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.

SECTION A: RESEARCHER / RESEARCH TEAM INFORMATION	
Lead Applicant Details	
Title	Dr.
First name (given name)	Lei
Surname (family name)	Xu
Gender	Male
Position at employing organisation/ institution	Associated Professor/PI
ORCID ID (https://orcid.org) or URL to academic profile	*0000-0003-2566-2118* *(if no ORCID or URL, please attach a short academic CV)*
Email	*jing.gao@ki.se*

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	*Tsinghua University*
City, Country	*Beijing, China*
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	*Åsa Wheelock*
1. Position / Role in analysis	*Associated Professor*
1. Organisation/Institution	*Respiratory Medicine Unit, Karolinska Institutet, Sweden*
2. Name	*Jing Gao*
2. Position / Role in analysis	*Researcher/ Co-investigator*
2. Organisation/Institution	*Respiratory Medicine Unit, Karolinska Institutet, Sweden*
3. Name	*Cui Zhou*
3. Position / Role in analysis	*PhD student/ Co-investigator*
3. Organisation/Institution	*Vanke School of Public Health, Tsinghua University, China*
4. Name	*Yikang Wang*
4. Position / Role in analysis	*PhD student/ Co-investigator*
4. Organisation/Institution	*Centre for Advanced Spatial Analysis, University College London, UK*

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)*
All applicant report no potential conflicts of interest in this study

Complete all sections of this form fully and return to covid19@iddo.org.

SECTION B: RESEARCH PLAN	
Title of Proposed Research	The interactive effect of clinical and non-clinical features in COVID-19 patients with chronic respiratory diseases (CRDs): a global perspective.
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)*

Chronic respiratory diseases (CRDs) such as chronic obstructive pulmonary disease, asthma, and interstitial lung disease are common comorbidities of COVID-19 (COVID-CRDs). Patients with these comorbidities constitute a vulnerable population in the pandemic, but risk of adverse outcomes varies worldwide. Beyond clinical and demographic factors, progression and outcome of COVID-CRDs may be influenced by social, economic, climatic, air pollution, and many other factors. Individual data from over 1500 centres worldwide are sufficiently spatially heterogeneous to investigate the interaction of environmental risk and clinical factors at scale. We propose a new approach to identify individual-level effects of such interactions, assess environment-associated differences in progression and outcome, and guide more precise intervention and treatment strategies.

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

Objectives

1.Explore interactions of clinical and non-clinical factors such as air pollution, climate, and built environment and how such interactions affect health outcomes of COVID-CRDs.

2. Evaluate differences in interaction of clinical and non-clinical factors among COVID patients with and without CRDs.

3. Identify environmental-risk-driven scenarios.

4. Investigate differences of COVID-CRD disease progression by time series analysis in different environmental-risk-driven scenarios.

5. Assess the effect of treatments (e.g. inhaled corticosteroids) on COVID-CRDs in different environmental-risk-driven scenarios.

Scientific value

Patients with COVID-CRDs are likely to constitute a large vulnerable population that has been affected by a complex of natural environmental, social environmental, and clinical factors since the pandemic began. Non-clinical factors such as age, deprivation indices, and air pollution have been confirmed as important risks for poor prognosis in the CRD population. Our previous studies presented significant regional variation in COVID-19 health outcomes, with country-level socio-environmental factors such as GDP, ageing, health burden, and PM2.5 impacting COVID-19 case fatality rate at the global scale. Accordingly, it is important to comprehensively analyse the characteristics of the COVID-CRD population and to predict health outcomes with a global perspective, and to identify multidimensional risk factors of COVID-CRDs at the individual level. Elucidating the interactions of clinical and non-clinical factors will help us in interpreting the significance of clinical indicators in this population in diverse scenarios.

Public health is inextricably linked to clinical health and geospatial context. Patient samples from multiple centres spanning the globe reflect geographical differences, and such a dataset will support us in teasing out the contributions of non-clinical factors to disease progression and outcome. We already have environmental remote sensing data (temperature, precipitation, pollutants, road networks, green space, etc.) and Twitter positioning-based mobility data, which could be useful in exploring health outcomes in the broader environmental context.

Further time series analysis of key clinical indicators of COVID-CRD progression will help reveal which environments are more associated with worse outcomes. Furthermore, it is worth evaluating intervention efficacy in the context of different environments. These findings will contribute to the development of precision medicine pipelines that guide prevention and treatment strategies to reduce adverse health outcomes in large populations with chronic conditions caused by infection, thereby reducing pressure on healthcare systems and excess deaths caused by the global run on healthcare resources.

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

Primary outcomes:

Case fatality rate, Need for intensive care.

Secondary outcomes:

Length of Intensive Care Unit (ICU) stay, Length of hospital stay, Need for treatment and support (oxygen therapy, invasive ventilation, non-invasive ventilation)

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

Data integration and description

We will statistically describe the clinical data, understand the data distribution, and deal with missing values. Environmental remote sensing data is obtainable to every country. Air pollution data including annual mean $PM_{2.5}$ and NO_2 (0.01 x 0.01 degree resolution) from Washington University inSt.Louis (https://sites.wustl.edu/acag/datasets/surface-pm2-5/). Climate data mainly includes daily temperature, precipitation, humidity, wind speed (0.1 x 0.1 degree resolution) from the ERA5 reanalysis database(https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5). Green space data extracted from the FROM-GLC database providing 30m resolution global land cover maps (http://data.ess.tsinghua.edu.cn). Road networks data extracted from Natural Earth database (https://www.naturalearthdata.com/). Elevation data from GTOPO30 database providing 1km resolution global maps. Environmental data will be matched by geocoding the hospital location of COVID-19 patients. To exclude noise from different SARS-CoV-2 variants, patient samples will be sequenced to identify variants and determine their classification.

Statistical analysis

Machine learning models such as random forest and eXtreme Gradient Boosting (XGBoost) will be used to identify important features that predict risk of outcomes, while SHapley Additive exPlanations (SHAP) will be used to interpret the machine learning models, identify risk and protective factors, and achieve risk stratification based on important risk factors. SHAP will also clarify risk strengths and thresholds and explain interactions between features at the individual level. In addition, causal forest models will be used to explore the causal pathways among environmental factors, clinical indicators, and clinical outcomes. Changes in key clinical

indicators over the course of the disease will also be described in different environmental scenarios. Finally, disease outcomes will be predicted through time series analysis methods such as a recurrent neural network (RNN).
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).*
This study has applied for an ethical exemption from Tsinghua University Human Research Ethics Committee (Project ID: 20220163). Data used for this project will be de-identified. All data acquired for this project will be stored in secure and confidential conditions in accordance with the Tsinghua University Research Data Management Policy, the approved study protocol and the Data Transfer Agreement.
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 proposed completion time of this work is February 2023. The research findings will be published in high-impact open access journals as the main channel of dissemination. We will express our great appreciation of the IDDO team for the valuable data and support for this research work in the acknowledgements.
Research Priorities Addressed (*suggested maximum 300 words)* *Provide details of how this research aligns with nationally or internationally set research priorities.*
Currently and for the foreseeable future, people remain at constant risk of COVID-19 infection. However, knowledge of the severity, clinical features, and prognostic factors of the COVID-19 with CRDs population in different environments and populations remains incomplete on a global scale. Moreover, the impact of the natural environment on individual disease progression and health outcomes is not yet clear. In this project, we expect to identify and assess adverse health outcomes in COVID-19 and CRD populations and, based on the findings, to develop predictive models to guide targeted interventions at the individual level, advance the realization of precision medicine, and guide national-level long-term policy development in the post-COVID-19 era to protect individual and population health and alleviate pressure on the healthcare system.
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 research will be conducted in collaboration with researchers from Tsinghua University (China), Karolinska Institutet (Sweden), and University College London (UK). The produced knowledge will be shared with them and our institution.
Funding *(suggested maximum 100 words)* *Provide details of how this research will be funded/resourced. Please name the source of funding.*
This project is funded by Vanke School of Public Health, Tsinghua University and National Natural Sciences Foundation of China (L2124008).

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.

The proposed work has been reviewed by the Vanke School of Public Health, Tsinghua University.

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.

Demographics:
AdmissionDate (454643)
SEX (477712)
COUNTRY (477712)
AGE (461648)
Ethnicity (if available)

Disposition:
DEATH (104487)
STILL IN HOSPITAL (11517)
TRANSFERRED (19248)
DISCHARGED (314907)
UNKNOWN (4105)

Healthcare Encounters:
HOSPITAL (468062)
INTENSIVE CARE UNIT (444013)
INTENSIVE CARE (65)

Clinical and Adverse Event:
Signs and Symptoms (451178) (symptoms and signs on admission)
Comorbidities (446278) (existing prior to admission and ongoing)

Vital Signs:
Temperature (199191)
Heart Rate (182871)
Pulse Rate (6387)
Respiratory Rate (188211)
Oxygen Saturation (189967)
Diastolic Blood Pressure (190739)
Systolic Blood Pressure (190902)
Mean Arterial Pressure (19931)
Weight (16635)
Height (14316)
Body Mass Index (8382)

Microbiology Specimen Test Result:
SARS CoV2 test result (333595)
Other respiratory pathogens (11977)
Other pathogens of public health importance (1657)

Influenza Virus (335236)
Malaria test result (371)
Plasmodium (406)
HIV-1/2 Antibody + HIV-1 p24 Antigen (602)
Human Immunodeficiency Virus (HIV) (644)
Bacteria test result (83127)
Adenovirus test result (86078)
Respiratory Syncytial Virus (RSV) test result (89994)
Coronavirus test result (92084)
Microbial Organism Identification (95573)

Laboratory Test Results:
White blood cell (WBC) count (177787)
Platelets (176693)
Hemoglobin (176349)
Lymphocyte count (171448)
Neutrophil count (170928)
C Reactive Protein (164301)
Urea Nitrogen (163116)
Creatinine (145621)
Sodium (145442)
Potassium (139668)
Bilirubin (124161)
Alanine Aminotransferase (119587)
Glucose (92389)
Prothrombin time (PT) (87980)
Lactic Acid (67954)
Hematocrit (53445)
Prothrombin INR (39063)
Hemoglobin A1C (38674)
Aspartate Aminotransferase (34565)
Lactate Dehydrogenase (32219)
Ferritin (23254)
Procalcitonin (20704)
Creatine Kinase (17952)
pH (8001)
Partial Pressure Carbon Dioxide (7996)
Bicarbonate (7916)
Partial Pressure Oxygen (7806)
Base Excess (7142)
D-Dimer (6862)
Activated partial thromboplastin time (APTT) (5349)
Lymphocytes (%) (4932)
Erythrocyte Sedimentation Rate (4913)
Neutrophil (%) (4755)
Troponin I (4578)
Erythrocytes (3448)
Erythrocytes Distribution Width (3448)
Mean Corpuscular Haemoglobin Concentration (MCHC) (3448)
Mean Corpuscular Hemoglobin (MCH) (3448)
Monocytes/Leukocytes (3448)
Monocyte count (3448)
Basophil (%) (3448)

Eosinophil (%) (3443)
Eosinophil count (3260)
CD4 (3035)
Mean corpuscular volume (MCV) (3032)
Fibrinogen (2953)
Protein (2804)
Basophil count (2629)
Albumin (2443)
Prothrombin time (PT) ratio (2167)
Direct Bilirubin (1540)
Urate (1309)
Interleukin 6 (1301)
Mean Platelet Volume (1265)
Platelet Hematocrit (1265)
Cholesterol (1103)
Alkaline Phosphatase (1097)
Amylase (924)
Activated partial thromboplastin time (APTT) ratio (922)
Iron (853)
Gamma Glutamyl Transferase (801)
Fibrinogen, Functional (498)
Chloride (482)
Calcium (406)
PaO2/FiO2 ratio (276)
Thrombin time (176)
Calcium, Ionized(146)
Calcium, Ionized pH Adjusted (146)
Troponin (132)
Carbon Dioxide (126)
Carboxyhemoglobin (124)
Deoxyhemoglobin (124)
Methemoglobin (124)
Oxyhemoglobin (124)
Magnesium (18)
Troponin T (4)

Interventions:
SUPPORTIVE CARE (13178),
CORTICOSTEROIDS FOR SYSTEMIC USE (10883),
ANTIBACTERIALS FOR SYSTEMIC USE (10824),
ANTIVIRALS FOR SYSTEMIC USE (10646),
OXYGEN (10242),
ANTIMYCOTICS FOR SYSTEMIC USE (10143),
MEDICATION (8423),
ANTIFUNGAL AGENTS (8247),
CARDIAC THERAPY (7838),
ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS,
NON-STEROIDS (7783),
AGENTS ACTING ON THE RENIN-ANGIOTENSIN SYSTEM (7609),
ANTITHROMBOTIC AGENTS (3794),
OTHER RESPIRATORY SYSTEM PRODUCTS (3578),
ANTIVIRAL AGENTS (3158),
MEDICAL HISTORY (3013),

ANTIMALARIALS (2885),
ANTIMALARIAL AGENTS (2765),
INVASIVE VENTILATION (2747),
EXTRACORPOREAL (2746),
PRONE POSITIONING (2746),
NON-INVASIVE VENTILATION (2746),
RENAL REPLACEMENT THERAPIES (2739),
ANTIBIOTIC AGENTS (1972),
CORTICOSTEROIDS (1865),
NSAIDS (268),
EXPERIMENTAL AGENTS (268)

Environmental risks (if available)

Potential available data:
prehospital medications related to COPD,
the need for home oxygen prior to admission

Population: United States of America (the), United Kingdom of Great Britain and Northern Ireland (the), South Africa, Malaysia, Brazil, France, Japan, Russian Federation (the), China, Italy, Germany,Netherlands (the), Canada, Ireland, Norway, Peru, Portugal, Colombia, Spain, Poland, Belgium, Australia, Malawi, Israel, Ukraine, Indonesia, Estonia, Kuwait, Argentina, Greece, Ecuador, Chile, Bolivia (Plurinational State of), Mexico

Since we need to match geographical features for our study, it will be helpful to obtain the name of the hospital where the patient is located. In terms of clinical indicators, it would be better if pulmonary function indicators such as FEV1 and FVC were available. Prehospital medications, the need for home oxygen, and other important treatment of patients prior to admission are also will help to improve the understanding of the severity of a patient's underlying disease prior to admission. Moreover, we would like to adjust for as many confounding factors as possible and build global-scale studies, so more variables and populations, larger data volumes, and longer time coverages are better for our research.