COVID-19 Data Platform - Data Access Application Form

Please review the Data Access Guidelines and the Data Transfer Agreement before completing this form. A complete application should address all of the Review Considerations outlined in the Data Access Guidelines. Note that the details of all approved applications will be made publicly available on the COVID-19 Data Platform website. Complete all sections of this form fully and return to covid19@iddo.org.

SECTION A: RESEARCHER / RESEARCH TEAM INFORMATION	
Lead Applicant Details	
Title	
First name (given name)	Smith
Surname (family name)	Heavner
Gender	Non-binary
Position at employing organisation/ institution	Scientific Director
ORCID ID (https://orcid.org) or URL to academic profile	https://orcid.org/0000-0003-0912-0407 *(if no ORCID or URL, please attach a short academic CV)*
Email	*sheavner@c-path.org*
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.***	
Institution name	CURE Drug Repurposing Collaboratory, Critical Path Institute
City, Country	*Tucson, AZ, USA*
Does your institution agree to execute the Data Transfer Agreement? (if your application is approved)	**YES** (delete as appropriate)
Co-applicants *ALL individuals accessing the data must be listed. Any additions must be notified to the COVID-19 Data Access Committee. Add rows as necessary.*	
1. Name	*Serghei Gorobet*
1. Position / Role in analysis	*Data Scientist*
1. Organisation/Institution	*NCATS/NIH*
2. Name	*Ewy Mathe*
2. Position / Role in analysis	*Data Scientist*
2. Organisation/Institution	*NCATS/NIH*
3. Name	*Serghei Gorobet*
3. Position / Role in analysis	*Data Scientist*
3. Organisation/Institution	*NCATS/NIH*
4. Name	*Mike Pauley*
4. Position / Role in analysis	*Statistician / Data manager*
4. Organisation/Institution	*CDRC*
5. Name	*Kerry Howard*
5. Position / Role in analysis	*Post-doc fellow / statistician*
5. Organisation/Institution	*CDRC / Clemson University*
Conflicts of Interest *List details of any existing or perceived conflicts of interest (financial or non-financial) that exist relating to the use of the requested data by the data requestor and/or co-applicants (see ICMJE.org for the definition of conflicts of interest)*	
None to disclose	

SECTION B: RESEARCH PLAN

Title of Proposed Research	Formative Evaluation of a Minimally Viable COVID-19 Dataset
Is this a re-submission of a previous application to the COVID-19 DAC? If yes, provide the submission date of the previous application.	No

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

We are working to understand the minimum data required to study key outcomes in COVID-19 including all cause, in hospital mortality, length of stay, and need for dialysis. This methodological evaluation will use data from the COVID-19 data platform to test analytic models and determine what clinical variables and data volumes are needed to predict outcomes.

Summary of Research Objectives and Scientific Value *(suggested maximum 400 words)*

The CURE-ID database is an internet-based repository that lets the clinical community report novel uses of existing drugs for difficult-to-treat infectious diseases. The platform enables the crowdsourcing of medical information from health care providers to facilitate the development of new treatments. The repository captures clinical outcomes when drugs are used for new conditions, in new populations, in new doses or in new combinations.

CURE ID is a collaboration between the FDA and the National Center for Advancing Translational Sciences (NCATS), part of the National Institutes of Health (NIH). FDA and NIH are also collaborating with the World Health Organization and the Infectious Disease Society of America to assess the global utility of the CURE-ID. The utility of the database is being explored in the context of novel treatments for COVID-19. To evaluate the design of the CURE-ID database for COVID-19, we will simulate a CURE-ID dataset using ISARIC data and model analyses to determine if the feasibility of assessing outcomes using the variables included.

This project aims to assess correlation of all covariates with key outcomes to inform the continued development of the CURE-ID dataset. By using the IDDO dataset to explore associations between data variables we will inform the design of the CURE-ID dataset and calculate the sample sizes needed to address the CURE-ID scientific objectives. The heterogeneity, quality and missingness of key data variables such as vital signs, demographics, signs and symptoms at presentation, and treatments, will inform our statistical design. The relationships between these variables and outcmes of interest such as in-hospital mortality, length of stay and need for dialysis will guide what variables are included in the CURE-ID dataset.

Ultimately, we will leverage the IDDO data to deliver a data-driven approach to the design of the CURE-ID dataset. This will ensuring the optimal methodological and statistical design through understanding of what can be expected from a global COVID-19 dataset.

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

This is not a study of clinical outcomes. This evaluation seeks to apply a data driven approach to defining a small dataset as part of the deliverables of the CURE-ID program. Our goal is to understand statistical relationships between the variables and outcomes requested. The primary clinical outcome we will examine is all cause mortality within 28 days of hospitalization, and the secondary outcome is hospital length of stay.

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

We will be using logistic regression with least absolute shrinkage and selection operator (LASSO) along with k-fold cross validation in order to select the variables for the publicly available dataset. The LASSO was selected over other variable selection methods (e.g., Ridge Regression, Artificial Neural Network) due to the higher performance from easy implementation. LASSO will be able to address overfitting and optimism bias result from standard methods.
The whole cohort will be randomly divided into training and testing sets at a 9:1 ratio. The application of LASSO regression to the training set will be used to identify the most relevant variables associated with the selected outcome measure. And the test set will be used to evaluate the performance of the selected model. This process will be conducted for each outcome of interest and the selected variables from these models will be used in developing the publicly available dataset.

Ethics *(suggested maximum 300 words)*
Provide details of any approvals required by your institution to undertake this work, list reference numbers of any approvals, or provide clear evidence as to why no approvals are required (e.g. an extract of relevant the policy from your institutional ethics review board).
In addition, please give examples of which ethics guidelines you will be following with respect to delivering this project (e.g. you may wish to refer to general guidance such as the CIOMS/WHO International Ethical Guidelines for Health-related Research Involving Humans, domain-specific guidance such as the FATML Principles for Accountable Algorithms, or guidance specific to public health emergencies such as the Nuffield Council on Bioethics Research in Global Health Emergencies: Ethical Issues (as applicable).

We are aware that data on the COVID-19 data platform undergo a number of anonymisation steps, including the removal of the 18 Safe Harbour variables. This qualifies analysis of these data as anonymised according to the NIH, which is exempt from Ethics Committee review. Regionally tailored approaches to research ethics review and consent were applied in each setting of data collection. Our proposed use of data focuses on better understanding COVID-19 treatment and is in line with the original purpose of data collection.
This research will be implemented in accordance with the CIOMS/WHO International Ethical Guidelines. In accordance with these guidelines, use of the COVID-19 platform data will strengthen the inclusion of data from vulnerable and under-represented groups, ensuring representation of these populations in the development of our dataset. The CURE-ID dataset is an international effort and will benefit from the international nature of the available data.

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.

Findings will initially be shared with the project leadership team. The final outputs will include an improved design of the CURE-ID dataset, which will be widely and openly disseminated within the 2 years of the project.
Any publications arising from the outputs of this evaluation will include acknowledgement of all data contributors and/or the DOIs associated with their datasets to accredit their contribution.
The research team will not share or make public any portion of the data provided by IDDO.

Research Priorities Addressed (*suggested maximum 300 words)*
Provide details of how this research aligns with nationally or internationally set research priorities.

Improving COVID-19 treatment is a key priority of every relevant research agenda. This work will contribute to that cause by developing the CURE-ID dataset design to capture

variables that can robustly assess signals of treatment efficacy or harm to inform new clinical trials.

Collaboration and Knowledge Sharing *(suggested maximum 300 words)*
Provide details of how this research will collaborate, support and/or share knowledge with appropriate partners. The platform is particularly interested in research that builds capacity in low-resource settings.

This project is an initiation of a collaboration with ISARIC and will support ISARIC's development of a core dataset. Outputs will be integrated into the CURE-ID dataset that has an international reach via the WHO and As the C-Path project focuses on These efforts will include ISARIC and C-Path partners from low- and middle-income collaboration with international partners.

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

Personnel and resources required to deliver this analysis are provided in kind by C-Path.

Scientific Review *(suggested maximum 200 words)*
If the project has been scientifically reviewed, please provide details. This could be by your institution, a funder/donor or review committee.

This project was reviewed by statisticians within the CURE Drug Repurposing Collaboratory. See co-applicants for names. This project will be conducted in collaboration with ISARIC and will leverage the input from the ISARIC Head of Data and Clinical/Analytic Team where required.

SECTION C: DATA

Data Variables
*Provide a list of the **data variables** required to achieve the research objectives.*
Note: Please go to www.iddo.cognitive.city to explore the interactive COVID-19 data inventory and to identify the variables, populations and data volumes required for your analysis. You can select the data variables from this inventory and copy it to this section.

We request access to all data variables included in the C-Path database as below.

We recognize that some variables may be sensitive (e.g., race, ethnicity, and country) and that there is a potential for racial bias in our findings. This study will only examine these variables' correlation with outcomes rather than drawing any conclusions about the effects of race and ethnicity on clinical outcomes in COVID-19.

Variable	Category	Overarching Variable
Priority Level 0 - Minimal Viable Product		
Year	Reference	
Date of Hospitalization	Reference	
CURE ID Assigned identification	Reference	
Subject ID	Reference	
Age	Demographics	
Sex	Demographics	
Ethnicity	Demographics	
Race	Demographics	
Country	Demographics	

HIV	Baseline Condition	Co-morbidities
Asthma	Baseline Condition	Co-morbidities
COPD	Baseline Condition	Co-morbidities
Other Chronic Lung Disease	Baseline Condition	Co-morbidities
Diabetes Mellitus	Baseline Condition	Co-morbidities
Hypertension	Baseline Condition	Co-morbidities
Cardiovascular Disease	Baseline Condition	Co-morbidities
BMI	Baseline Condition	Co-morbidities
Weight	Baseline Condition	Co-morbidities
Nucleic Acid Test	Diagnosis	Laboratory Assessment
COVID-19 Clinical Codes -List these out**	Diagnosis	Clinical Assessment
Pulse Oximetry	Clinical Presentation	Vital Signs
Respiratory Rate	Clinical Presentation	Vital Signs
Systolic Blood Pressure	Clinical Presentation	Vital Signs
Diastolic Blood Pressure	Clinical Presentation	Vital Signs
Mean Arterial Pressure	Clinical Presentation	Vital Signs
SpO2/FiO2	Clinical Presentation	Vital Signs
PaO2/FiO2	Clinical Presentation	Vital Signs
Heart Rate	Clinical Presentation	Vital Signs
Drug	Interventions and Treatment	Medication Administration
Oxygen Support Device	Interventions and Treatment	Oxygen
Dialysis (as outcome)	Interventions and Treatment	Life Support Functions
Hospitalization	Outcomes	
Death in Hospital	Outcomes	
Priority Level 1 - For Pilot (Needed)		
Gender	Demographics	

Chronic Renal Disease	Baseline Condition	Co-morbidities
Chronic Liver Disease	Baseline Condition	Co-morbidities
Immunocompromised Condition – Note: would have to map to specific ICD-10 codes	Baseline Condition	Co-morbidities
Stroke	Baseline Condition	Co-morbidities
Dementia	Baseline Condition	Co-morbidities
Smoker	Baseline Condition	Co-morbidities
Height	Baseline Condition	
Pregnancy status	Baseline Condition	Pregnancy
Bulk Flow of O2	Clinical Presentation	Vital Signs
Temperature	Clinical Presentation	Vital Signs
Date of Laboratory Diagnosis of Infection (Relative Day)	Laboratory	
Hemoglobin	Laboratory	Laboratory Assessment
WBC	Laboratory	Laboratory Assessment
Total Lymphocyte Count	Laboratory	Laboratory Assessment
Neutrophil count	Laboratory	Laboratory Assessment
Platelet count	Laboratory	Laboratory Assessment
Eosinophil count	Laboratory	Laboratory Assessment
APTT	Laboratory	Laboratory Assessment
PT	Laboratory	Laboratory Assessment
INR	Laboratory	Laboratory Assessment
Fibrinogen	Laboratory	Laboratory Assessment
ALT/SGPT	Laboratory	Laboratory Assessment
Total Bilirubin	Laboratory	Laboratory Assessment
AST/SGOT	Laboratory	Laboratory Assessment
Glucose	Laboratory	Laboratory Assessment

Urea (BUN)	Laboratory	Laboratory Assessment
Lactate	Laboratory	Laboratory Assessment
Serum creatinine	Laboratory	Laboratory Assessment
Sodium	Laboratory	Laboratory Assessment
Potassium	Laboratory	Laboratory Assessment
Procalcitonin	Laboratory	Laboratory Assessment
CRP	Laboratory	Laboratory Assessment
LDH	Laboratory	Laboratory Assessment
Creatinine kinase	Laboratory	Laboratory Assessment
Troponin I	Laboratory	Laboratory Assessment
D-dimer	Laboratory	Laboratory Assessment
Ferritin	Laboratory	Laboratory Assessment
IL-6	Laboratory	Laboratory Assessment
Glomerular Filtration Rate (GFR)	Laboratory	Laboratory Assessment
Vasopressor Use	Interventions and Treatment	Medication Administration
Route of Administration	Interventions and Treatment	Medication Administration
Days of Drug Administration	Interventions and Treatment	Medication Administration
History of COVID-19 Vaccination as reported by the patient	Interventions and Treatment	Vaccinations
Time to Improvement	Outcomes	
WHO Ordinal Scale Score	Outcomes	
Priority Level 2 - For Full 2 Year Grant (Nice to Have)		
Immunosuppressant drugs	Baseline Condition	Co-morbidities
Down Syndrome	Baseline Condition	Co-morbidities
Autism	Baseline Condition	Co-morbidities
Gestational age, if pregnant	Baseline Condition	Pregnancy

Population density of census tract	Social Determinants	Social Determinants
Area Deprivation Index Score	Social Determinants	Social Determinants
Insurance Status	Social Determinants	Social Determinants
SOFA score	Clinical Presentation	Vital Signs
QSOFA score	Clinical Presentation	Vital Signs
Glasgow Coma Scale (GCS)	Clinical Presentation	Vital Signs
Richmond Agitation Severity Score (RASS)	Clinical Presentation	Vital Signs
ICD Codes for other syndromes/presentation	Clinical Presentation	Clinical Syndrome and Presentation
Admitting Diagnosis	Diagnosis	
Discharge Diagnosis	Diagnosis	
Antigen Test	Diagnosis	Laboratory Assessment
Bacterial culture conducted	Microbiology	Bacterial culture
Type of Bacterial culture	Microbiology	Bacterial culture
Bacterial Organism	Microbiology	Bacterial culture
Viral assay conducted	Microbiology	Viral assay
Type of Viral Assay	Microbiology	Viral assay
Viral respiratory panel	Microbiology	Viral assay
Viral respiratory panel result	Microbiology	Viral assay
Adenovirus	Microbiology	Viral assay
Coronavirus (229E, HKU1, NL63, OC43)	Microbiology	Viral assay
Human metapneumovirus	Microbiology	Viral assay
Rhinovirus/enterovirus	Microbiology	Viral assay
Influenza A	Microbiology	Viral assay
Influenza B	Microbiology	Viral assay
Parainfluenza 1-4	Microbiology	Viral assay
Respiratory syncitial virus	Microbiology	Viral assay
PCR for Herpes viruses	Microbiology	Viral assay
Fungal culture conducted	Microbiology	Fungal culture
Type of Fungal culture	Microbiology	Fungal culture
Fungal culture growth	Microbiology	Fungal culture
Beta D-glucan	Microbiology	Fungal biomarkers
Galactomannan (serum)	Microbiology	Fungal biomarkers
Galactomannan (bronchoalveolar lavage)	Microbiology	Fungal biomarkers
Histoplasma antigen (serum)	Microbiology	Fungal biomarkers
Histoplasma antigen (urine)	Microbiology	Fungal biomarkers
Coccidioides IgM antibody	Microbiology	Fungal biomarkers

Coccidioides IgG antibody	Microbiology	Fungal biomarkers
Viral pneumonia	Complications	
Bacterial pneumonia	Complications	
Acute Respiratory Distress Syndrome	Complications	
Pneumothorax	Complications	
Pleural effusion	Complications	
Cardiac arrest	Complications	
Myocardial infarction	Complications	
Cardiac arrhythmia	Complications	
Myocarditis/Pericarditis	Complications	
Endocarditis	Complications	
Congestive heart failure	Complications	
Seizure	Complications	
Stroke / Cerebrovascular accident	Complications	
Meningitis / Encephalitis	Complications	
Bacteremia	Complications	
DIC	Complications	
Pulmonary embolism	Complications	
Deep Vein thrombosis	Complications	
Acute renal injury / Acute renal failure	Complications	
Gastrointestinal hemorrhage	Complications	
Recent Outpatient Drugs	Interventions and Treatment	Medication Administration
Dose	Interventions and Treatment	Medication Administration
Frequency	Interventions and Treatment	Medication Administration
Tracheostomy inserted	Interventions and Treatment	Other Interventions
ICU Stay	Outcomes	
Priority Level 3 - For Full 2 Year Grant (Aspirational)		
Postpartum (within 6 weeks of delivery)	Baseline Condition	Pregnancy
Pregnancy Outcome	Baseline Condition	Pregnancy
SARS-CoV-2 Strain	Diagnosis	Laboratory Assessment
Symptom Onset Date (relative to admission)	Diagnosis	Clinical Assessment
Smear	Diagnosis	Laboratory Assessment
Culture	Diagnosis	Laboratory Assessment
PCR or other nucleic acid test	Diagnosis	Laboratory Assessment

Clinical Assessment	Diagnosis	Clinical Assessment
Imaging	Diagnosis	Clinical Assessment
Pathology	Diagnosis	Laboratory Assessment
Serology	Diagnosis	Laboratory Assessment
ALK PHOS	Laboratory	Laboratory Assessment
Inhaled nitric oxide	Interventions and Treatment	Other Interventions
Surgery	Interventions and Treatment	Surgery
Death Outside of Hospital	Outcomes	
Adverse Events	Outcomes	Adverse Events
Infant stillborn	Outcomes	Pregnancy Outcome
NICU Admission	Outcomes	Pregnancy Outcome
Infant complications	Outcomes	Pregnancy Outcome
Congenital anomaly	Outcomes	Pregnancy Outcome
Neonatal outcome	Outcomes	Pregnancy Outcome
Primary cause of neonatal death	Outcomes	Pregnancy Outcome
Primary cause of neonatal death, causal organism if infection	Outcomes	Pregnancy Outcome
Follow-up Culture	Outcomes	Laboratory Outcome
Follow-up Smear	Outcomes	Laboratory Outcome