution	EPFL: École Polytechnique Fédérale de Lausanne • Head of Machine Learning and Optimization laboratory (MLO) • iGH: Intelligent Global Health group co-lead
3. Name	Felix Hans Michel Grimberg
3. Title	M.Sc. candidate
3. Organisation/Institution	EPFL: École Polytechnique Fédérale de Lausanne

SECTION B: RESEARCH PLAN

Title of Proposed Research	Collaborative privacy: Derivation and validation of robust, and personalised clinical insights for Ebola Virus Disease—A decentralised machine learning framework incentivising real time collaboration through securing data privacy.
Is this a re-submission of a previous application that has been reviewed by the Ebola DAC? If so, please provide the surname of the Lead Requestor and submission date of the previous application.	No

Summary of Research in Lay Language *(suggested ~ 100 words)*

Our research aims to develop a platform for easily and robustly deriving diverse clinical insights on Ebola Virus Disease (EVD) from fragmented datasets distributed across several geographies and users whilst maintaining patient privacy and local ownership. We will evaluate new decentralized machine learning (ML) methods to crowdsource models personalised to the user's context. These methods will enable clinical researchers to generate reliable models using patient data that is distributed across sites, but without the need for data owners to share any original data or provide their patient records to a centralized repository. This will strengthen patient privacy and incentivise collaboration and interoperability between competing research parties, which is especially relevant during health emergencies and in rural settings without reliable centralized medical infrastructure.

Scientific Summary of Research *(suggested maximum 300 words)*

In the wake of the recent epidemics of emerging infections such as Ebola, which disproportionally impact low-resource settings, there is a clear need for the availability of rapidly updatable predictive models that adapt to changing environments. For EVD, such tools can have important applications in patient triage, stratification, and management. However, this requires real-time data sharing and open collaboration. During the 2014-16 EVD epidemic, the challenge of ensuring privacy and interoperability, as well as non-collaborative attitudes amongst competing academic parties and issues around gaining ethical clearance for sharing original data, created many barriers and delays to curating this invaluable and unique interoperable central repository (which is an—unfortunately—rare success of open collaboration).

Recent work from the applicants and collaborators [1-5] have shown that even with limited access to data, diagnostic and prognostic models can be highly predictive of EVD outcomes and implemented as easy-to-use tools offering clinicians a valuable adjunct for the purpose of cohorting patients to avoid nosocomial infections and to stratify clinical severity for better allocation of limited resources. Furthermore, we have evidence [4] that these models can be externally validated against expert observational wellness assessments, and also be used to provide informed access to recommended evidence-based guidelines. However, all of these published models were derived retrospectively, after lengthy processes of manual hypothesis-based testing on small heterogenous datasets and did thus not have immediate application during the outbreak and today, have limited potential to generalise to other contexts.

We thus deeply appreciate the work of the IDDO and aim to use the EDP to address the core issues of why this repository was created in the first place: following on in its spirit by providing a framework and platform to facilitate such efforts in the future and incentivise rapid interoperable collaboration and intelligent cooperative data collection.

The study will create an open-source collaborative data analysis platform that allows users to crowdsource diverse clinical insights from rapidly accumulating data across fragmented datasets, but whilst preserving patient privacy to better incentivise sharing. Although this platform would assume and incentivise interoperability of variables, decentralised learning does not require a central repository nor does it result in the exchange of any original data. We will use the EDP to offer evidence on the applicability and performance of these new decentralized ML methods to create robust real time data analysis during an outbreak.

The various models created in this study via the platform will not only seek to address key clinical insights on natural history/pathophysiology of EVD, but it will specifically seek to build evidence-based trust in elements of automated analysis possible with the latest techniques in ML. This will be done by cross-validating results of ML models to traditional statistics approaches used currently.

Thus, we aim to better democratize data analytics by providing an accessible pathway to visualising and interpreting data: recognizing public health data as a public good and delivering a tool which could be used to achieve this in the future.

Summary of Research Objectives *(suggested maximum 200 words)*

1) to create an **easy-to use, open-source collaborative data analysis platform** that allows users to **crowdsource diverse clinical insights** from rapidly accumulating data across fragmented datasets using decentralised machine learning
2) to design, test, and evaluate this framework using the EDP, by simulating **prompt, incentivised collaboration** between diverse sites during an Ebola outbreak
3) to quantify the **impact of missing data** on decentralized ML models, and develop models that **crowdsource robust imputations** and identify systematic gaps early in the response to **better inform representative data capture**.
4) to develop and **compare the predictive performance of the decentralized models with the "gold-standard"** determined by a global model centrally trained on the entire EDP dataset or **using traditional statistic methods**.
   Specifically, the models developed in this last objective, will aim to answer several key clinical questions related to
   - understand the factors predictive of EVD infection and mortality risk for improved risk stratification to continue and expand on the previous work of applicants for EVD diagnosis and prognosis prediction [1-5] as well as adding and comparing various ML approaches
   - predictive tools to estimate biomarker evolution for probabilistic resource allocation

- improved data quality through anomaly detection and intelligent imputation
- intelligent data collection to generate data-driven questionnaires that improve data quality and collaborative predictive power
- Outbreak detection (using unsupervised machine learning), to differentiate malaria (and if possible other febrile diseases) from Ebola as was the aim of [2]

Primary and Secondary Outcome Measures *(suggested maximum 200 words)*

Primary outcome measures:

(1) An easy to use **open-source platform for private decentralised data analysis**
- publication of web-based platform
- Functionalities of decentralised learning and interpretable data visualisations

(2) **Improved performance** measures of **predictive algorithms** for EVD diagnosis, prognostication, biomarker evolution estimation compared to existing predictions
- Accuracy, sensitivity, specificity, AUC, F1 score, calibration, etc.

(3) **Imputation with certainty scores**
- Generating **a list of high-priority questions** to add to questionnaires in order to further improve imputation accuracy through data-driven data collection .

(4) An **evaluation of the predictive performance of decentralized models** across various indices vs traditional centralised approaches.

Secondary outcome measures:

(5) **Robustness** of centralized and decentralized models to **missing data**

(6) **Identify and quantify challenges associated to the use of decentralized ML methods** in rural settings
- Communication efficiency, scalability, elasticity as well as resilience to unreliable and heterogeneous participating devices (for data collection and computation)

(7) **Ecological reporting on the carbon footprint** of each ML model
- kWh (kilowatt hours)

Proposed Methodology and Statistical Analysis Plan *(suggested maximum 400 words)*

Analysis Platform: We have already developed a functional prototype for a web-based application with an easy to use interface able to rapidly generate interpretable and comparable ML predictive models on centralised datasets. This platform will be adapted to host decentralised learning described below.

Intelligent data collection applications: We have already developed a working prototype of two flexible interoperable data capture mobile apps: 1) **EbolaCare Guidelines:** (publicly available here: https://github.com/broadinstitute/ebola-care-guidelines) for use at the point-of-care as a clinical decision support tool, providing tailored access to treatment recommendations [4] and 2) **CLINICapp:** (publicly available soon) for use in structured collaborative research data collection. These apps will be expanded to integrate into the interoperable decentralised architecture to facilitate rapid collaboration in the field, ranging from the calculation of summary statistics from patient cohorts across sites, to more sophisticated model building and updatable treatment recommendations.

Decentralized ML : Technically, we will apply algorithms developed at the MLO lab at EPFL led by applicant Prof. Jaggi. These algorithms are based on Stochastic Gradient Descent (SGD). Compared to classic gradient descent, SGD performs updates to the model parameters more efficiently and more frequently, based on just a single training datapoint per step or a small mini-batch of training points. A critical challenge in realizing a scalable method to use across heterogenous sites, especially in rural settings, is to develop efficient, privacy-preserving methods for communicating and coordinating information between distributed nodes. Model compression becomes a crucial ingredient in these settings in order to satisfy device capacity constraints e.g. on mobile user

devices. Recent work from Jaggi's lab has shown the first decentralized SGD algorithm that converges with arbitrary compression for convex cost functions, with applications to many machine learning models [6], such as those that are typically used in clinical prediction (e.g. linear and logistic regression). In recent follow-up work [7], it was further shown that the algorithm converges under arbitrary high compression ratio on general non-convex functions (such as for deep learning).

Centralised traditional statistical approach: We will apply the analysis methodologies used in our previous publications. In particular, we will carry out internal validation of models using bootstrap resampling, and performance will be characterized by the area under the curve (AUC), McFadden's pseudo-R2, Brier score, accuracy, sensitivity and specificity. Confidence intervals (CI) of all the performance estimates will be calculated using Fisher's transformation, while Odds ratios (OR) from logistic regression models will be converted to risk ratios (RRs) according to Zhang and Yu method [8].

Handling of missing data: We will test the accuracy of imputation models for each feature, by generating "fake" missing data and using the real values as ground truths. A range of imputation techniques will be used to predict missing data and then we will assign appropriate techniques to each feature "type" (based on variance, range, % missingness and similarity to collaborating datasets from which the feature is being imputed). Features with poorly performing imputations will be suggested to be prioritised for collection (if they are useful in predicting clinical outcomes).

Data Privacy methodology: The kind of data represented by the EDP dataset is extremely sensitive, as it comprises patient records affected by a highly stigmatizing disease, for which informed consent was exceptionally waived in many cases. Our decentralized algorithms will use secure multi-party computation (MPC), which is a is a family of cryptographic techniques allowing several parties to compute a function of their private inputs in a way that only reveals the intended output to each of the parties [9]. Recent advancements have led to successful deployments in industry and are receiving increasing interest for ML use-cases [10]. This project will rely on MPC and secure aggregation to protect data of all EVD patients, while making the resulting models available openly, facilitating inference and quality control of models.

Ethics *(suggested maximum 300 words)*
Provide details of any ethical considerations relating to the research proposal.
Additionally, list any approvals required by your institution to undertake this work, list reference numbers of any approved proposals, or explain why no approvals are required.

We anticipate the EPFL and Harvard Committees on the Use of Human Subjects to provide ethical approval for this study and exemption from informed consent, as our past studies similar to this one have obtained such approval and exemption.
The research questions addressed in this project will not go outside the EDP research agenda and rather aim to fulfil it.

Publication and Dissemination Plan *(suggested maximum 300 words)*
Provide details of plans for authorship/acknowledgement of data contributors.
Provide details of timelines for publication and dissemination of research findings.

The core aim of this study is to **enhance open collaboration** and **build capacity and concrete tools to better democratise data analysis.** It is obvious that this policy of inclusivity would continue to authorship and acknowledgement in publications where data contributors fulfilling the ICMJE standards of co-authorship will be contacted and included as per these objective and transparent international guidelines.

Our data analysis code will be made open source to better invite collaboration and transparent validation. Access to the EDP data to which we will apply said code will obviously be controlled by the IDDO-EPFL DTA.

The platform for analysis will also be made open source and open access, allowing users to plug in their own datasets for analysis.

We aim is to have initial publishable insights on diagnostic and prognostic models within 6months of receiving the data (preparing publications and passing them through peer review should take another 6m to 1 year, as per usual).
Once the β version of our analysis platform is complete (6m-1year) the insights will be much faster to generate and validate and several papers will follow regarding imputation, data quality and intelligent questionnaire building etc.

Thus, we are working on a 2-3 year timeline for the completion of most milestones, however, this work is part of a bigger platform for outbreak detection (with large scale roll-out in several African countries). We have committed to maintaining the platform over the next 5 years and plan to secure funding for maintaining it thereafter.

Addressing Knowledge Gaps (*suggested maximum 300 words)*
Provide details of how this research will address knowledge gaps of importance to those affected by or at risk of emerging and poverty-related diseases.

Speed is critical in understanding the nature and spread of viral infections like Ebola; a rapid response can save thousands of lives. Paramount to timely outbreak response is the ability to share information in real-time. From the moment the first cases of an outbreak are reported, scientists need to look at the available data to identify clusters of cases, detect the pathogen, characterize its severity and clinical manifestations, and understand the patterns of transmission.

By applying machine learning on this data, scientists can further create models to predict how the virus is going to affect new patients, and find the best treatment and containment strategies before the outbreak becomes a devastating epidemic. However, most data needed by scientists to stop an outbreak can be traced back to individual patients, making protection of patient privacy a fundamental need in surveillance and response efforts. Patient health data is always sensitive, but even more so for outbreak survivors who can face stigma in their communities after recovery. Patient privacy is one key reason why data owners—hospitals, humanitarian organizations, ministries of public health, among others—are reticent to share their data. Currently, when even possible, accessing patient records involves a lengthy process where data owners and research institutions establish data sharing agreements and protocols to ensure ethical use of the patient's information. Even in the best cases, this can be a time-consuming process. Most research clinicians today still use simple anonymization (i.e. removing names, addresses and identity numbers), which is becoming increasingly outdated, as it puts patient privacy at risk of re-identification attacks that integrate multiple data sources. Furthermore, anonymization offers no protection for the researchers' intellectual property. This is critical in the context of competitive research, and leads to data hoarding, incomplete sharing and significant delays in dataset publication; all of which go against the FAIR principles of open research [11]. Our decentralized ML tools will provide a framework that allows for multi-party collaboration by sharing models and by guaranteeing the privacy of each user in the network, thus enabling rapid collaboration in the event of an outbreak.

Thus, we aim to develop a tool that is able to address knowledge gaps in real time whilst securing data privacy and improving data integrity. In developing this platform we will also directly address several specific knowledge gaps within the dataset i.e. features predictive of death, infection, differential diagnoses and biomarker evolution. We aim to integrate these ML models into a tool that will allow them to evolve with data captures in the next outbreak

Equity and Capacity Building *(suggested maximum 300 words)*
Provide details of how this research will support health equity and/or capacity building in endemic regions affected by or at risk of emerging and poverty-related diseases.
Please refer to the Ebola Data Platform Approaches to Capacity Building for guidance.

This project is closely in line with the IDDO objectives for capacity building
In particular

IDDO objective	Our planned activities
Create training and mentorship opportunities to strengthen clinical research skills among researchers in countries of data origin	We currently have underway a funded academic exchange program in Tanzania and Rwanda (to be expanded to other African countries at risk of EVD). It is called MLx (machine learning academic exchange programme) and serves to promote ML capacity in Africa (See annexed advert for this effort in Tanzania).
Provide training in data management, statistics, study design, research ethics, community engagement, manuscript writing and/or grant applications	Dr Colubri, applicant on this request has developed a participatory epidemiology tool simulating an outbreak. We aim to secure funding to adapt this tool for African contexts and implement it in collaboration with local schools in affected countries to create community engagement that can be leveraged in future outbreaks: http://andrescolubri.net/projects/o2
Enable access to fellowships and training schemes	The MLx programme specifically gives bursaries for online Machine learning courses with practical training and an EPFL-diploma.
Curate a list of recommended online courses and resources applicable to EVD-affected countries	Our tools will be available online and specifically will include tailored access to treatment recommendation guidelines.
Deliver in-person training	Students excelling in the EPFL diploma course, will be invited to participate in a 3-month internship at EPFL to contribute to the tools we are developing. We will further host a hackathon to build tools for health care in resource limited settings: open to the Tanzanian public.
Answer research questions aligned with the priorities of countries affected by or at risk of Ebola outbreaks	We are developing a platform that will not only generate models relevant to the EDP data, but also allow these models to evolve with new data for future outbreaks.
Reinforce the integration of research in public health policy and response	Our platform aims to make ML data analysis more accessible through interpretability, giving users insights into how individual inputs affect the outcome of their trained model.
Build, support and/or sustain frameworks to bridge public health and research response	Our point of care tool is designed to integrate official treatment recommendations at the point of care, thus bridging public health policy and field actors in treatment and research.
Improve the quality of public health data	Our focus on imputation and interoperable data entry is a clear aim to this end. Both the mobile app and decentralized ML framework will be released as open-source, properly documented, promoted, and beta tested, to ensure accessibility and usage. Thus, more health workers and organizations active in regions affected by EVD outbreaks will be incentivized to develop diagnostic and prognostic machine learning models. The mobile app will give its users recommendations for data collection to improve interoperability and better leverage collaborative ML models.
Develop evidence-based data capture forms	See above.

Funding *(suggested maximum 100 words)*
Provide details of how this research will be funded/resourced.

Dr Hartley and Prof Jaggi's work is funded by the Botnar foundation, which aims to improve diagnostics in resource limited settings. Salaries are funded and secured on long term contracts covering the scope of this project.
The group of the applicant (EPFL-MLO) has over 20 members already employed and capable of undertaking these analyses. Notably, the work includes several persons of African origin in key positions to best ensure agency and insights from resource-limited settings.

Dr Colubri is a key collaborator and has developed the EbolaCare guidelines app as well as provided funding and support for the CLINICapp tool. His salary is funded independently in a permanent position.

While **no further funding is required to undertake the analyses in this study**, we are actively seeking funding to expand its scope.

Scientific Review *(suggested maximum 200 words)*
Provide details of how the details of the project outlined above have been scientifically reviewed. This could be by your institution, a funder/donor or review committee.

This work is part of a larger project that has been funded with an 8 million CHF grant from the Botnar foundation, to improve health for children in resource limited settings. The protocol was rigorously reviewed by several ethics' committees and a scientific team of several experienced clinicians (from South Africa, Germany, Switzerland and Tanzania) as well as accomplished data scientists (including Prof Martin Jaggi: applicant in this request and group leader at EPFL, an institution ranked 6th in the world). Our team at EPFL and MIT is comprised of a diverse mix of clinicians, data scientists and computational scientist who have all approved this application.

SECTION C: DATA

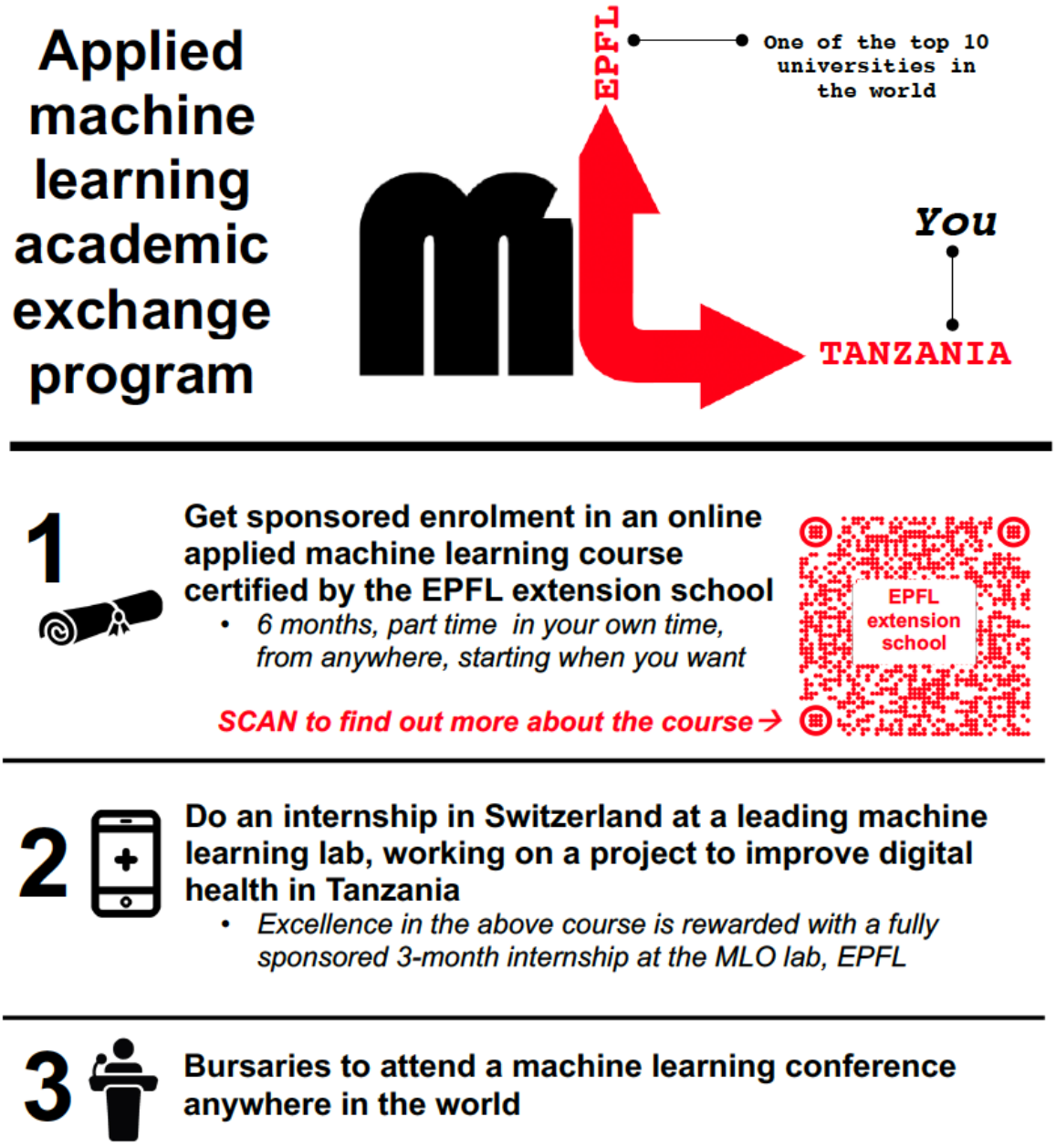
Data Variables
*Provide a list of the **data variables and data sources** required to achieve the research objectives.*
Note: Data sources can be listed as populations (e.g. all EVD-positive pregnant women, or all children under 16 years of age from Liberia) or as datasets from a source listed on the Accessing Data web page (these should be named by 'Contributing organisation, Country, City' as listed in the table). Get in touch if you have any questions about this ebolaDAC@iddo.org

To perform this study, we require all patient records and all available variables.
Rationale:

- The ML techniques we are going to use perform best on large datasets.
- We have the specific aim of exploring the value of decentralised models on heterogenous data, and thus each dataset provides critical information on the effect of heterogeneity and the value of model personalisation.
- Using unsupervised learning, we wish to explore hypothesis-free model building (thus without manual preselection of variables).
  This study explores the importance of missing data and the value of imputation. It is critically important that we get the entire dataset with all variables, to evaluate the performance of proposed imputation techniques and the resilience of collaborative ML methods to missing data.

Annexe

MLx: The machine learning academic exchange programme to build ML capacity in Africa centred around global health issues.



REFERENCES

1. Levine AC, Shetty PP, Burbach R, Cheemalapati S, Glavis-Bloom J, Wiskel T, et al. *Derivation and Internal Validation of the Ebola Prediction Score for Risk Stratification of Patients With Suspected Ebola Virus Disease.* Ann Emerg Med, 2015. **66**(3): p. 285-293 e1.
2. Hartley MA, Young A, Tran AM, Okoni-Williams HH, Suma M, Mancuso B, et al. *Predicting Ebola infection: A malaria-sensitive triage score for Ebola virus disease.* PLoS Negl Trop Dis, 2017. **11**(2): p. e0005356.
3. Hartley MA, Young A, Tran AM, Okoni-Williams HH, Suma M, Mancuso B, et al. *Predicting Ebola Severity: A Clinical Prioritization Score for Ebola Virus Disease.* PLoS Negl Trop Dis, 2017. **11**(2): p. e0005265.
4. Colubri A, Hartley MA, Siakor M, Wolfman V, Felix A, Sesay T, et al. *Machine-learning Prognostic Models from the 2014-16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications.* EClinicalMedicine, 2019. **11**: p. 54-64.
5. Colubri A, Silver T, Fradet T, Retzepi K, Fry B and Sabeti P. *Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients.* PLoS Negl Trop Dis, 2016. **10**(3): p. e0004549.
6. Koloskova A, Stich SU and Jaggi M *Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication*. arXiv e-prints, 2019. arXiv:1902.00340.
7. Koloskova A, Lin T, Stich SU and Jaggi M *Decentralized Deep Learning with Arbitrary Communication Compression*. arXiv e-prints, 2019. arXiv:1907.09356.
8. Zhang J and Yu KF. *What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes.* JAMA, 1998. **280**(19): p. 1690-1.
9. Yao AC. *How to generate and exchange secrets*. in *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*. 1986.
10. Chen V, Pastro V and Raykova M *Secure Computation for Machine Learning With SPDZ*. arXiv e-prints, 2019. arXiv:1901.00329.
11. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.