CODA-19 Data Security Framework

Authors:

Michaël Chassé
Maxime Lavigne
Louis-Antoine Mullie
Bruno Lavoie

Version 0.2.2
December 2020

Table of contents

Table of contents 1

1. Background 3

2. Definitions 3

3. Overall architecture 4

4. Access control 4

4.1 Authentication component 5

4.2 Authorization system 6

5. Data de-identification 8

5.1 De-identification of tabular data 8

5.2 De-identification of DICOM data 8

5.3 De-identification of textual data 9

5.4 Site responsibilities and auditing 9

6. Protection of data at rest 9

7. Protection of data in transit 10

8. Application security 10

9. Auditing 11

10. Data security committee 11

Appendix A. Basic statistical queries 13

Appendix B. Advanced statistical queries 14

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 9 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 describe 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.

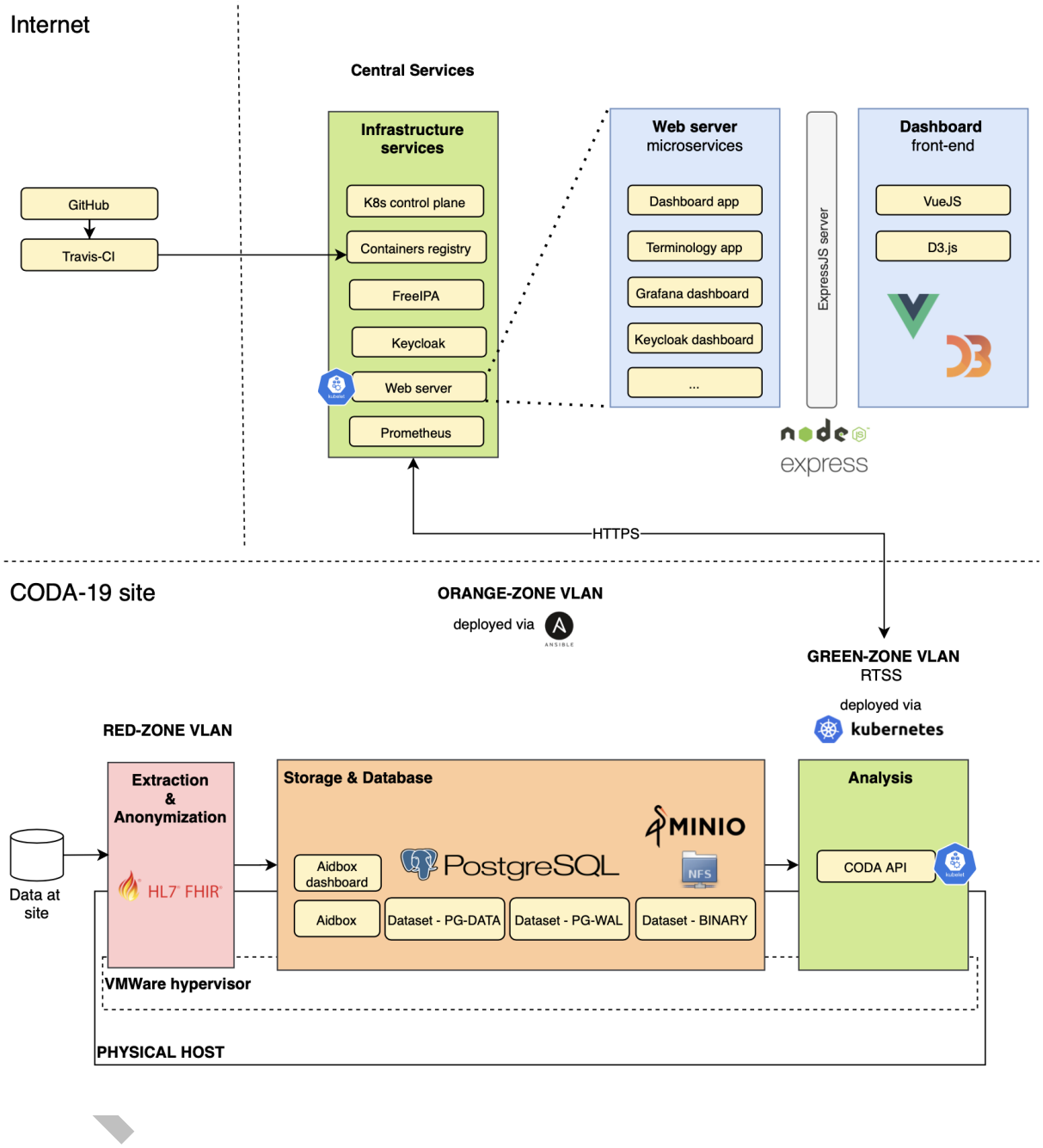
"CODA-19 application server" refers to a virtual machine deployed on a physical server, which communicates with the CODA-19 network and serves the applications that are built as part of the CODA-19 network.

"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.

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)

[1] https://owasp.org/www-pdf-archive/OWASP_SCP_Quick_Reference_Guide_v2.pdf
[2] https://pages.nist.gov/800-63-FAQ/#q-b05

- 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.

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.
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.

[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.

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, day 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.

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

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.

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

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

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 to 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.
- **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

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

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	Count, mean Standard deviation Median, 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.