CODA-19: a Collaborative Data Analysis Platform to Improve Clinical Care in Patients with COVID-19

Research Protocol, version 6.0

Principal Investigator
Michaël Chassé MD PhD FRCPC, CHUM, University of Montréal

Co-principal Investigators
David Buckeridge, MD, PhD, McGill University Health Centre Research Institute, McGill
Jonathan Afilalo, MD, MSc, Jewish General Hospital, McGill University
Han Ting Wang MD, FRCPC, Maisonneuve-Rosemont Hospital, University of Montreal
Yiorgos A. Cavayas, MD, MSc, FRCPC, Hôpital Sacré-Coeur de Montréal, University of Montreal
Alexis Turgeon MD, MSc, FRCPC, CHU de Québec-Université Laval Research Center, Université Laval
Patrick Archambault, MD, MSc, FRCPC, CISSS Chaudière-Appalaches, Université Laval
Joelle Pineau, PhD, Mila, McGill University

Co-Investigators
Marc Afilalo, MD, Jewish General Hospital, McGill University
François Martin Carrier, MD, MSc, FRCPC, CHUM, University of Montreal
Emmanuel Charbonney, MD, PhD, University of Montreal
Carl Chartrand-Lefebvre, MD, MSc, FRCPC, CHUM, University of Montreal
Joseph Paul Cohen, PhD, Mila, University of Montreal
Audrey Durand, PhD, Mila, Université Laval
Madeleine Durand, MD, MSc, FRCPC, University of Montreal
Shane W. English, MD, MSc, The Ottawa Hospital, University of Ottawa
Philippe Jouvet, MD, PhD, MBA, Hôpital Sainte-Justine, University of Montreal
Louis-Antoine Mullie, MD, FRCPC, CHUM, University of Montreal
Esli Osmanlliu MD, MSc (cand), FRCPC, McGill University Health Centre Research Institute, McGill
Guillaume Plourde, MD, PhD, CHUM, University of Montreal
Brent Richards, MD, MSc, Jewish General Hospital, McGill University
Antony Robert, MD, MASc, FRCPC, McGill University Health Center, McGill University
Michaël Sauthier, MD, MBI, Hôpital Sainte-Justine, University of Montreal
Nicolas Sauthier, MD, MSc (cand), CHUM, University of Montreal
An Tang, MD, MSc, FRCPC, CHUM, University of Montreal
Martin Girard, MD, MSc, FRCPC, CHUM, University of Montreal
François Lamontagne, MD, MSc, CHUS, University of Sherbrooke

Participating Sites
Centre Hospitalier de l'Université de Montréal (CHUM)
Hôpital Maisonneuve-Rosemont (CIUSSS de l'Est-de-l'Île-de-Montréal)
Hôpital Général Juif (CIUSSS du Centre-Ouest-de-l'Île-de-Montréal)
Centre Universitaire Santé McGill (CUSM/MUHC/Montreal Children Hospital)
Hôpital Sacré-Cœur de Montréal (CIUSSS du Nord-de-l'Île-de-Montréal)
Centre Hospitalier Universitaire Sainte-Justine (CHU Sainte-Justine)
Centre Hospitalier Universitaire de Québec - Université Laval (CHU de Québec)
CISSS de Chaudière-Appalaches
Centre hospitalier universitaire de Sherbrooke (CHUS)

Project Summary

Objective: To build a multi-centre data infrastructure enabling the rapid development and prospective validation of predictive models to aid the clinical management of Coronavirus Disease 2019 (COVID-19) and optimize healthcare resource utilization in response to the COVID-19 pandemic.

Rationale: An evidence-based Canadian response to the ongoing COVID-19 pandemic requires the collection and analysis of high-quality, structured data on patients in whom COVID-19 is suspected or confirmed. Data infrastructures are urgently needed in order to develop and prospectively validate predictive models aimed at facilitating early diagnosis of COVID-19, adapting management to distinct disease presentations, identifying patients at risk of adverse outcomes, and forecasting resource usage.

Preliminary Results: We have developed **CODA-19**, a data repository of all patients with suspected or confirmed COVID-19 at 9 hospital sites in Québec and 1 in Ontario. In addition to standard clinical characteristics and outcomes, this repository contains **multi-modality biological signals**, including **clinical parameter trends** (e.g. laboratory tests, vital signs, ventilator settings over time), **2D and 3D chest imaging data**, and other sensor recordings (e.g. electrocardiograms).

Aims: We will develop a collaborative analysis platform to conduct **real-time, prospective clinical validation** of epidemiological and machine learning models developed using **CODA-19** under 4 domains: **1) Diagnostic risk stratification**: To rapidly estimate the probability of COVID-19 in patients presenting with compatible symptomatology; **2) Clinical phenotypes**: To identify distinct clinical phenotypes in patients with confirmed COVID-19, and assess whether phenotypes influence the response to supportive treatments; **3) Early warning system**: To identify early warning signs that predict the time to an adverse outcome among inpatients with confirmed COVID-19; and, **4) Health system resource use**: To forecast the need for beds, materials and staff, identify at-risk thresholds for equipment shortages, and optimize the delivery of care for patients with and without COVID-19.

Team: We have established partnerships with **10 hospital sites** – including 6 sites in Montréal, the COVID-19 epicenter in Canada – to recruit patients into **CODA-19** and conduct prospective clinical validation of the models developed. Our multi-disciplinary team combines **leading expertise** in the fields of machine learning, data science, epidemiology, and biostatistics, with **strong clinical health data research experience** in radiology, internal medicine, emergency medicine, and critical care.

Knowledge Translation: We will develop and distribute a web-based, interactive **diagnostic risk stratification tool** as well as a **resource-planning forecasting and simulation tool**. We will implement and validate an **early warning system** to identify patients at risk of deterioration in real-time, as well as a **COVID risk dashboard** to simplify bed management at each site. We will distribute an **open-source library** to facilitate access to CODA-19 data by authorized researchers.

Relevance: Results will help mount an evidence-based response to the COVID-19 pandemic by providing tools to inform triage at the frontlines, to select the optimal care setting for each patient, and to ensure equitable delivery of health care services for patients with and without COVID-19. Our data analysis platform will enhance the readiness of Canadian hospitals to future pandemics and help catalyze further multi-centric research efforts between participating sites.

1. Background

COVID-19 is a highly contagious acute respiratory illness that has undergone rapid global spread in the beginning of 2020. It is a major national public health threat that has disproportionately affected vulnerable Canadians, such as immunocompromised patients, older patients, and those living in long-term care facilities [1,2]. There is a pressing need to develop predictive models in large patient cohorts to enable early diagnosis of COVID-19, adapt management strategies to individual risk profiles, identify early warning signs for clinical deterioration, forecast resource usage, and optimize health system organization. Scalable data sharing and distributed analysis infrastructures are urgently needed to enable the development and clinical validation of robust predictive models across multiple sites, and translate models into impactful decision support tools.

In collaboration with **10 hospital sites** at the Canadian epicenter of the pandemic, we have built a **large repository** of anonymized, **multi-modality data** from patients with suspected or confirmed COVID-19. Our team of **leading experts** in **clinical research** and **health data science** is uniquely positioned to tackle the challenge of putting big data at the service of clinicians and administrators managing COVID-19. Using a **scalable collaborative analysis platform** to conduct multicenter **prospective model validation**, we will develop **point-of-care decision support tools** and **provide actionable insights** to support Canada's response to COVID-19.

1.1 Clinical Domains of Research

CODA-19 will be used to develop and **prospectively validate** epidemiological and machine learning prediction models, both supervised and unsupervised, answering **4 key clinical needs**: **1)** achieving early diagnostic risk stratification, in order to improve early patient triage, and reduce nosocomial transmission; **2)** understanding how distinct **clinical** phenotypes **respond to alternative treatments, to provide individualized risk assessment and management;** 3) **identifying** early warning signs **that herald clinical deterioration, to enable proactive monitoring and treatment; and** 4) **forecasting** resource usage**, to optimize health system resource use.**

1.2 Prospective Clinical Validation

We will conduct **ongoing, prospective clinical validation** of the epidemiological and machine learning models developed in the 4 clinical domains. A **scalable data analysis platform** has been developed and will be deployed across sites to continuously **aggregate, monitor and analyze** the results of model validation (**Figure 2**). Patients meeting the inclusion criteria for each model will be identified on an **automated, real-time basis** and will be used to prospectively validate model performance. New patient data will be incorporated on an ongoing basis to refine existing models. State-of-the-art **distributed and federated learning** strategies will be used to train statistical models across sites, while minimizing the amount of patient-level data sharing [3-5].

2 Methods

2.1 Methods common to all domains

We will conduct a multi-center, retrospective and prospective observational data collection study. The study will initially be conducted at 10 hospital sites in Montreal, Quebec, Canada (Centre Hospitalier de l'Université de Montréal; Hôpital du Sacré-Coeur de Montréal; Hôpital Maisonneuve-Rosemont; Hôpital Général Juif; Centre Hospitalier de l'Université McGill; CHU de Québec-Université Laval; CISSS Chaudière-Appalaches; and Centre Hospitalier Universitaire Sainte-Justine-CHU de Sainte-Justine, Centre hospitalier universitaire de Sherbrooke) and 1 in Ontario (The Ottawa Hospital).

All patients in whom a PCR test for COVID-19 was performed after December 31st, 2019 at one of the participating institutions, or for which there is a flag stating that they were infected by the SARS-COV2 virus, whether or not they were admitted to the hospital, will be included in this study. Given the ongoing nature of the COVID-19 pandemic, no set end date is currently fixed.

A data specification has been published (http://www.coda19.com/) in order to standardize data collection at each site and will be updated. Standard procedures for data de-identification will also be published and enforced by participating hospitals. Each of the participating institutions will be responsible for the local development and maintenance of a database at their site, as well as for data de-identification procedures. The data infrastructure and data flow is presented in Figure 1 and Figure 2.

Figure 1. Model development via secure data sharing infrastructure

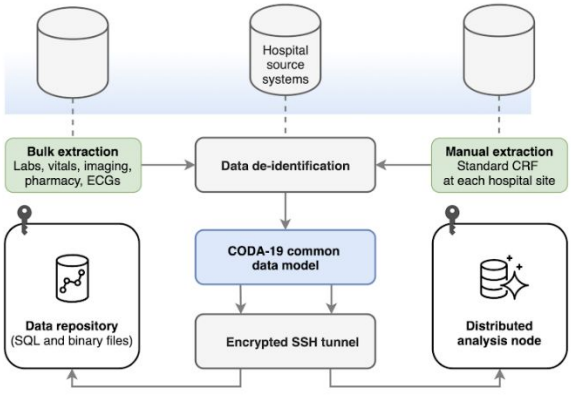
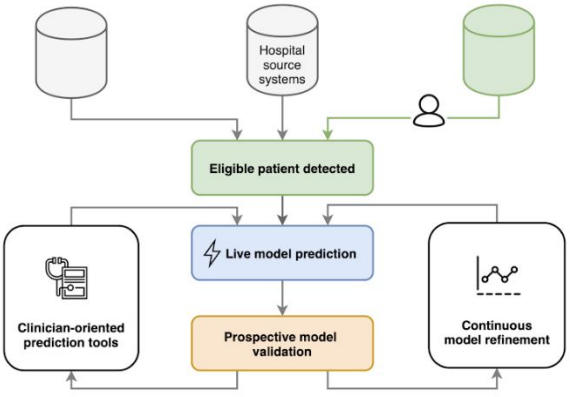


Figure 2. Continuous prospective clinical validation and model refinement



A governance framework will be adopted in order to regulate access to de-identified data from the database. Projects requesting the use of the database will be approved by an executive committee, consisting of a representative at each site. Data sharing agreements will be adopted for projects requiring access to data from multiple sites. (See Appendix 2 for governance)

Outcomes and covariates

An exhaustive search of microbiology registries will be performed at each participating hospital to identify patients who underwent testing for COVID-19 by PCR. Tests will be labeled according to the status of the patient at the time of testing (outpatient / inpatient), and according to whether or not the test result was transferred from an outside institution.

The following variables will be extracted in all patients when available, from two year prior to COVID testing up until 1 year following the date of the first COVID test:

- Non-identifying demographics, such as age, sex, ethnicity
- Other non-identifying patient characteristics, such as height, weight
- History of present illness, including exposure/travel and symptom history
- Medical history, including major comorbidities and frailty assessment
- Home medications, allergies and pertinent habits
- Clinical / nursing observations, such as vital signs
- Oxygenation and ventilation parameters
- Radiological data (X-ray, CT scan or ultrasound)
- All laboratory tests performed
- PCR tests (including COVID testing)
- Microbiological cultures
- Standard and continuous electrocardiogram data
- Other continuous monitoring signals, such as pulse oximetry
- Treatments (conventional and experimental)
- In-hospital complications and outcomes
- Mortality and cause of death

Data collection

Patient characteristics and outcome data will be extracted from the participating hospital's electronic health records, supplemented by manual chart review if necessary.

A data specification has been published and is being maintained (http://www.coda19.com/) detailing the types, attributes, and interrelationships of objects found in CODA-19. Each participating site will be responsible for encoding data into a storage system that conforms to this data specification. The data sources used to encode data into different sections of the database may vary at each site. An SQL schema corresponding to the data specification has been shared with the participating institutions, in order to ensure the consistency of data across sites.

2.2 Methods specific to each domain

Domain 1: Diagnostic Risk Stratification

Aim: To develop and validate methods to estimate the probability of COVID-19 in patients presenting to an acute care setting for symptoms or complications of COVID-19.

Overview: Polymerase chain reaction (PCR) testing for COVID-19 has a 6-19% false negative rate, a variable turnaround time, and is vulnerable to delays when the number of tests performed increases [6,7]. Predicting COVID-19 status from readily available clinical information would improve early patient orientation, prior to the availability of a PCR test result. Automated prediction systems can also assist in identifying missed cases and nosocomial transmission on an ongoing basis.

Methods: Epidemiological and machine learning models will be developed to identify determinants and predictors of COVID-19 status among patients presenting to the ER or transferred from another hospital for symptomatology compatible with COVID-19, in whom a COVID-19 PCR test was sampled within ± 5 days of presenting to care, and in whom laboratory tests and chest X-ray were performed within 24h of arrival. Patients with any positive test in the window period will be considered positive. A. Epidemiological model. In order to control for site-specific differences in patient populations, and variations in COVID-19 treatment designation status over time, COVID-positive patients will be matched in a 1:4 ratio to controls for age, sex, calendar week, hospital site [8]. Conditional logistic regression will be performed in the matched patient cohort, with predefined independent variables identified via a review of the literature [9]. The most abnormal result in the first 24 hours will be sampled for discrete-time data. Missing data will be imputed using multiple imputation by chained equations (MICE) with random forest (RF) regressors [10,11]. All analyses will be stratified for significant comorbidities that may affect COVID-19 outcomes, such as diabetes and immunocompromising conditions at baseline [12]. Machine learning model. Gradient boosting techniques will be used to identify features of interest among numerical and categorical input variables in the matched patient cohort [13, 14]. Synthetic Minority Oversampling Technique (SMOTE) will be used to balance classes in the training data set [15-17]. Supervised classifiers will be trained to estimate the probability of a COVID-19 diagnosis, based on imaging features and other selected input variables. Convolutional autoencoders will be used in order to learn a feature map representation for chest X-rays, using our data and publicly available chest X-ray data sets [18,19]. Model performance will be assessed using k-fold validation, with careful separation of training, validation and test sets. Predicted probabilities will be calibrated using Platt's method [20-22]. Prediction explanation techniques will be applied to gain insights into model reasoning [23]. An alternative project previously submitted in a previous version of this study is detailed in Appendix 1 and continues to be conducted using the proposed infrastructure.

Expected output: We will develop and validate clinically an interactive diagnostic risk stratification tool that clinicians can use to estimate the probability of a COVID-19 diagnosis at the point-of-care. Using our predictive model, we will deploy a COVID-19 risk dashboard providing a color-coded overview of risk categories for patients in the ER. Based on epidemiological modeling, we will develop and validate a simple clinical decision rule to facilitate patient triage by first responders.

Domain 2: Diagnostic Risk Stratification

Aim: To identify clinical phenotypes in patients with COVID-19, assess their interaction with the response to specific supportive management strategies, and evaluate their association with outcomes.

Overview: At least two clinical phenotypes of COVID-19 pneumonia have been described, on the basis of distinctive lung mechanics and radiological findings, and implications for clinical management have been inferred [24-25]. Multiple disease phases have also been described, according to whether viral pathogenicity or host inflammatory response is predominant [26]. In a subset of patients with COVID-19, a hypercoagulable phenotype has been observed, prompting recommendations for intensified antithrombotic therapy [27-29]. Although it is not yet known if the existence of distinct disease phenotypes is borne out by clinical data, the identification of such entities may have important implications for routine patient management and trial design [30].

Methods: Clinical phenotypes will be derived in an unsupervised fashion among patients with an inpatient admission at one of the participating sites for ≥ 72 hours for an acute respiratory illness due to COVID-19, in whom a chest X-ray is available within ± 72 hours of admission. Patients meeting inclusion criteria will be split into a derivation and a validation cohort (75%/25%) [31]. Inputs for phenotype derivation will include cross-sectional data (e.g. age, sex), time-varying data (e.g. laboratory tests, vital signs), and chest X-ray data. Missing data will be imputed using MICE-RF [10]. Clustering will be performed using deep embedded clustering [32]. Silhouette coefficients will be assessed, and the reproducibility of phenotypes will be evaluated [33]. Chord diagrams will be used to visualize the relation between phenotypes and organ dysfunction, as assessed by Sequential Organ Failure Assessment (SOFA) sub-scores (pulmonary, cardiovascular, renal, hepatic, coagulation) [34-35]. The association between phenotypes and biomarkers of infection and inflammation will be assessed. The effects of phenotype assignment on patient-centered outcomes (e.g. all-cause mortality, ICU-free days), as well as organ failure (as assessed by SOFA score), will be assessed using linear mixed effects models, with participating centers as random effects and phenotypes as fixed effects. Phenotype × treatment interactions will be evaluated to determine if phenotypes are associated with differential responses to specific supportive therapies. The association between phenotype assignment and the response to candidate treatments (e.g. optimal ventilation strategies) will be evaluated using a generalized mixed effects model [36].

Expected output: We will develop and validate clinically the first large-scale, multi-dimensional phenotypic analysis of patients with COVID-19. The identification and validation of candidate disease phenotypes will help our understanding of the pleomorphic clinical expression patterns of COVID-19. Deriving distinct associated organ dysfunction profiles will identify opportunities for intervention through adjustments in supportive care, and drive hypotheses for future interventional trials.

Domain 3: Early Warning System

Aim: To identify factors that predict and determine the time to an adverse clinical outcome in patients with COVID-19, and find early warning signs that predict deterioration in the acute care setting.

Overview: Epidemiological data suggests that older adults are disproportionately affected by complications of COVID-19 [37]. In addition to age, other risk factors may influence patient evolution and predict clinical deterioration. Recent evidence suggests that patients who at first look stable may suddenly deteriorate, the so-called "*Happy hypoxemic patients*" [38]. Accordingly, clinicians may preemptively admit patients to higher acuity units, leading to suboptimal use of scarce ICU resources. Risk stratification tools are urgently needed to determine which patients are likely to die or have an unfavourable outcome while hospitalized with COVID-19.

Methods: Predictors and determinants of adverse clinical outcome, defined as the need for invasive mechanical ventilation or death, will be analyzed among patients admitted for an acute respiratory illness due to COVID-19 to an inpatient unit at one of the participating sites. **A. Epidemiological model.** Continuous live demographic, treatment and outcome data will be collected, presented and analyzed in a descriptive and univariate manner. Marginal Cox regression models will be fitted to determine the association between predefined independent variables and time to occurrence of an adverse clinical outcome, using robust variances to take into account center effect and clustering [39]. **B. Machine learning model.** Machine learning models will be developed to assess the incremental value of using multiple sampling time points and incorporating imaging data into predictions. A semi-supervised anomaly detection model, using long-short term memory networks (LSTMs) and deep autoencoders, will be used to predict the occurrence of an adverse clinical outcome within the next 24 hours, based on data sampled in the preceding 72 hours [40,41]. Nonparametric regression models (recurrent neural networks, random forests) will be evaluated to numerically estimate the time to adverse clinical outcome. Mean absolute percentage errors will be determined using stratified k-fold validation, with stratification by site [42].

Expected output: We will implement and validate clinically an early warning system for inpatients with COVID-19, using continuous predictions to identify patients at risk of clinical deterioration in real time. Live and accurate Descriptive epidemiological/phenotypic data and their subsequent modelling will help identify factors that determine an unfavourable clinical evolution early in the course of illness, and help clinicians adapt treatment to goals of care.

Domain 4: Health System Resource Use

Aim: To develop and validate models to predict resource demand by patients with suspected or confirmed COVID-19, and to optimize allocation of resources when the demand exceeds capacity.

Overview: The COVID-19 crisis has led to unprecedented demand for health system resources, resulting in at times catastrophic situations in locations where demands exceeded available capacity. Models that forecast demand for different types of resources (e.g. ward beds, ICU beds, ventilators, protective equipment) are needed urgently to inform clinical and administrative decision-making during the pandemic response, and make the most efficient and equitable use of hospital resources.

Methods: We will first provide and report live descriptive statistics of resource usage related to COVID19 activity in the participating centers. Under a mandate from the Québec government, we have already developed a Markov state-transition model, trained using data on all hospital admissions for COVID-19 in Quebec, to project demand for health system resources (i.e., ward beds, ICU beds, ventilators) at the provincial and regional level (https://covid.mchi.mcgill.ca) [43]. The

model takes as input the projected rate of outside arrivals and estimates state-transition parameters from the data for movement through the ED, ward beds, ICU beds, ventilators, and discharge or death. The first extension to this model will be to stratify the patient states by features that influence patient trajectories (from Domain 3) and to expand the resources measured, including human and physical resources. The second extension will incorporate capacity constraints into the more richly-stratified Markov model and estimate parameters for decline in clinical status when patients cannot receive the required level of care due to capacity constraints. In a third step, we will explore machine learning methods such as LSTMs for predicting demand and optimizing resource allocation [44]. To simulate patient arrivals, we will sample from historical admissions, accounting for hospital closures and policies to allocate patients across hospitals [45]. Monte Carlo simulation models will be used to identify key thresholds that predict resource shortages.

Expected output: Models that anticipate disease activity peaks could be used to predict when demand is likely to exceed capacity, triggering actionable responses to balance load across sites. An online forecasting dashboard will be created and disseminated. A web-based resource planning simulation tool will be created, allowing clinicians and administrators to visualize bed occupancy projections under adjustable constraints, thus facilitating the optimization of bed capacity in anticipation of surges, informing public health policies, such as confinement, and optimizing scheduled clinical activities.

3 Limitations

Given the imperfect sensitivity of the PCR test for COVID-19, misclassification could occur in patients who were only tested for the disease once. This source of error is of lesser concern in the subset of hospitalized patients, in whom PCR should be repeated following a first negative result. Given that the PCR test is currently the gold standard for diagnosing COVID-19, this source of error is not modifiable.

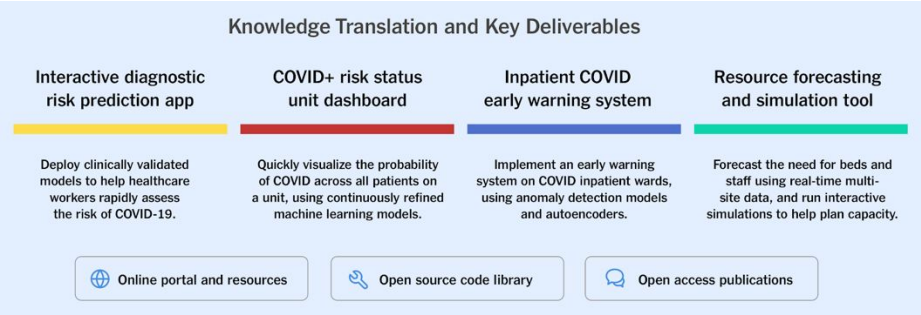
Data duplication is a potential source of error that will need to be addressed at each site, using local best practices. Automated scripts will be developed to identify potential duplicate records and flag them for manual review.

Variations in image cropping, noise, artefacts, and quality are real-world challenges that need to be considered when building models using diagnostic medical imaging. However, the appropriate strategies to tackle these challenges are specific to the learning problem being solved; these strategies will be detailed in sub-study protocols.

Since the care of COVID-19 patients implies that contact with patients is minimized, test ordering may not follow usual practices. Thus, interpretation of results should bear in mind that patients who underwent a particular test (e.g. chest CT) are likely to differ in highly significant ways from the overall study population. Where relevant, care should be taken in substudies to examine the effect of confounding by indication for the imaging test being studied by assessing for systematic differences between patients who underwent the test and patients who did not.

4 Knowledge Translation

We will develop and distribute a web-based, interactive **diagnostic risk stratification tool** as well as a **resource-planning forecasting and simulation tool**. We will implement and validate an **early warning system** to identify patients at risk of deterioration in real-time, as well as a **COVID risk dashboard** to simplify bed management at each site. We will distribute an **open-source library** to facilitate access to CODA-19 data by authorized researchers. We will publish in peer-reviewed journals all findings regarding descriptive epidemiological/phenotypic models as well as the research output produced from this collaborative.



5 Ethical and privacy considerations

Ethics approval will be sought from the Ethics Review Board at each participating hospital. Given the non-interventional nature of the study, individual patient consent will not be required.

Multiple strategies will be deployed to ensure that patient privacy is protected during the development of the database (see also figure 2):

- Patient identifiers will be scrambled using a strong hashing procedure (PBKDF2 with 100,000 iterations of SHA-512) with unique, per-patient random salt. Random salts will be securely stored at each institution, and will not be shared between sites.
- Metadata will be stripped from DICOM imaging studies (e.g. X-rays and CT scans) prior to storage for analysis. Only a whitelisted subset of technical, non-identifying DICOM headers will be kept in the database. De-identified (anonymous at the researcher level, see above) scan images will be tagged using a unique study identification number, which will be used to match imaging studies. De-identified data will be stored separately from clinical identifiers.
- If required, data transfers between participating institutions will consist solely of de-identified data, and will be performed using a secure tunnel for confidential data transport with verifiable integrity (SSH), with 4096-bit RSA keys.
- For any research output resulting from this repository, elements of a unique type (e.g. a specific laboratory test) with a count less than 5 in the overall output will be censored and replaced with “<=5).

6 Team Expertise and Roles

Dr. Michaël Chassé will lead the team. MC is an intensivist, health data scientist at the Centre Hospitalier de l'Université de Montréal (CHUM), and Associate Professor at University of Montreal. He is also an IVADO.ca professor and the scientific director of the CHUM Center for Integration and Analysis of Medical Data (CITADEL). He has led several national clinical trials, including multi-centre data science projects. He brings together a group of scientists and professionals specialized in health data science (Pineau, Buckeridge, Cohen, Osmanlliu, Mullie, Sauthier), biostatistics (Buckeridge, Carrier, Turgeon), bioinformatics and machine learning (JAfilalo, ADurand, Pineau, Mullie, Sauthier), epidemiology (JAfilalo, Carrier, English, MDurand, Plourde, Turgeon), knowledge transfer (Archambault, MAfilalo), clinical informatics (Robert, Sauthier), as well as physicians with expertise in radiology (Chartrand-Lefebvre, Tang), adult/pediatric critical care (Wang, Cavayas, English, Jouvet, Turgeon, Archambault, Charbonney, Carrier, Plourde, Mullie, Jouvet, Sauthier) and emergency medicine (Archambault, Esli, Robert, MAfilalo). Our collaborative effort brings together the networks of University of Montreal, Université Laval, University of Ottawa and McGill University, as well as the Quebec Institute of Artificial Intelligence (Mila) (Cohen, Pineau, Buckeridge, ADurand) Reasoning and Learning Lab co-lead (Pineau). The team will be managed by an executive committee (PIs, Mullie, Plourde), and for daily management, by a steering committee (PIs and co-Is, domain-specific coordination, clinical validation).

7 Timeline

Data will be analyzed on an ongoing basis, beginning immediately once approval from the Ethics Review Board is granted. A public informational website (www.coda19.com) will be set up to promote a collaborative effort on the development of CODA-19 and recruit other centers. The results will be promoted at machine learning and medical imaging conferences such as Machine Learning for Healthcare (ML4HC), Medical Imaging with Deep Learning (MIDL), and Radiological Society of North America (RSNA).

Although the rapidly evolving situation of the COVID-19 pandemic makes it difficult to provide time estimates for publication, work on the manuscripts for sub-studies is expected to begin in June 2020, with publications projected in December 2020. Work on the main database manuscript is expected to begin in September 2020, with publication projected in January 2021.

8 References

1. Wang, Wang, et al. "Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures." Journal of Medical Virology, Mar. 2020.
2. COVID-19 Daily Epidemiology Update [Internet]. [cited 2020 May 11]. Available from: https://www.canada.ca/content/dam/phac-aspc/documents/services/diseases/2019-novel-coronavirus-infection/surv-covid19-epi-update-eng.pdf
3. Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc JAMIA. 2015 Nov;22(6):1212–9.
4. Choudhury O, Park Y, Salonidis T, Gkoulalas-Divanis A, Sylla I, Das AK. Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning. AMIA Annu Symp Proc AMIA Symp. 2019;2019:313–22.
5. Ma J, Zhang Q, Lou J, Ho JC, Xiong L, Jiang X. Privacy-Preserving Tensor Factorization for Collaborative Health Data Analysis. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management [Internet]. Beijing China: ACM; 2019 [cited 2020 May 12]. p. 1291–300. Available from: https://dl.acm.org/doi/10.1145/3357384.3357878.
6. He J-L, Luo L, Luo Z-D, Lyu J-X, Ng M-Y, Shen X-P, et al. Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. Respir Med. 2020 Apr 21;168:105980.
7. Kim H, Hong H, Yoon SH. Diagnostic Performance of CT and Reverse Transcriptase-Polymerase Chain Reaction for Coronavirus Disease 2019: A Meta-Analysis. Radiology. 2020 Apr 17;201343.
8. Rasmy L, Wu Y, Wang N, Geng X, Zheng WJ, Wang F, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. J Biomed Inform. 2018;84:11–6.
9. Kuo C-L, Duan Y, Grady J. Unconditional or Conditional Logistic Regression Model for Age-Matched Case-Control Data? Front Public Health. 2018;6:57.
10. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min. 2017 Dec;10(6):363–77.
11. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007 Jun;16(3):219–42.
12. Guo W, Li M, Dong Y, Zhou H, Zhang Z, Tian C, et al. Diabetes is a risk factor for the progression and prognosis of COVID-19. Diabetes Metab Res Rev. 2020 Mar 31;e3319.
13. Kilic A, Goyal A, Miller JK, Gjekmarkaj E, Tam WL, Gleason TG, et al. Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery. Ann Thorac Surg. 2019 Nov 7;
14. Xu Y, Yang X, Huang H, Peng C, Ge Y, Wu H, et al. Extreme Gradient Boosting Model Has a Better Performance in Predicting the Risk of 90-Day Readmissions in Patients with Ischaemic Stroke. J Stroke Cerebrovasc Dis Off J Natl Stroke Assoc. 2019 Dec;28(12):104441.
15. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009 Sep;21(9):1263–84.
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 1;16:321–57.

17. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics. 2010 Dec;11(1):523.
18. Abiyev RH, Ma'aitah MKS. Deep Convolutional Neural Networks for Chest Diseases Detection. J Healthc Eng. 2018 Aug 1;2018:1–11.
19. Irvin, Jeremy, et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." AAAI Conference on Artificial Intelligence, 2019, http://arxiv.org/abs/1901.07031.
20. Chen W, Sahiner B, Samuelson F, Pezeshk A, Petrick N. Calibration of medical diagnostic classifier scores to the probability of disease. Stat Methods Med Res. 2018;27(5):1394–409.
21. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. J Am Med Inform Assoc JAMIA. 2012 Apr;19(2):263–74.
22. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth Isotonic Regression: A New Method to Calibrate Predictive Models. AMIA Summits Transl Sci Proc. 2011 Mar 7;2011:16–20.
23. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv160204938 Cs Stat [Internet]. 2016 Aug 9 [cited 2020 May 11]; Available from: http://arxiv.org/abs/1602.04938
24. Gattinoni L, Coppola S, Cressoni M, Busana M, Rossi S, Chiumello D. Covid-19 Does Not Lead to a "Typical" Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med. 2020 Mar 30;
25. Gattinoni L, Chiumello D, Caironi P, Busana M, Romitti F, Brazzi L, et al. COVID-19 pneumonia: different respiratory treatments for different phenotypes? Intensive Care Med [Internet]. 2020 Apr 14 [cited 2020 May 11]; Available from: https://doi.org/10.1007/s00134-020-06033-2
26. Siddiqi et al. "COVID-19 Illness in Native and Immunosuppressed States: A Clinical-Therapeutic Staging Proposal" Journal of Heart and Lung Transplantation, Mar. 2020.
27. Kollias A, Kyriakoulis KG, Dimakakos E, Poulakou G, Stergiou GS, Syrigos K. Thromboembolic risk and anticoagulant therapy in COVID-19 patients: emerging evidence and call for action. Br J Haematol. 2020 Apr 18;
28. Spiezia L, Boscolo A, Poletto F, Cerruti L, Tiberio I, Campello E, et al. COVID-19-Related Severe Hypercoagulability in Patients Admitted to Intensive Care Unit for Acute Respiratory Failure. Thromb Haemost. 2020 Apr 21;
29. Tang N, Li D, Wang X, Sun Z. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. J Thromb Haemost JTH. 2020;18(4):844–7.
30. Prescott HC, Calfee CS, Thompson BT, Angus DC, Liu VX. Toward Smarter Lumping and Smarter Splitting: Rethinking Strategies for Sepsis and Acute Respiratory Distress Syndrome Clinical Trial Design. Am J Respir Crit Care Med. 2016 15;194(2):147–55.
31. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics. 2011 Apr 8;4:31.
32. Enguehard J, O'Halloran P, Gholipour A. Semi-Supervised Learning With Deep Embedded Clustering for Image Classification and Segmentation. IEEE Access. 2019;7:11093–104.
33. Pourahmad S, Pourhashemi S, Mohammadianpanah M. Colorectal Cancer Staging Using Three Clustering Methods Based on Preoperative Clinical Findings. Asian Pac J Cancer Prev APJCP. 2016;17(2):823–7.

34. Seymour CW, Kennedy JN, Wang S, Chang C-CH, Elliott CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA. 2019 28;321(20):2003–17.
35. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996 Jul;22(7):707–10.
36. Diaz FJ. Measuring the individual benefit of a medical or behavioral treatment using generalized linear mixed-effects models. Stat Med. 2016 15;35(23):4077–92.
37. McMichael TM, Currie DW, Clark S, Pogosjans S, Kay M, Schwartz NG, et al. Epidemiology of Covid-19 in a Long-Term Care Facility in King County, Washington. N Engl J Med. 2020 Mar 27;
38. Archer SL, Sharp WW, Weir EK. Differentiating COVID-19 Pneumonia from Acute Respiratory Distress Syndrome (ARDS) and High Altitude Pulmonary Edema (HAPE): Therapeutic Implications. Circulation. 2020 May 5;
39. Chen Y, Chen K, Ying Z. ANALYSIS OF MULTIVARIATE FAILURE TIME DATA USING MARGINAL PROPORTIONAL HAZARDS MODEL. Stat Sin. 2010;20(33):1025–41.
40. Laptev N, Yosinski J, Li LE, Smyl S. Time-series Extreme Event Forecasting with Neural Networks at Uber. :5.
41. Maya S, Ueno K, Nishikawa T. dLSTM: a new approach for anomaly detection using deep learning with delayed prediction. Int J Data Sci Anal. 2019 Sep;8(2):137–64.
42. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–1143. (IJCAI'95).
43. Iskandar R. A theoretical foundation for state-transition cohort models in health decision analysis. Gontis V, editor. PLOS ONE. 2018 Dec 11;13(12):e0205543.
44. Zhu X, Fu B, Yang Y, Ma Y, Hao J, Chen S, et al. Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinformatics. 2019 Nov;20(S18):575.
45. Merad M, Martin JC. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. Nat Rev Immunol. 2020 May 6;1–8.

Appendix 1: Alternative analysis (Theme 1)

Predicting key clinical outcomes in patients with confirmed COVID-19 using deep neural networks

Project lead: Joelle Pineau, PhD, Mila, McGill University

This project will focus on developing machine learning methods to predict ICU admission, prolonged mechanical ventilation, and in-hospital mortality in hospitalized patients with confirmed COVID-19. Patients with confirmed COVID-19 who underwent a chest X-ray within 48 hours of diagnosis will be included.

The population will be divided into training, validation and testing groups according to standard proportions. For each clinical outcome, relevant clinical and paraclinical variables will be selected using feature selection algorithms (e.g. XGBoost). Convolutional neural networks will also be trained to predict clinical outcomes (eg. length of hospital stay, admission to intensive care, hospital mortality) from imaging data. Several architectures of pre-trained convolutional neural networks as well as untrained convolutional neural networks will be tested in an exploratory cohort consisting in a subset of the training cohort.

Supervised classification models (eg. multi-layer perceptron) will be used to combine the radiological data with other clinical and paraclinical data, to obtain predictive models. Deep recurrent neural networks (e.g. long-short term memory networks, gated recurrent units) will be developed to create sequential models that will aim to predict the transitions over time of patients through different unit types (e.g. emergency room, to ward, to intensive care unit).

Descriptive statistics of the clinical, laboratory and imaging characteristics corresponding to each of the disease phenotypes will be tabulated and analyzed. Learning curves and receiver efficiency function curves will be used to assess the performance of the models. State transition diagrams will be used in order to visualize the variables influencing transfers between different unit types.

APPENDIX 2: CADRE DE GOUVERNANCE

PROPRIÉTÉ

Les données ont été obtenues grâce à une collaboration étroite entre les investigateurs principaux: Michaël Chassé et Louis Mullie au CHUM, Brent Richards au Jewish General Hospital, David Buckeridge au McGill University Health Center, Alexis Turgeon au CHU de Québec Université Laval, Patrick Archambault au CISSS Chaudière-Appalaches, Han Ting Wang à l'Hôpital Maisonneuve-Rosemont, Yiorgos A. Cavayas à l'Hôpital Sacré-Cœur de Montréal, Philippe Jouvet/Michaël Sauthier au Centre Hospitalier Universitaire Sainte-Justine-CHU de Sainte-Justine, Shane English à l'Hôpital d'Ottawa et François Lamontagne au Centre hospitalier universitaire de Sherbrooke. Puisque chaque institution est responsable de l'extraction et de la collecte des données dans son établissement, et que la banque sera stockée de façon distribuée dans chacun des établissements, chaque institution demeure propriétaire de son sous-ensemble de données et peut procéder aux analyses désirées et approuvées par son institution sur son sous-ensemble. L'établissement de rattachement des chercheurs et les chercheurs principaux resteront propriétaires de tout matériel acquis ou produit dans le cadre de cette banque.

Dans les situations où une équipe de recherche désire effectuer des travaux à l'aide de données provenant de plus d'un centre, un résumé du projet devra être soumis au comité de gouvernance qui devra approuver l'utilisation des données multicentriques. La demande devra être détaillée au plan méthodologique et devra planifier les coûts d'extraction et de manipulation des données qui seront associées à l'accès à la banque.

INSTANCES DE GOUVERNANCE ET DE GESTION

Les instances impliquées dans la gouvernance et la gestion de la banque seront les chercheurs principaux de la banque.

Chercheurs

Les chercheurs principaux auront les responsabilités suivantes:

1. Établir des règles de fonctionnement en conformité avec le cadre de gestion des données propre à chaque établissement;
2. S'assurer que les règles et procédures énoncées dans le présent cadre et toute politique ou ligne directrice élaborée par les sites participants soient respectées;
3. Définir les orientations et priorités de développement de CODA-19;
4. Autoriser les accès à la banque pour leur site.

Responsabilités des sites

Le projet CODA-19 est développé dans le cadre d'un partenariat avec les établissements de santé identifiés. Les sites auront les responsabilités suivantes :

1) Soutenir les activités de collecte et d'extraction des données nécessaires pour alimenter la banque;

2) Encadrer au plan légal, éthique et confidentialité la gestion des données pour leur site;

3) Encadrer la gestion des ressources humaines nécessaires à l'extraction des données et au maintien de l'infrastructure de données;

4) Promouvoir l'utilisation des résultats des analyses de données de la banque à des fins d'amélioration des pratiques.

Coordination de la banque

La coordination des activités de CODA-19 sera assurée par un/une coordonnatrice à la recherche désignée par les chercheurs principaux. Elle aura pour fonctions de

1) Gérer les demandes des utilisateurs (recevoir les demandes d'accès; s'assurer que les demandes d'accès sont évaluées selon le cadre de gestion; informer les utilisateurs des décisions relatives à leurs demandes; guider les utilisateurs quant aux procédures d'accès;

2) Assurer les liens entre les sites;

3) Assurer les liaisons avec les comités d'éthique;

4) Veiller au respect de la confidentialité des informations, des données et de la sécurité de la banque;

5) Veiller au respect, par les utilisateurs de la banque, des règles inscrites dans ce cadre de gestion;

6) Assurer un suivi de l'utilisation des données

Financement de CODA-19

Le développement initial de CODA-19 a été assuré grâce à des fonds de recherche discrétionnaires des chercheurs principaux et est partiellement supporté par une bourse de démarrage IVADO (Chassé) et du Réseau de Bio-imagerie du Québec. Cependant, à la fin de la subvention et de la phase de développement, l'exploitation de CODA-19 impliquera des coûts liés au salaire de la coordination de la banque, l'optimisation continue des infrastructures de données, l'analyse des données et le soutien. Pour les chercheurs principaux, ces frais seront imputés à des subventions de recherche en cours ou à venir. Pour les autres utilisateurs, les demandes d'accès aux données et de services sont sujettes à un recouvrement de coûts des opérations. Le projet a obtenu une approbation de financement des Instituts de Recherche en Santé du Canada (IRSC).

CHANGEMENT DE VOCATION, VENTE, FUSION OU TRANSFERT DE LA BANQUE

Les données ne pourront être utilisées à des fins autres que des activités de recherche et d'amélioration de la qualité reliés à la pandémie de COVID19. Toute procédure visant à changer la vocation de la banque devra faire l'objet d'une autorisation formelle des instances suivantes : a) le comité de gouvernance de la banque ; b) les établissements d'attache des chercheurs principaux pour chacune des sous-ensembles de CODA-19 ; c) le comité d'éthique de la recherche du CHUM qui agit en tant que Comité d'Éthique à la Recherche évaluateur.

CODA-19 Data Security Framework

Preliminary

Authors:

Michaël Chassé
Maxime Lavigne
Louis-Antoine Mullie
Bruno Lavoie

Version 0.2.0
September 2020

Table of contents

1. Background

This document outlines the data security policies governing the use of CODA-19, a platform for decentralized, privacy-preserving analysis of data on patients with suspected or confirmed COVID-19 at 8 hospital sites in Québec and 1 in Ontario. The overarching principle of the CODA-19 project is that **no individual patient data will be exchanged between hospital sites.** The objective of this document is to document the procedures that will be deployed in order to protect confidential information across all aspects of the CODA-19 project.

2. Definitions

"CODA-19 network" refers to the entirety of the devices and systems deployed as part of the CODA-19 project.

"CODA-19 computation node" refers to a virtual machine deployed on a physical server at one of the participating sites, which provides computing resources for decentralized analyses, and is accessible through the CODA-19 standard interface.

"CODA-19 storage node" refers to a virtual machine deployed on a physical server at one of the participating sites, which stores de-identified data meeting the inclusion criteria for the cohort.

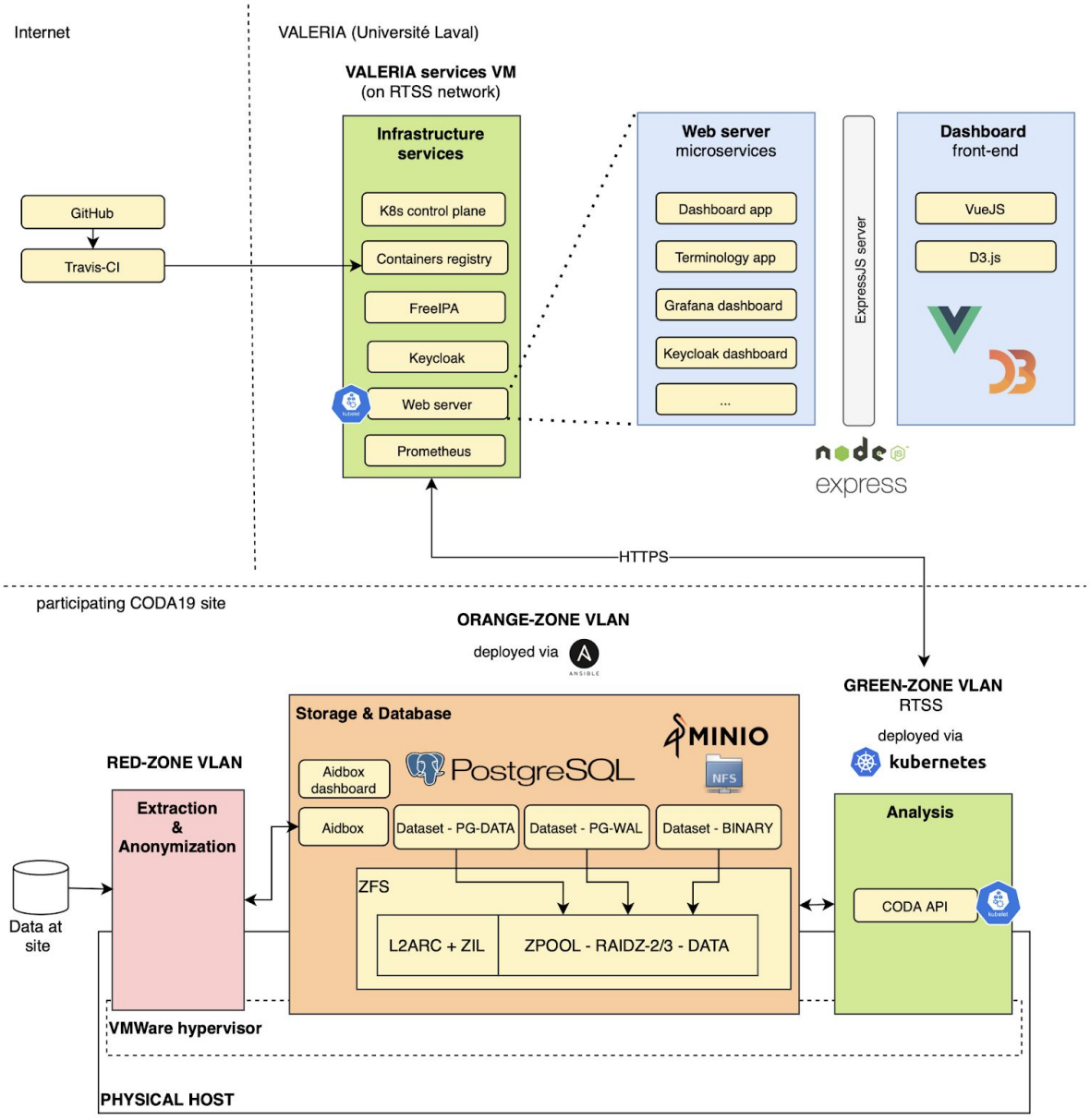
"CODA-19 application server" refers to a virtual machine deployed on a physical server, which communicates with the CODA-19 network and serves the applications that are built as part of the CODA-19 network.

"CODA-19 dashboard" refers to a web-facing application, which is deployed on the CODA-19 application server, and enables authenticated and authorized users to obtain a high-level overview of key statistics that are computed by the CODA-19 network.

"CODA-19 standard format" refers to the standard format used to represent and store the data. In this document, the Fast Healthcare Interoperability Resources (FIHR) format will be assumed.

"Data server" refers to a local data server interface, operating outside the CODA-19 network, which may be used to deploy and/or manage the CODA-19 storage and computation node.

3. Overall architecture



4. Access control

Access control encompasses both authentication (ensuring that only legitimate users are able to access the system) and authorization (ensuring that legitimate users are only able to access the data that they are authorized to view). CODA-19 implements the principle of "least privilege," which restricts users to only the functionality, data and system information that is required to perform their tasks[1]. In order to accomplish this, a central **authentication and authorization server** shall be deployed.

4.1 Authentication component

The authentication component will be responsible for:

- Keeping an up-to-date list of user accounts who will be authorized to access certain CODA-19 applications and other system components;
- Managing creation, update and revocation of authentication credentials associated with the aforementioned users;
- Recording a timestamped log of successful and unsuccessful authentication attempts, and notifying the system administrator of any suspicious login attempts.

End-users will be authenticated via the use of the following combination of credentials:

- An institutional e-mail, associated with a whitelisted e-mail domain or sub-domain at one of the participating institutions (e.g. @chum.ssss.gouv.qc.ca).
- An account password, which will be supplied by the user at the time of account creation. Standard rules for the creation of secure passwords will be enforced. Users will be required to change the password that is provided to them upon account creation.
- A one-time verification code, which will be sent to the user's e-mail or phone for each login attempt.

[1] https://owasp.org/www-pdf-archive/OWASP_SCP_Quick_Reference_Guide_v2.pdf

The creation of user passwords will be according to the NIST digital identity guidelines[2]:

- Minimum of 8 characters and maximum of 128 characters
- Ability to use all special characters, but no special requirement to use them
- Restriction of context-specific passwords (e.g. name of the site)
- Restriction of commonly used passwords and dictionary words
- Restriction of passwords from previously breached corpuses

4.2 Authorization system

A central **authorization system** will be implemented to limit what operations users can perform in the CODA-19 network. This system will be responsible for:

- Defining access roles, and the permissions associated with each role;
- Granting or revoking one or more roles to a user account;
- Recording a timestamped log of any data accessed by a user.

Note that the central authorization system **does not** govern access control outside of the CODA-19 network (e.g. existing servers and storage infrastructures at participating sites). Access control for these external components remains under the local responsibility of each site.

The **roles** and corresponding permissions that will be granted to users within the CODA-19 network are listed in **Table 1.** Roles will be granted on a per-site basis.

The following **types of roles** will be implemented:

- Administrator roles: administrator roles enable the bearer to grant and revoke other roles. This includes the site administrator role, which enables the bearer to grant user roles for a specific participating site, and the master administrator role, which enables the bearer to grant and revoke site administrator roles.
- User roles: user roles enable the bearer to perform specific user operations for a specific site, such as viewing the dashboard for that site. A full list of user roles is provided in Table 1.

[2] https://pages.nist.gov/800-63-FAQ/#q-b05

The following **access permissions** will be implemented:

- Site dashboard access permissions: gives access to the CODA-19 dashboard application, providing a high-level overview of the project's key statistics, but does not enable the user to perform API queries.
- Site basic API access permissions: the user with the site basic API access role will be entitled to perform a restricted, pre-defined set of custom queries to obtain summary statistics on the patient population in the site's data repository (e.g. mean, median, standard deviation at single time points and over time). The full list of basic queries is detailed in **Appendix A**.
- Site advanced API access permissions: the user with the site advanced access role will be entitled to perform distributed and federated analyses running on the site's CODA-19 computation node inside a sandboxed environment with access to a whitelisted set of statistical queries. The full list of advanced queries is detailed in **Appendix B.**

Table 1. Roles, with their associated role type, description, and access permissions

Name	**Role type**	**Description**	**Permissions**
Master administrator	Administrator role	Allows the bearer to grant and revoke site administrator roles, when the data security committee authorizes a new site administrator.	Grant site administrator roles. No access permissions.
Site administrator	Administrator role	Allows the bearer to grant and revoke roles for a given site, following the approval of a project by the CODA-19 executive committee.	Grant user roles within a particular site. No access permissions[3].
Principal investigator	User role	Allows the bearer to access aggregated statistics from all sites in the dashboard.	Dashboard access permissions for all sites.
Site data scientist	User role	Allows the bearer to access the site's section in the dashboard, and grants that user basic and advanced API access permissions.	All access permissions for the site.

[3] The master administrator and site administrator roles have no access permissions. Thus, neither role can act as a single point of failure in the event of an account breach.

Site researcher	User role	Allows the bearer to access the site's data on the dashboard, with additional API permissions depending on the needs of the researcher.	Dashboard access permissions for the site, plus other permissions on a per-request basis.
Health administrator	User role	Allows the bearer to access the site's data on the dashboard.	Dashboard access permissions for the site.
Other end-user	User role	Allows the bearer to access the site's data on the dashboard.	Dashboard access permissions for the site.

Note that one user may have multiple roles, including both an administrator and a user role, and that roles are defined on a per-site basis. For example, a "site lead data scientist" may wish to perform distributed analyses across 3 additional sites. This user already has API access permissions, but is not authorized to query the CODA-19 computation nodes in the other sites. Upon project approval by the governance committee and ethics board, each of the "site administrators" at the 3 additional sites will grant the "site researcher" role to this user. This will enable the user to perform distributed analyses involving these sites.

5. Data de-identification

The CODA-19 project is committed to applying the highest standards for the protection of the patient's privacy and confidentiality. **No personally identifying information shall, under any circumstance, be made available on any component of the CODA-19 network.** This section describes the data de-identification policies that will be enforced in order to ensure that confidential information is handled with the utmost rigor.

5.1 De-identification of tabular data

All personally identifying information (e.g. patient identifiers, address of residence, date of birth) will be removed during data extraction, prior to storage of the standardized data format on the CODA-19 storage node. Elements of a unique type (e.g. a specific laboratory test) with a count less than 5 in the overall output will be censored.

Database row identifiers (non-sensitive identifiers that uniquely identify a database record in local source systems; this excludes public identifiers such as RAMQ number, which are *not* stored in the CODA-19 database) will be hashed using PBKDF2 with 100,000 iterations of

SHA512, using a 128-character hexadecimal salt generated using a secure cryptographic random number generator.

5.2 De-identification of DICOM data

DICOM files will **not** be stored on the CODA-19 storage device. Only a whitelisted subset of technical, non-identifying DICOM headers will be extracted from the DICOM files and stored in the database. Pixel data will be extracted from DICOM imaging studies and stored in HDF5 format. Pixel data will only be extracted only from imaging slices, corresponding to the following modalities in the DICOM standard: CR, CT, US.

5.3 De-identification of textual data

Textual data (such as imaging reports, free text clinical observations or nursing notes) is not currently included in the CODA-19 standard format. If such data is included at a later date, a formal process will be agreed upon to identify the optimal de-identification strategy, and the data security policy will be amended accordingly.

5.4 Site responsibilities and auditing

De-identification of data will be under the ultimate responsibility of the site lead data scientist, who will seek assistance from the data security committee in the event of any questions. In addition, each site will nominate a privacy auditor, who will receive a summary report of each "build" of the site's CODA-19 database. This summary report will contain descriptive statistics for each table and column, and implement automated rules to flag potentially problematic information. The privacy auditor will review each summary report in order to identify any potential anomalies in the data de-identification process.

6. Protection of data at rest

"Data at rest" refers to information that is stored in a permanent or semi-permanent fashion on any component of the CODA-19 network. The protection of data at rest in the context of the CODA-19 project is based on the following principles:

- Data isolation: CODA-19 storage nodes, inside which de-identified data will be stored, will be implemented as virtual machines that are used solely for the purpose of data storage and retrieval.

- Access limitation: CODA-19 storage nodes may only be accessed from a whitelisted list of IP addresses.
- Server hardening: CODA-19 storage nodes will run AppArmor, a Linux application security system that protects the operating system and applications by providing mandatory access control[4].
- Data preservation: data collected as part of the CODA-19 will be preserved for 10 years, unless otherwise specified by an ERB-approved study protocol.

7. Protection of data in transit

No personally identifying information shall, under any circumstance, transit between components of the CODA-19 network. "Data in transit" refers to any information exiting a component of the CODA-19 network over a network interface. Such exchanges may occur either:

1. Between multiple CODA-19 computation nodes, e.g. when a federated learning query is performed, and a trained machine learning model is passed between nodes.
2. Between a computation node in the CODA-19 network and an authorized device in the local area network of one of the participating sites, e.g. when an authorized researcher programmatically submits an API query to the local CODA-19 computation node.
3. Between the application server and the CODA-19 network, e.g. when an authorized researcher performs a distributed analysis via the CODA-19 dashboard.
4. Between the application server and an authorized device in the local area network of one of the participating sites, e.g. when an authorized researcher consults the CODA-19 dashboard from the hospital computer network at one of the participating sites.

All data in transit will be protected using Secure Sockets Layer/Transport Layers Security (SSL/TLS), with 4096-bit RSA keys[5]. Standard SSL/TLS will be used for applications deployed over a web interface (scenario 4), which only display a highly restricted subset of aggregate statistics. Two-way SSL/TLS, also known as "TLS with client certificate authentication," will be used for all other situations of data in transit (scenarios 1-3).

Two-way SSL/TLS involves the use of both a client certificate and a server certificate, such that the server can restrict which clients are allowed to communicate with it. By contrast with a symmetric shared secret key, this provides mutual authentication: the API server must authenticate itself to the API client, and the API client must authenticate itself to the API server.

[4] https://arxiv.org/pdf/1501.02967.pdf

[5] https://tools.ietf.org/html/rfc8446

Each CODA-19 computation node and each user account will be associated with a unique self-signed X.509 SSL/TLS certificate. Users will be able to generate and download their SSL/TLS certificate via the account section of the dashboard. An OSCP server will be deployed in order to manage certificates and revocation.

8. Application security

The CODA-19 application server will be responsible for serving web-facing applications, such as the CODA-19 dashboard. Applications deployed on this server will implement 2-factor authentication, as described in section 3, and be served over SSL/TLS, as described in section 6.

In addition to these measures, CODA-19 applications will be subject to review by the data security committee in order to ensure that they follow guidelines for secure coding practices[6]. This includes, but is not limited to, the following:

- Ensuring that appropriate validation of all user-supplied inputs is performed;
- Ensuring that session management is performed using strong random tokens;
- Ensuring that logout functionality fully terminates the associated session or connection;
- Ensuring that server-side components segregate privileged logic from other code;
- Ensuring that error handling does not disclose sensitive information;
- Ensuring that the SSL/TLS keys are stored in a secure location.

"Scripting" will not be permitted through the dashboard. Researchers who build custom statistical and/or machine learning models using scripts, and wish to train them across sites, will be required submit their applications for code review; such models will be run centrally by the coordinating organization after appropriate oversight has been conducted.

9. Auditing

- **Independent security audit:** this document will be formally reviewed by an independent security consultant, and a security audit will be conducted once the network is in place and multicenter analyses are ready to be performed. Subsequent audits may be necessary after significant changes to the security framework, as deemed relevant by the data security committee.

[6] https://owasp.org/www-pdf-archive/OWASP_SCP_Quick_Reference_Guide_v2.pdf

- **Duty to report:** all members of the CODA-19 project have a duty to report any security anomalies to the data security committee: for example, if a user was given an incorrect role, or if a site found multiple users were using the same credentials.

10. Data security committee

A data security committee will be created in order to coordinate and supervise the implementation of the data security framework described in this document. This committee will contain at least one independent consultant with formal expertise in information security. In addition, each site will nominate a privacy auditor, who will report to the data security committee.

Appendix A. Basic statistical queries

This appendix will be approved by the executive committee prior to deployment, and will be updated over time.

Type of variable	**1-sample operations**	**2-sample operations**
Continuous	Mean, median Standard deviation Interquartile range 95% confidence interval	Mean difference Standardized mean difference
Binary	Count Mode Marginals	Risk ratio Odds ratio Risk difference
Discrete/categorical	Count Mode Marginals	Risk ratio Odds ratio Risk difference

Appendix B. Advanced statistical queries

This appendix will be developed for phase 2 of the CODA-19 project, and approved by the executive committee prior to deployment.