Algononymous & DeepChest

EPFL Semester Project Report Fall 2020

Authors:	Nicolas Martinod (master of Robotics)
Supervisor:	Mary-Anne Hartley
Affiliation:	Machine Learning and Optimization Laboratory Intelligent Global Health

EPFL

Contents

Part 2: DeepChest

Abstract

Medical data is notoriously sensitive. The necessarily strict laws and ethical standards to protect patient privacy are rendering the ethical sharing of medical data fastidious and can even prevent it. Yet, having a global access to medical data for researchers is key in understanding diseases and finding cures. This lack of sharing is due to legitimate concerns on data privacy both for the patient and for the intellectual property of the doctor but it is greatly impacting the pace of medical research.

The Infectious Disease Data Observatory (IDDO), with whom iGH is collaborating, is working on solving this issue. They curate fragmented data sets and ensure their ethical sharing, by manually anonymizing the data sets for each data request. A non scalable, time-consuming process.

Moreover, traditional practices of anonymization by removing identifiable features from the dataset is not sufficient anymore, as novel machine learning techniques enable complex re-identification attacks. Privacy, therefore, becomes ultimately a trade-off between re-identification risk and information loss.

The aim of this project is to provide IDDO with a personalized tool to analyze the privacy risk created by a new third party medical research. Based on the research needs in terms of data and precision. To then, apply anonymization techniques that would fit our dynamic privacy threshold.

1. INTRODUCTION

Medical data is notoriously sensitive. The (necessarily) strict laws and ethical standards to protect patient privacy have created data silos of poorly interoperable data and fragmented statistical power. A major roadblock to obtaining ethical clearance for data sharing is ensuring that the data is truly private. But what is considered "private"? With the proliferation of machine learning techniques, the traditional practices of manually anonymizing data sets by removing features deemed to be identifiable by human users is no longer sufficient. The definition of privacy also changes over time either due to changing standards or improved abilities to launch re-identification attacks. [1]

Privacy is ultimately a trade-off between what is considered sensitive and the tolerance we have to the utility cost of protecting the data to a certain standard. The problem of optimizing the trade-off between disclosure risk and information loss is known as the **statistical disclosure control (SDC)** problem.

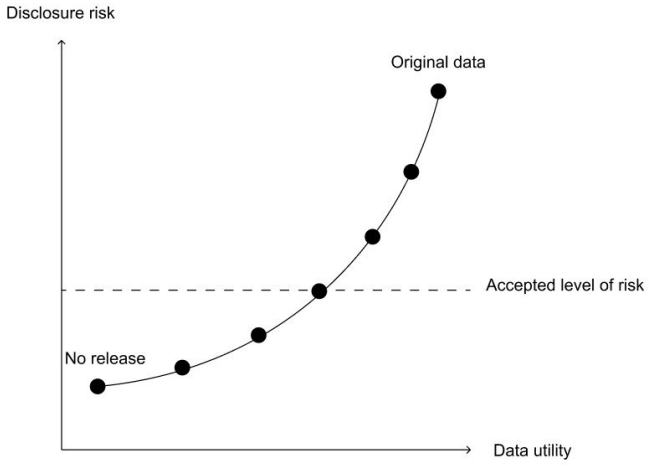


Fig 1: Simplified visualization of the risk–utility trade-off. The goal of the SDC process is to find the optimal point where utility for end users is maximized at an acceptable level of risk.

This project results from the collaboration between iGH (intelligent Global Health), which is affiliated to the MLO (Machine Learning & Optimization) laboratory of EPFL, and IDDO (Infectious Disease Data Observatory).
IDDO "*assembles clinical, laboratory and epidemiological data on a collaborative platform to be shared with the research and humanitarian communities. The data is analysed to generate reliable evidence and innovative resources that enable research-driven responses to the major challenges of emerging and neglected infections." [14]*

Currently, IDDO follows traditional practices of manually anonymizing data sets by removing features deemed to be identifiable by human users. However, these "anonymized" data sets

can still be vulnerable to more complex algorithms that can identify unique patterns in the remaining features able to individualise or even identify a patient.
We aim to implement an anonymization tool as a proof of concept on IDDO's curated Ebola data set, which is the first multi-country repository of data from individual Ebola virus disease patients. [1]

2. AIM & OBJECTIVES

The aim of this project is to provide IDDO with a personalized tool to analyze the privacy risk created by a new third party medical research. Based on the research needs in terms of data and precision. To then, apply anonymization techniques that would fit our dynamic privacy threshold.
This undeveloped tool received the name of Algononymous from the previous project. [1]

Objectives:

1. To review the literature on state of the art privacy techniques and available platforms and describe evaluate and compare their key features, utility vs privacy metrics

2. To identify the appropriate anonymization framework to use for Algononymous

3. To implement the anonymization framework on IDDO dataset

4. To select the development framework for Algononymous

5. To design, with minimal key features and build a Proof-of-Concept platform of Algononymous.

3. LITERATURE

Prior to this project, a state of the art analysis of anonymization method was done by Mahmoud Said [1]. Therefore, this literature review does not intend to go as wide and deep as [1] but will only focus on **SDC (Statistical Disclosure Control)**. This section introduces the basic concepts of SDC methods and the trade-off between disclosure risks and information loss.

3.1 What is Disclosure

Disclosure, also known as “re-identification,” occurs when the intruder reveals previously unknown information about a respondent by using the released data. Three types of disclosure are noted here [2]:

- Identity disclosure, which occurs if the intruder associates a known individual with a released data record.
- Attribute disclosure, which occurs if the intruder is able to determine some new characteristics of an individual based on the information available in the released data.
- Inferential disclosure, which occurs if the intruder is able to determine the value of some characteristic of an individual more accurately with the released data than would otherwise have been possible.

3.2 Classification of variables

SDC methods are applied to variables whose values might lead to re-identification.

- Direct identifiers: reveal directly and unambiguously the identity of the respondent. Examples are names, passport numbers, social identity numbers,... They are removed from the dataset prior to release.
- Quasi identifiers: contain information that, when combined with other quasi-identifiers in the dataset, can lead to re-identification of respondents.
- Non-identifying variables, they cannot be used for re-identification of the respondent. They can either be sensitive (contains confidential information) or non-sensitive.

For the SDC process, it is also useful to further classify the quasi-identifiers into **categorical**, **continuous** and **semicontinuous** variables. This classification is important for determining the appropriate SDC methods for that variable, as well as the validity of risk measures.

3.3 Type of data release

The trade-off between risk and utility in the SDC process depends greatly on who the users are and under what conditions the data are released. There is generally two types of release [3]:

- Public Use File (PUF): the data “are available to anyone agreeing to respect a core set of easy-to-meet conditions.
- Scientific Use File (SUF): the “dissemination is restricted to users who have received authorization to access them after submitting a documented application and signing an agreement governing the data’s use.

3.4 Measuring risk

The risk assessment is based on risk assigned to each and every individual in the record (individual risk) and for the overall database (global risk) based on their uniqueness. For categorical variables, we consider the uniqueness of combinations of values of quasi-identifiers. This concept, however, is not relevant for continuous variables, as it is likely that most individuals will have unique values. Risk measures for continuous variables are a posteriori measures; they are based on comparing the data before and after anonymization.

Measuring uniqueness in the sample: Every single individual in the database is assigned a re-identification risk based on uniqueness in the sample. The sample uniqueness is determined based on the combinations of quasi-identifiers used.

Measuring uniqueness in the population: Obtaining information on how unique a given combination of quasi-identifiers is in the population is difficult. Therefore, population uniqueness is generally based on the size of the population at a given location time.

3.4.1 Risk metrics

To benchmark the risk of a dataset, the standard is to use the following metrics.

- k-anonymity: assuming that sample uniques are more likely to be reidentified, one way to protect confidentiality is to ensure that each distinct pattern of key variables is possessed by at least k records in the sample. A typical practice is to set k = 3. [4]
- l-diversity: for each set of quasi identifiers in a k-anonymized dataset there must be l different values for the sensitive attribute. [6]

SUDA: The previously discussed risk measures depend on identifying key variables for which there may be information available from other sources. In practice, however, it might

not always be possible to conduct an inventory of all available datasets and their variables and thus assess all risks.
To overcome this, an alternative heuristic measure based on special uniques has been developed. This algorithm is called SUDA (Special Uniqueness Detection Algorithm).
In its first step, all unique attribute sets are located. Then, SUDA considers only Minimal Sample Uniques **(MSUs)**, which are unique attribute sets without any unique subsets within a sample. Once all MSUs have been found, a SUDA score is assigned to each record indicating how "risky" it is, using the size and distribution of MSUs within each record.
The final SUDA score for the record is computed by adding the scores for each MSU. In this way, records with more MSUs are assigned a higher SUDA score. [8]

For the sake of the reader, other risk metrics exist such as t-closeness, δ-Disclosure privacy and β-Likeness, but they are considered out of scope for this report as it intends to cover only the basic principle of SDC.

3.5 Anonymization Methods

There are two main kinds of SDC techniques:

1. Non-perturbative techniques, which suppress or reduce the detail without altering the original data

    - **Recoding:** the idea of recoding is to combine several categories into one with a higher frequency count and less information. [11]

    - **Local suppression:** If unique combinations of categorical key variables remain after recoding, local suppression could be applied to the data to achieve k-anonymity. Not all values for all individuals of a certain variable are suppressed, only certain values for a particular variable. [11]

2. Perturbative techniques, which distort the original micro-dataset before release

    - **Post-Randomization Method PRAM:** it is a probabilistic method that reclassifies the values of variables, such that if the intruder is able to match individuals between external files and the released data files, he cannot be sure whether these matches are to the correct individual.
      PRAM is defined by the transition matrix P, which specifies the transition probabilities, i.e., the probability that a value of a certain variable stays unchanged or is changed to any of the other $k-1$ values.
      Given that the transition matrix is known to the end users, it is possible to correct statistical analysis of the data for the distortions introduced by PRAM. [10]

    - **Micro aggregation:** The method first partitions records into groups, then assigns an aggregate value (typically the mean) to each variable in the group.

To preserve the multivariate structure of the data, the most challenging part of micro-aggregation is to group records by how “similar” they are. [12]

- **Noise addition:** The idea is to add or multiply a stochastic or randomized number to the original values to protect data from exact matching with external files. [3]

- **Shuffling:** it generates new values for selected sensitive variables based on the conditional density of sensitive variables given nonsensitive variables. It uses an underlying regression model for the variables to determine which variables are swapped. [9]

3.6 Measuring Information Loss

After SDC methods have been applied to the original dataset, it is critical to measure the resulting information loss. There are two complementary approaches to assessing information loss: (1) direct measures of distances between the original and perturbed data, and (2) the benchmarking approach comparing statistics computed on the original and perturbed data.

3.6.1 Direct measures

- **IL1s information loss measure**: it is the scaled distances between original and perturbed values.
  Let $X = \{x_{ij}\}$ be the original dataset, and $X' = \{x'_{ij}\}$ is a perturbed version of X. Suppose both datasets consist of n records and p variables each. The measure of information loss is defined by

$$IL1s = \frac{1}{pn} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{|x_{ij} - x'_{ij}|}{\sqrt{2} S_j}$$

  where S_j is the standard deviation of the j-th variable in the original dataset. [13]

- **Eigenvalues:** A second measure is the relative absolute differences between eigenvalues of the co-variances from standardized original and perturbed values of continuous key variables [11].

- **Lm:** A third measure is the differences between estimates obtained from fitting a pre-specified regression model on the original data and the perturbed data:

$$lm = \left| \frac{\bar{\hat{y}}_w - \bar{\hat{y}}'_w}{\bar{\hat{y}}_w} \right|$$

where $\overline{\widehat{y}}$ denotes the estimated values using the original data, $\overline{\widehat{y}'}$ the estimated values using the perturbed data. [11]

3.6.2 Benchmarking indicators

The idea behind this approach is fairly simple as it consists in choosing a set of indicators and to compare their value on the original data vs the anonymized data. Examples of typical benchmarking indicators are listed below:

- Number of record change: it gives a good indication of the impact of the anonymization methods on the data.
- Use of contingency tables, which displays the frequency distribution of variables. To maintain the analytical validity of a dataset, the contingency tables should stay approximately the same.
- Basic statistical indicators: mean, covariance, correlation. This statistics should not change after the SDC process.
- Histograms and density plots are also useful for quick comparisons of variable distribution before and after anonymization.

3.7 Re-identification attacks

In the context of applying SDC on a dataset, it is important to identify the most probable type of re-identification attacks:

- Linkage attacks: these are attacks where the adversary uses a secondary data set to learn their identity or sensitive attributes. [15].
- Graph/node attack: in which the data graph nature is used to de-anonymize data (nodes) belonging to a person. Primary examples involve social network data where an individual's inter-node connections are uniquely identifiable [16].
- GIS/spatial attack: this type of attack occurs when we have spatial data in our dataset such as address. [17]

Attacks can also either target specific individuals, where a user has the explicit identifiers of someone in the data set and tries to connect it to the dataset, or it can be a generalized attack where many people are being re-identified at once

3.8 Acceptable risk threshold

IDDO is currently applying the following guideline to determine the risk threshold: [18]

The European Medicines Agency (EMA) advises a conservative risk threshold of 0.09 for the public disclosure of the clinical reports [5].The EMA further suggests that the most appropriate way to measure the risk of re-identification for an entire dataset, in the context of public disclosure, is through the maximum risk, which corresponds to the maximum probability of re-identification across all records [5]. The usage of maximum risk assumes that the adversary is attempting to re-identify a single person in the data set, and that the adversary will look for the target record that has the highest probability of being re-identified. For release of scientific use files, the risk is deemed to be lower than the publicly released files and there are precedents of usage of higher maximum risk threshold (See Figure 2-3 of [6]). A maximum individual risk threshold is considered conservative of sharing scientific use files (private sharing) and an average risk approach might be more useful [7].

3.9 Existing SDC tools

A review of all relevant SDC tools has been conducted:

		UTD-AT	CAT	Amnesia	ARX	µArgus	Parat	SDCmicro
Developer support	**Open source**	Yes	Yes	Yes	Yes	No	No (commercial)	Yes
	Active	No	No	Yes	Yes	No	Yes	Yes
	Public API	No	No	No	Yes	No	Limited info access due to its commercial nature	Yes
	Cross-platform	Yes	Yes	Yes	Yes	No		Yes
	Prog-language	Java	C++	Java	Java	C++		R
	Release date	Feb 2010	June 2009	Dec 2018	May 2016	Dec 2012		Sep 2015
Usability	**Graphical User Interface**	None	Full	Full	Full	Full	Limited info access due to its commercial nature	Full
	Hierachy creation	No	No	Yes	Yes	No		Yes
	Visualization	No	Data, risk	Data, risk and solution	Data, risk and solution	Risk		Data, risk and solution
	Standalone	No	Yes	Yes	Yes	Yes		No
Anonymity method	**Automatic solution**	Yes	Yes	Yes	Yes	No	Limited info access due to its commercial nature	Yes
	Privacy criteria	k, ℓ, t	ℓ, t	k	k, ℓ, t, δ	None		k, ℓ, SUDA
	Generalization	Yes	Yes	Yes	Yes	Yes		Yes
	Risk assessment	No	Limited	Yes	Yes	Limited		Yes
	Descriptive statistic	Basic	Basic	Basic+	Full	Basic		Full
	Anonymization method	6	2	2	7	2		8

Fig 2: Comparison table of existing SDC solution

- PARAT is the leading commercial de-identification software but only limited information is available to the public. [23]
- In our context, we are focusing on non-commercial tools. Such as the UTD Anonymization Toolbox [19] and the Cornell Anonymization Toolkit (CAT) [20], which are research prototypes that have mainly been developed for demonstration purposes. Problems with these tools include scalability issues when handling large datasets, complex IT configuration, and incomplete support of methods of data transformation.
- µ-Argus24 is a closed-source application that implements a broad spectrum of techniques, but it is no longer under active development. [22]

- Amnesia has its background at the Athena Research Center, it is in active development but as of right now, only has limited privacy criteria and anonymization methods implemented. [21]
- sdcMicro is a package for the R statistics software, which implements many features required for data anonymization and has a very active developer support. [25]
- ARX is a privacy software developed by TUM (Munich), it is a desktop app with active development and many privacy features. [24]

4. METHODS

To be able to provide IDDO with a personalized tool (Algononymous) to analyze the privacy risk created by a new third party medical research and to anonymize the dataset accordingly, we first need to understand what procedure is following IDDO in that context.

4.1 The IDDO process

IDDO is a global medical data sharing platform called the Infectious Disease Data Observatory. IDDO curates fragmented data sets and ensures ethical sharing from a privacy point of view.

Curation process: When they receive a new dataset, it first goes into a standardized curation process. The data managers (such as Sam Strudwick) are in charge of taking the raw data and to curate it through a commercial platform called Tamr. They use a general standard for data curation (unknown name but it has 15 files, 10 on variables about the patients, 5 on trial data)
Then, as a final step, their statisticians (such as Prabin Dahal) will manually remove features deemed
to directly identify patients (names, dates of birth, telephone numbers and addresses etc.)

Dataset request: When a researcher requests a specific dataset for a study they first have to make a formal request by explaining the goal of their research, what they need in terms of data, what they will do with it and who will have access to it. Then the data managers of IDDO will analyze it. If the request is accepted, it is sent to the statisticians who will gather the selected data from the request and apply SDC to meet the acceptable risk threshold. Once anonymized the data managers judge if the anonymization is sufficient and if the information loss does not undermine the potential quality of research too much. This process is therefore iterative between the data managers and the statisticians, and is repeated for every data request.

What takes time in this process is not only the SDC process in itself but also the preparation of the file, i.e: selection of the right sample, merger of the curated dataset into a single file and selection of key variables.

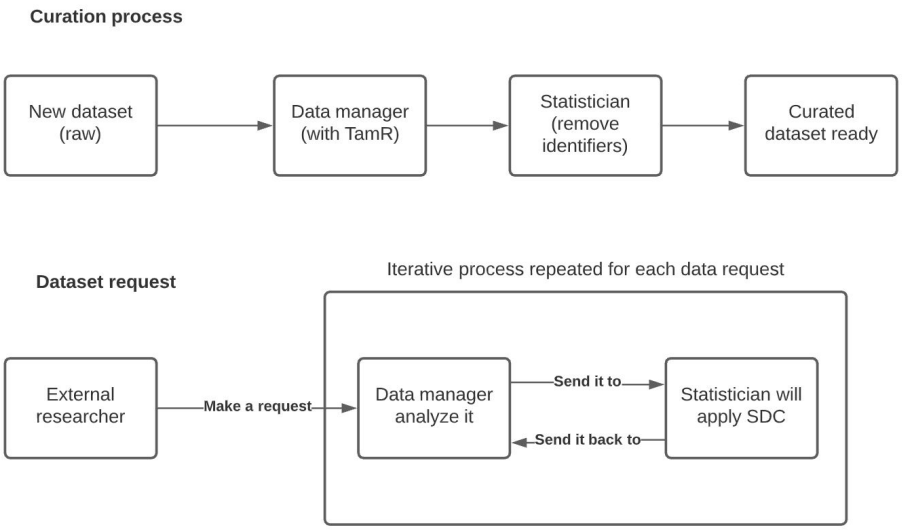


Diagram summarizing IDDO process for data curation and dataset request

4.2 Algononymous features

The platform should have the following features implemented:

- Visualization of the dataset and repartition of key variables
- SDC risk metrics: k-anonymity and l-diversity
- Visualization of the risk metrics and information loss
- Descriptive statistic for information loss quantification
- Anonymization method: recoding and suppression as minimal features. Then PRAM, noise addition, micro aggregation and shuffling as additional ones.
- Export the altered data set with an updatable privacy report

Overall it is not intended for people with programming skills and should therefore be user-friendly and as automatic as possible, while still staying flexible.

We want to benefit from the active open source community developing the SDC framework we will use. As the privacy definition always evolves as well as re-identification techniques, we want to be able to update the platform with the latest version of the used SDC framework.

4.3 SDC framework

The choice of framework for the SDC is the most restrictive as there are not many, still active, frameworks that exist. From the literature review of SDC tools, two options retained our attention:

- **ARX:**

The platform is open-source and developed in Java, which is good for app development. Nonetheless, it is a desktop app only and it was not designed as an open library one can use in their own script. Therefore, we could use the ARX platform to anonymize the dataset, but adapting the platform to our needs would have turned the project into an advanced java app development, where delivering a Proof-of-concept platform would have taken a lot of time. Also ARX capabilities extend further than what we need, and lack user-friendliness, making it not very intuitive when anonymizing the dataset.

- **SDCmicro:**

SDCMicro is a free, R-based open-source package for the generation of protected microdata for researchers and public use, developed by the Vienna University of Technology. It is under active development and is widely used in statistics offices in the European Union [11], but the user needs some knowledge of R to use sdcMicro
This framework is already used by IDDO and R is their main programming language.
As it is a package, it can be used in our own scripts, and therefore, can allow a quick implementation for a proof-of-concept tool.

It is for the above reasons that we decided to use SDCmicro as the main SDC framework for Algononymous.

4.4 App development framework

We intend to develop a simple proof-of-concept app, but we have to keep in mind that it will be further developed depending on our result. Thus it is important to choose wisely the app development framework now. Something too general can lead to time inefficiency of development, while too specific can get us in a dead-end later-on.

For user-friendliness reasons we first decided to move toward a web-based app rather than a desktop one.

When developing a web app, one needs to consider framework for both front-end (user side) and back-end (server side):

The front-end is composed of:

- HTML, which contains all the content of your webpage
- CSS: to style your HTML and make it look fancy
- Javascript: to make your website dynamic

While the back-end have 3 components:

- Server: where all the website files, the database, and other components are stored.

- Database (most of the time SQL is used)
- Server side programming: To write the function and logic of your app. Several languages are possible, the most used is Java but Python has a growing popularity with the Django and Flask framework

In our situation, going for one of these classical web frameworks poses a major interoperability problem between the back-end language and the R framework SDCmicro.

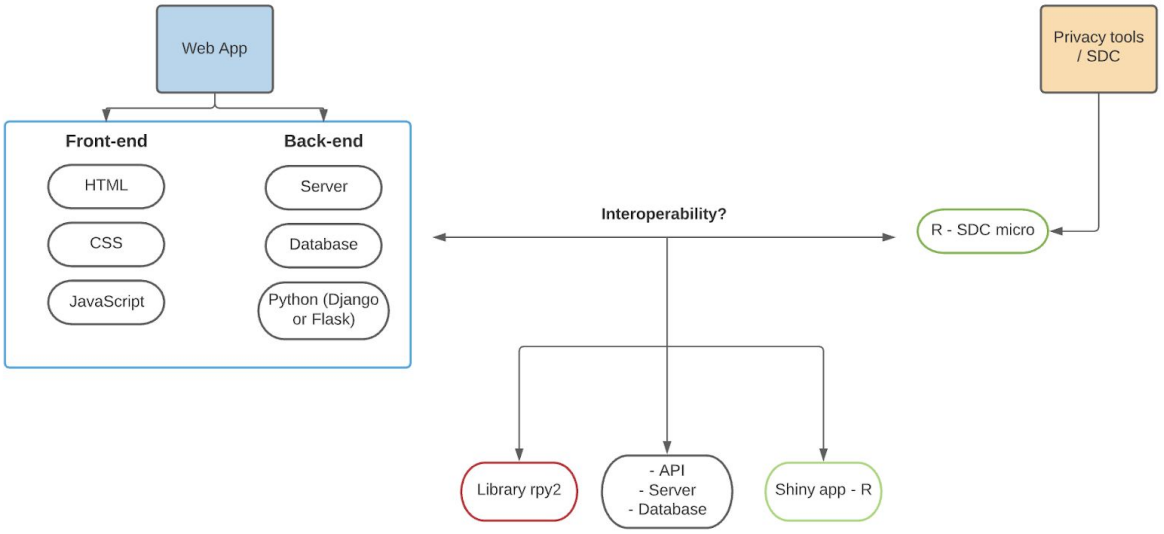


Fig 3: Illustration of the web app framework vs SDC tool issue

To solve this we found three possible scenarios

1. **Rpy2:** it is a R to Python interface which allows users to conveniently call R functions and methods from python, as well as exploiting existing R modules
   Yet, it is not very mature, nor user-friendly. It is complexe to adapt for many different functions so it is not a reliable long term solution in any way.

2. **API:** One can develop an R API in order to run R script on a server and access the result through the API. However you need to set up the API first, as well as the server and database, which is a waste of time for a proof-of concept app.

3. **Shiny App:** It is a low code tool in the form of a R package that allows the development of interactive web apps straight from R. Both the back-end and front-end is done through R and therefore we would have no interoperability problem with SDCmicro. It is an actively used framework by R programmers and can then be deployed on any cloud application platform, such as Heroku, which is free, and removes the need of setting up a server and a database for the back-end. [26]

Based on those scenarios we decided to move on with option 3 and to develop Algononymous on a Shiny web app.

5. RESULTS

5.1 The SDC process

Based on the SDC process and the context of IDDO, we have identified the following workflow when anonymizing a dataset:

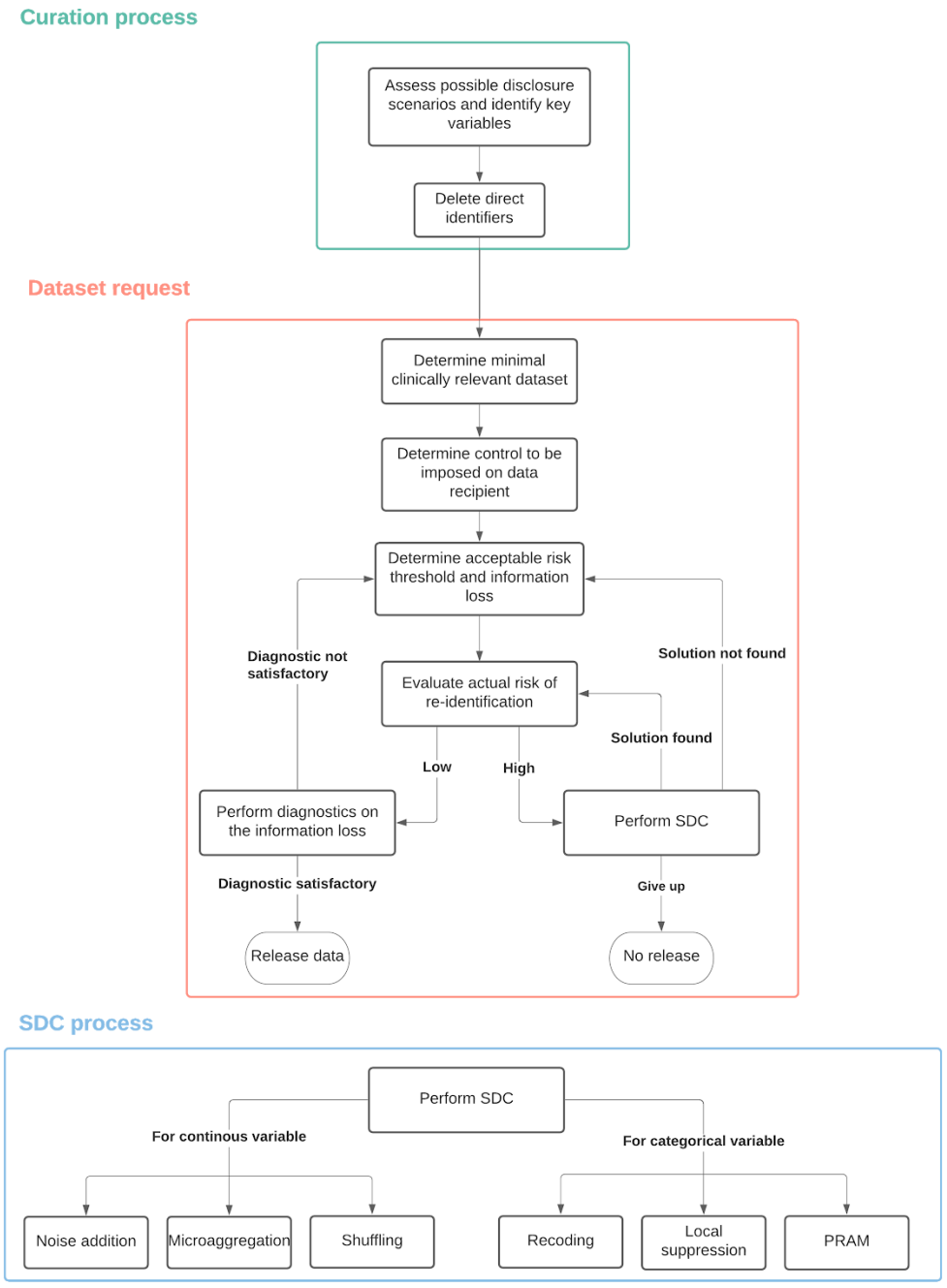


Fig 4: A guideline workflow when applying SDC

The selection of key variables is very subjective as any publicly available information can be a key variable, but if too many are selected, doing SDC will not make sense as there will always be uniqueness. Therefore they need to be hand-picked based on their apparent riskiness. Determining the minimal clinically relevant dataset depends on the request issued by the researcher. The level of control imposed to the data recipient establishes the condition to access the dataset and the way it can be handled (shared, published, used,...).
The risk threshold, risk and loss evaluation as well as the SDC process were explained in the literature section previously.

5.2 SDCmicro and Ebola dataset

The sdcMicro package is built around objects of class *sdcMicroObj.* You create an object based on a dataset to then specify its variables and their types. From the object, one can then access its privacy report and then apply anonymization algorithms. A detailed tutorial and guideline about SDCmicro can be found in [3]
The SDC analysis on the ebola dataset was covered by IDDO in [18]. A summary of their findings can be found below:

	Model
Details	Continuous key: Age Categorical key: COUNTRY, SEX, RFSTDTC, DTHDTC, PREFECTURE, DTHFL, DSDECOD, DSSTDY Weight variable: sample weight
Risk measures	
Violated: 2- anonymity	3078 (18.7%)
Violated: 3-anonymity	7670 (46.6%)
Violated: 5-anonymity	11280 (68.5%)
Maximum risk	0.011
Average risk	0.0014
Expected number of re-identifications	24.06 (0.15 %)
Maximum risk > 0.09	0 (0%)

Re-identification risk measures of Ebola dataset

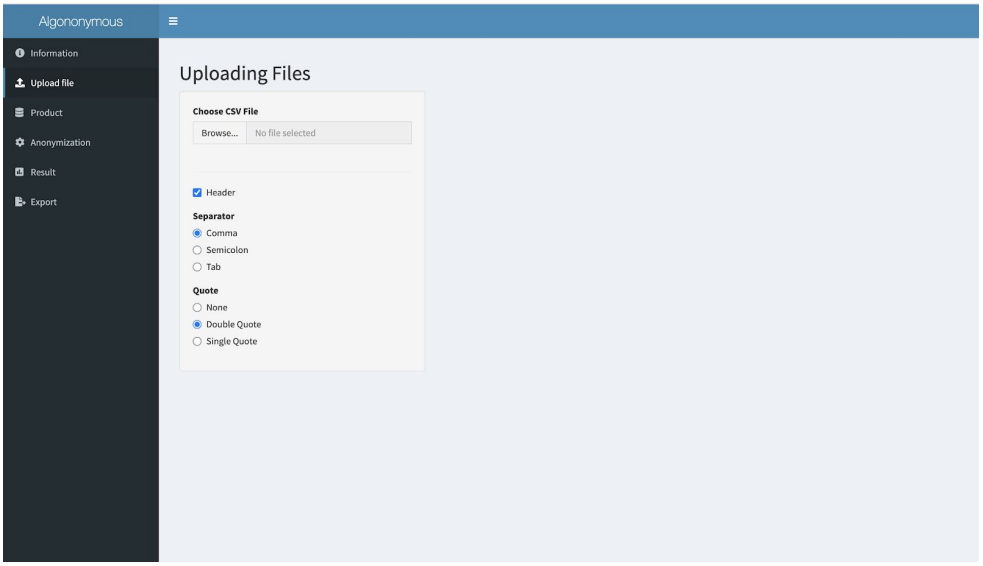
Contrary to a large proportion violating k-anonymity measures, the individual re-identification risks are very small. The maximum estimated risk for both models were within the EMA recommended threshold of 0.09 [18].
The code presenting the basics of SDCmicro is provided in the Github.

5.3 The premise of Algononymous

We selected the Shiny dashboard template to build our proof-of-concept Algononymous. The initial idea was to offer a no-code version of the SDCmicro package, which would implement the basic features we have listed in 5.3 and would guide the user through the SDC workflow outlined in 5.1

The Algononymous dashboard would have a tab for each SDC step

- 1st tab: to download the raw data,
- 2nd tab: to visualize the dataset, some key indicators, the selection of key variables and to display the privacy report with k-anonymity, l-diversity and SUDA score.
- 3rd tab: dedicated to the anonymization process
- 4th tab: to visualize the result of the anonymization, i.e the information loss and updated privacy report
- 5th tab: to export the anonymized dataset



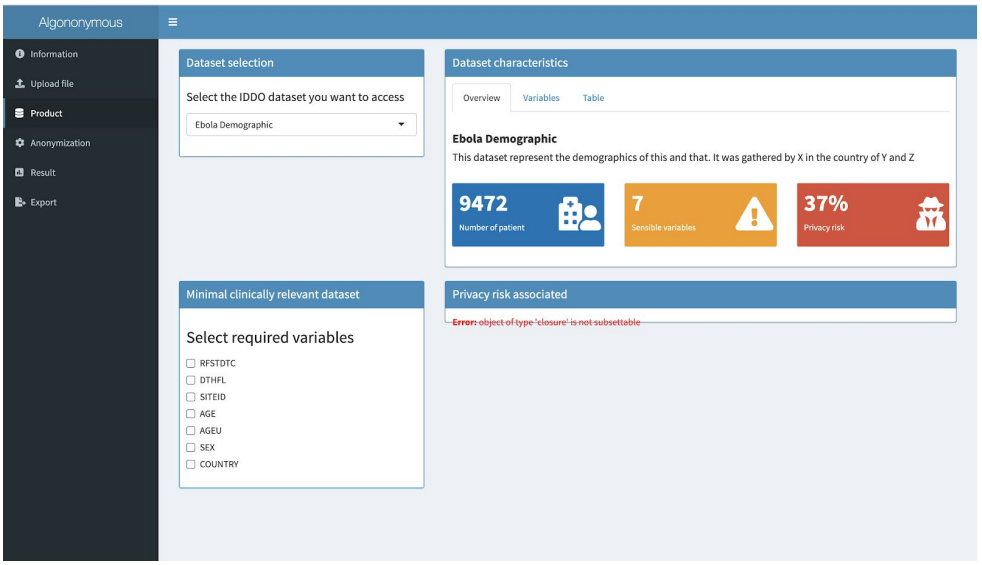


Fig 5: Draft of Algononymous build with Shiny. Tab Upload and tab Product are shown

5.4 SDCApp

On week 9 of this project, we discovered the sdcApp(). It is a shiny web app application of the SDCmicro package. It is fairly user friendly and it delivers a complete no-code access to all the SDCmicro features, which is, to some extent, what we had in mind when starting the development of Algononymous. It is more advanced than the proof-of-concept we initially wanted to develop but its implementation and process follow almost what we drafted previously.

This web app was released 3 years after the initial SDCmicro package and is not yet very popular hence maybe why we did not find it in our initial SDC tool overview.

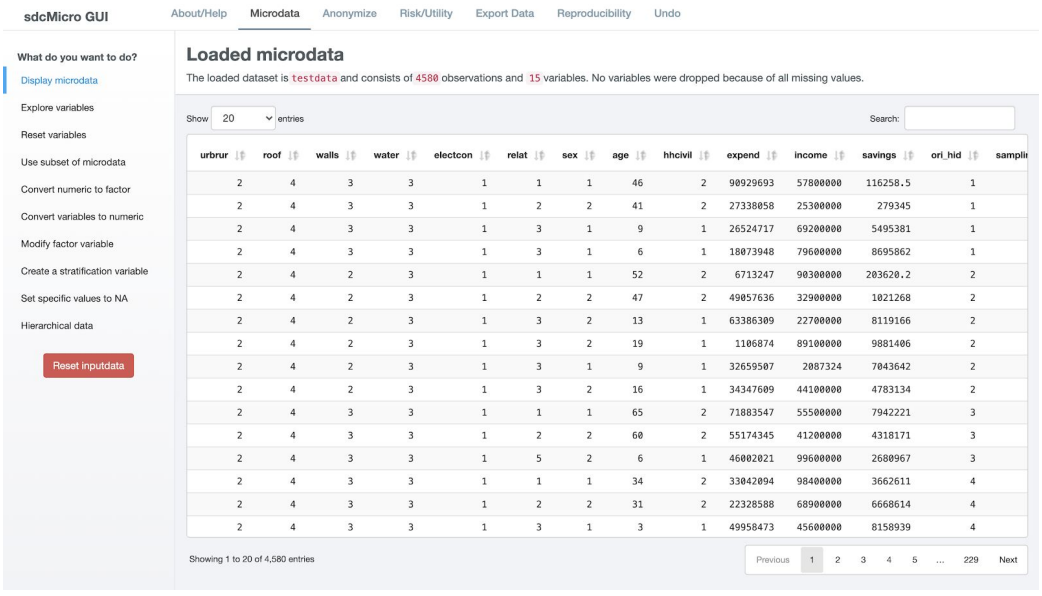


Fig 6: SDCmicro web based app

5.5 The shift to a more urgent project

With the discovery of SDCApp, which implements Algononymous' main feature and can therefore serve as a proof-of-concept in the scope of this project, we faced a decision.

1. Either to continue in the same direction, taking into account SDCapp, testing it fully and delivering the result of its implementation on the Ebola dataset.
2. Or to start a new project by taking up on an urgent task for COVID and to work on the project called DeepChest.

We decided to go with option 2, this work is covered in the part 2 of this report, on DeepChest.

6. DISCUSSION

6.1 Limitations

The SDCmicro app is a web app but not really standalone as it lives locally on your computer and you need to open it through R. It has more features than what IDDO directly needs and it is not tailored for IDDO data sharing process so it lacks user-friendliness as using the platform for someone not familiar with SDC, would require a bit of training.
SDCmicro and SDCapp are both developed entirely with the R language, even if it is a popular language it is not aimed for web app development and therefore lower the flexibility of the web framework Shiny.

6.2 Future work

Initially the objective of this project was to provide IDDO with a proof-of-concept tool for them to anonymize medical dataset with only a limited loss in information. This, in order to speed up the data sharing process in the scope of external scientific research.
But it appears IDDO are already able to apply SDC to a medical dataset to achieve a defined risk threshold using SDCmicro.

Future work for Algononymous will then depend on which direction we decide to go:

1. Either further assist IDDO in the data sharing task by trying to tackle the problems specific to their internal process. Which means not only to focus on the application of SDC but also on the other time consuming tasks they have in their pipeline, i.e: the back and forth analysis between data managers and statisticians for each and every request, the merger of dataset and the selection of sample according to the data request (see 4.1).

2. Or focus on the SDC tool, and try to solve more general issues with medical data sharing. Going down this path, would require to better understand the issue of medical dataset sharing:

    - What procedure are they following?
    - Are people aware of the SDC tool that already exists, i.e: SDCmicro and ARX?
    - If they are using it, does it solve their problems?
    - If not, then why: is it too complexe, or too demanding in pre-processing, or there is no concrete guideline regarding risk threshold from governments, or no satisfactory trade off between anonymization risk and information loss?

Depending of the previous decision, we have three different scenarios:

1. ***Integration:*** we just fully integrate a SDC tool in the process (either SDCmicro or ARX),
   Pros: we embrace an actively developed framework that will evolve with the updated definition of privacy and re-identification attacks. No work to be done regarding development, only training people to use the platform.
   Cons: the solution is not tailor-made to IDDO's specific needs, and therefore less user-friendly. We also have to accept the possible lack of features and absence of flexibility.

2. ***SDC platform:*** We fully develop a new SDC platform to better tackle the data sharing problem that the already existing tool cannot manage. It could either be the lack of user-friendliness, complexity of usage, lack of advance anonymization algorithm, a better assistance in the curation and pre-processing pipeline, etc.
   This platform can be built upon the active open source SDC framework, such as SDCmicro or ARX.
   Pros: We provide a tailor-made solution which aims at solving medical data sharing problems for multiple organizations.
   Cons: Even if we can build upon one of the open source frameworks, if we fork it, then it will be difficult to benefit from their later updates. This is a more ambitious solution which requires a lot of time and manpower.

3. **Hybrid:** We use one of the existing SDC tools (SDCmicro or ARX) but try to better integrate it in the IDDO process. It means to build a platform that specifically follows the IDDO process while trying to fully integrate the SDC tool. Therefore, we try to develop something more user-friendly around the already existing SDC tool. The objective being then for the data manager to handle the SDC process by themselves.
   Pros: we benefit from the active framework and the future updates.
   Cons: we have less development than in option 2 but it will still take time to identify the right solution and to implement it.

6.3 Conclusion

The Statistical Disclosure Control (SDC) in the scope of medical data sharing is far from being mature. Official guidelines regarding procedure to follow and risk threshold to apply depending on a specific sharing scenario do not really exist yet, or at least, are unclear.
However, SDC open source tools are being developed and could offer a solution for medical data owners looking to share them for scientific reasons.

DeepChest: A deep learning powered mobile application to assess COVID-19 pathology from lung ultrasound images

1. CONTEXT

As explained in 5.5, having fulfilled most of the objectives set for the original project on Algononymous on week 9, we shifted the project toward a more urgent task to work on a new project called DeepChest.

This part of the report does not intend to have the same scientific report structure as in Part 1 but rather more of a Readme of was built during the 4 last weeks of the semester.

1.1. DeepChest

This abstract is taken from Hugo Schmutz master thesis, who started DeepChest [27]

> **DeepChest:** A neural attention model for interpretable, missingness-resilient diagnosis and risk stratification of COVID-19 from lung ultrasound images
>
> **BACKGROUND.** On September 1st, 2020, 25, 251, 334 cases of COVID-19 have been confirmed worldwide with 840, 000 deaths. The rapidity of transmission quickly saturated response capacity, leading to global stock outs, and slow turn-around times of gold standard diagnostic kits, forcing care workers to rely on clinical skills to make triage decisions, however such expertise was spread thin in decentralised screening and triage units. Lung ultrasound (LUS) is a valuable non-invasive clinical skill to rapidly and inexpensively assess respiratory pathology at point-of-care. However, its interpretation is user-dependent and lacks standardisation. Recently, deep learning methods have shown promise to better standardise evaluations and guide decision making but currently available models are derived from uncontrolled fragmented cohorts that are not representative of triage populations and also rely on an image-by-image assessment that fails to robust represent patient anatomy, especially in the very frequent situation of image acquisition of variable quantity.
>
> **METHODS**. We apply a novel attention-based model on a dataset of 2,342 LUS images acquired from a cohort of 162 adult patients attending a COVID-19 triage unit at the University of Lausanne Hospital in Switzerland between 6 February 2020 and 5 August 2020. We discriminate between two outcome groups (diagnostic: COVID positive vs negative by gold standard RT-PCR; and prognostic: ambulatory vs hospitalised with oxygen support at day 7). The model is a

composite feed forward neural network extracting representations via a Resnet, and including positional encodings as an intuitive way to create distances in the latent representation space between the embedding vectors of images from different positions. In order to aggregate classifications at the patient level, we leverage a BERT model using a multi-head self-attention module and a feed forward network before a 3-layer neural network classifier. Thus, the network predicts either the diagnosis or risk stratification at the patient level from a variable amount of ultrasound images.

FINDINGS AND INTERPRETATION. For discriminating COVID-19 diagnosis, we achieve 98.62% area under the receiver-operator curve (AUCROC) and 93.13% of accuracy using the attention-based model. Discriminating prognosis achieves 77.28% AUCROC with 70.67% of accuracy. The model reached state-of-the-art models with a better potential for interpretation and a stronger robustness to missing data.

1.2. DeepChest app

The aim of DeepChest is to deliver a tool to clinicians to help them analyse lung ultrasound (LUS) images and make diagnostic and prognostic predictions. While LUS could be used to predict many kinds of lung pathologies, this first effort focuses on COVID-19

The end product will consist of a phone app which guides the clinicians through the acquisition and uploading of LUS images, to then analyze them and provide a prediction. This tool aims, for now, to inform clinicians only and not to drive the diagnosis or even to make it. Therefore it falls in the category of a class IIa medical device.

The DeepChest app project can be broken down into three work packages:

- Mobile app
- Image reprocessing
- DeepChest model

2. RESULT

2.1. Preprocessing

DeepChest takes as input raw LUS images and then apply three preprocessing steps:

- Remove duplicate images
- Crop polluting information from the images
- Reshape convex images to rectangular images

However the issue with the real world application of DeepChest is the access to the LUS images in the first place. Todate, there is no universal API or standard protocol to retrieve the raw data of the LUS image directly from the cloud or the ultrasound machine. Indeed, we can not know in advance the type of ultrasound probe the clinician will use and many older versions have no digital export system. Which means that we have no easy way to access the raw data and transfer it to DeepChest.

To solve this, we thought of asking the clinician to take a photo of the LUS images with their phone from whatever ultrasound device they are using, which could then be directly processed in the DeepChest App

2.1.1. Image scanner

We need to extract the LUS image from the photo taken by the clinician and to restore its original format so it can be fed into DeepChest (i.e. removing background objects outside of the image).
To this end, we used open cv2 to extract the image and applied a Gaussian blur filter to improve the contrast and then a canny filter to find the edges. We then look for all rectangles among our edges, order them by size and then check if there is a LUS image among them (see 2.1.2 for explanation).
A four point transform is applied to obtain a top-down view of this LUS image.

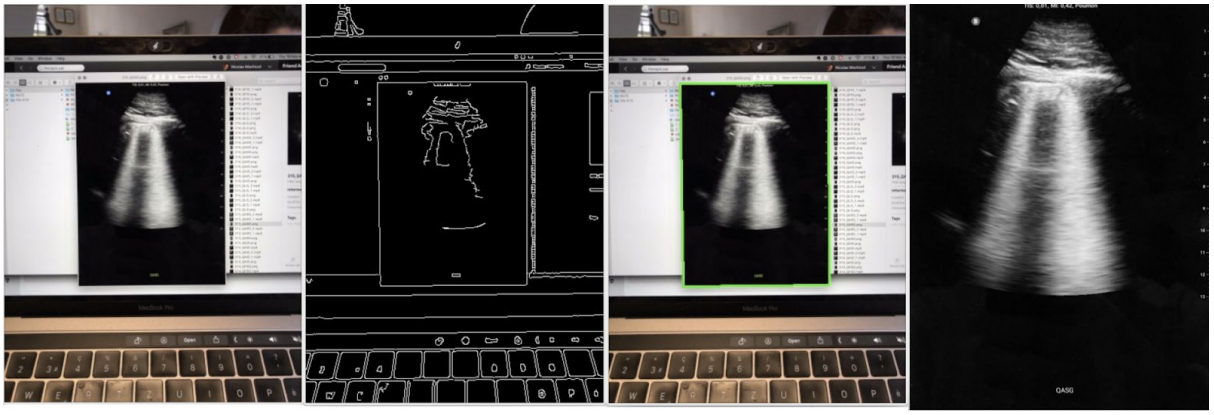


Fig 7: Pictures of the different steps in the image scanner. From a photo of a LUS image on a computer screen we can extract the LUS image perfectly

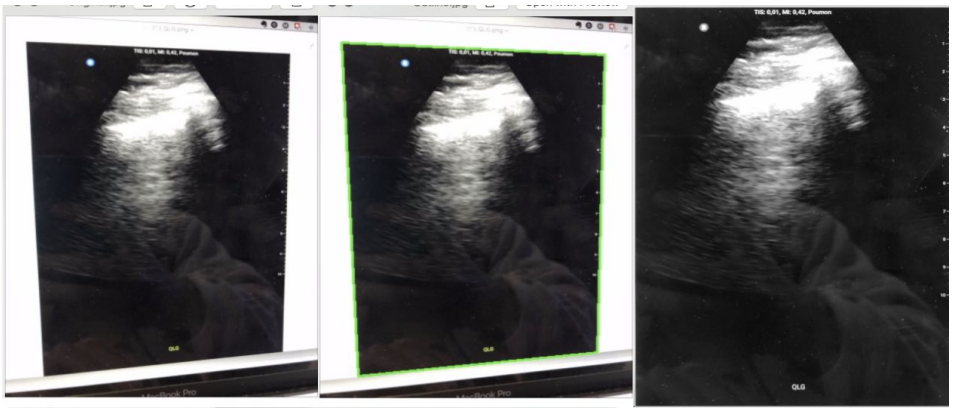


Fig 8: Illustration of the image scanner with a photo taken from the side and corrected with four-point transformation. Noticeable contamination from reflections remain.

Here, we can see the four point transform succeeds in reorienting the image to fill a rectangular space without deformation. However, visible reflections on the image remain..

To better ensure that the selected rectangle is indeed the one containing the LUS image (as we can see below). A CNN classifier was trained as described in the next section.

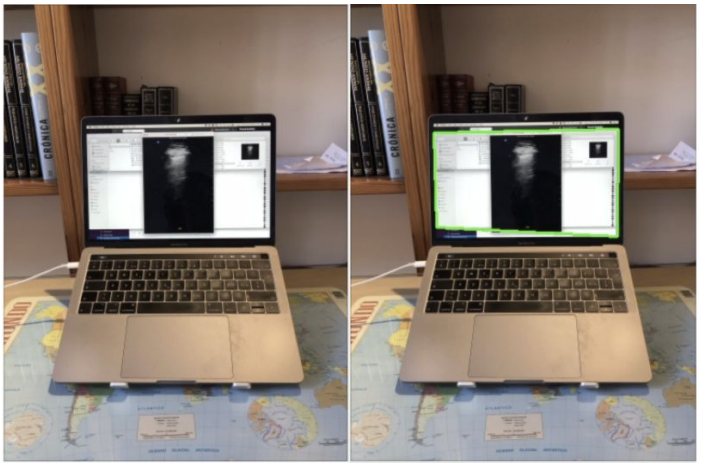


Fig 9: The image scanner took the wrong rectangle

2.1.2. CNN classifier

To solve the issue of selecting the incorrect rectangle we implemented a CNN classifier to check if the image selected is indeed a LUS image. A basic CNN was designed using Keras framework within tensorflow. The dataset to train and test the CNN was built by taking photos of the LUS image from the DeepChest database displayed on a laptop screen with background noise vs the same photo but with the screen displaying random web pages without a LUS image visible [28]. The random webpage was chosen because it is most likely that if the image scanner takes the wrong rectangle, it would be a computer screen.

The dataset folder needs to be organized as follow:

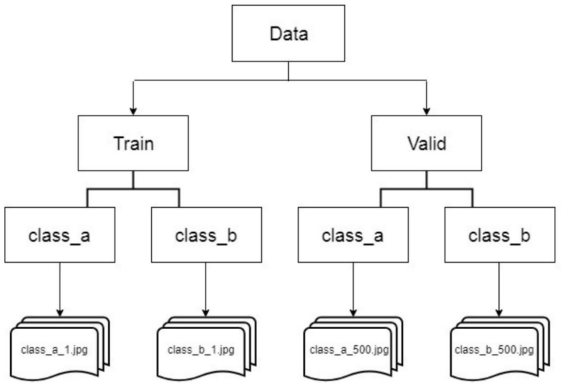


Fig 10: Dataset organization for the CNN

To make the CNN more robust, we applied some data augmentation to the dataset (shear, zoom and flip all randomly)

Not surprisingly, the model quickly (5 epochs) converged to a 99.9% accuracy as this is a very simple task.

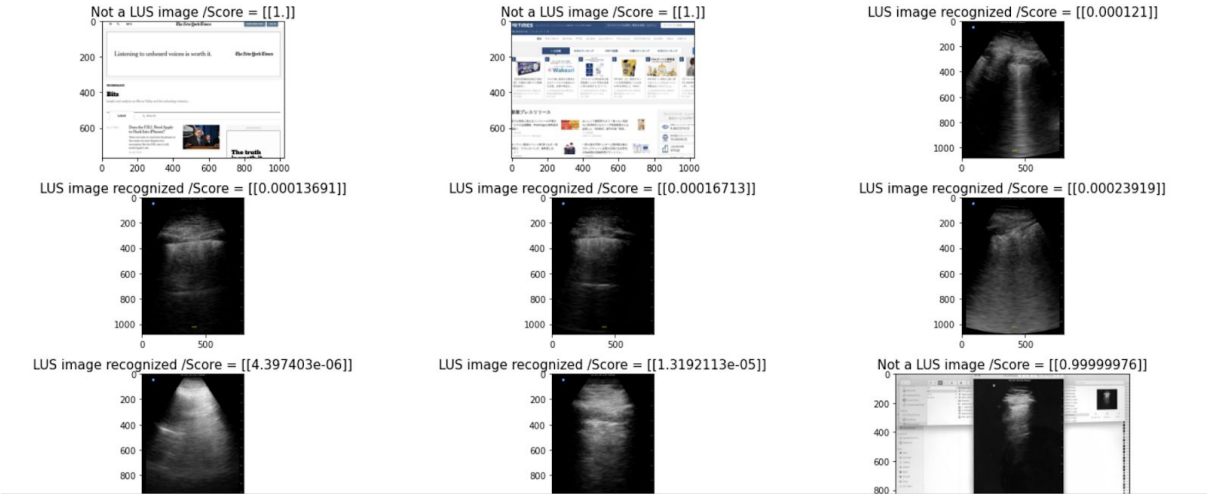


Fig 11: Result of the CNN

By implementing the CNN into the image scanner, we can select the LUS image among multiple rectangles, as you can see below.

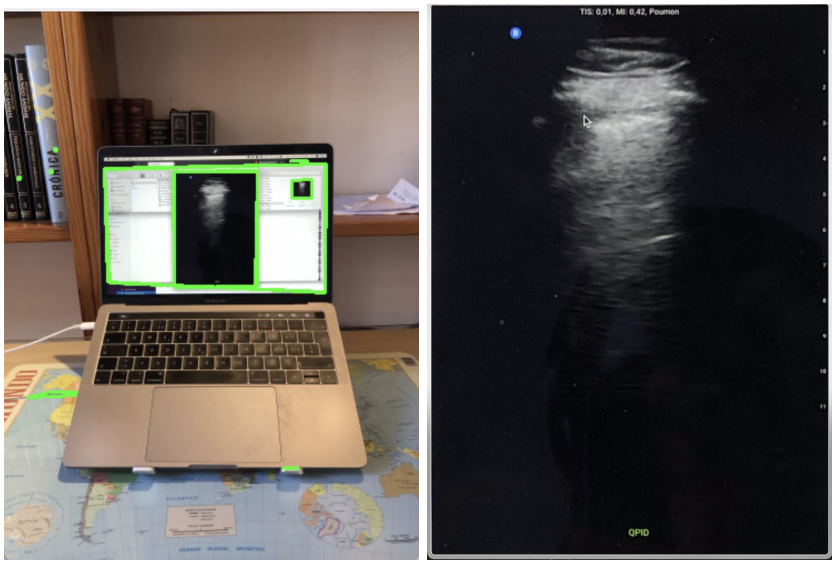


Fig 12: Result of implementing the CNN to select the rectangle with the LUS image.

2.1.3. Testing and debugging the preprocessing

To test the result of the scanned images within DeepChest, we needed to compare the prediction of DeepChest on both the original image vs the scanned ones. A dataset of 62 images was created for this purpose.

However the scanned images were not compatible with the preprocessing step of image reshaping (from convex to rectangular images). The reshaping works by finding the best cone of non-black pixels in the LUS image. In the case of scanned images, the black areas are not “real black”.

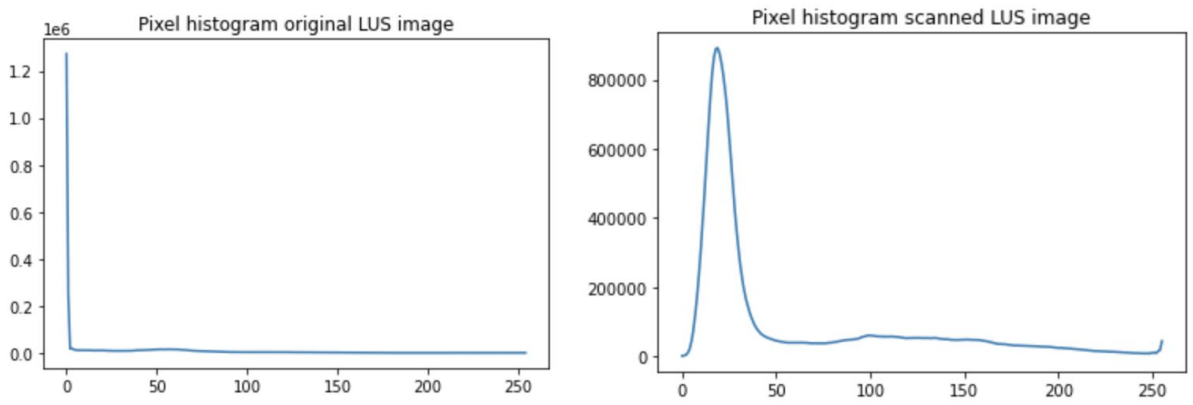


Fig 13: Pixel histogram of an original LUS image and a scanned one

One idea to solve this was to reset the first 50 pixels of the scanned LUS image to 0, in order to obtain the same spike as in an original image. Unfortunately, at the time of this report, the DeepChest model was not yet completed and the tests have not yet been run.

2.2. Android app

The interface for the DeepChest tool will be on a phone app. It is very likely that the app will be developed in an Android environment, and not iPhone, as the app intends to be delivered to low-resource settings countries, where iPhones are less common.
iGH is planning to have the EPFL Junior Entreprise develop the app, but this is yet to be started.

A preliminary draft of the UI of the app has been started, using Figma (see figure below). The app could guide clinicians on acquisition of images (key anatomical regions needed), uploading of the images (screenshot or photo), as well as capturing complimentary clinical data that could further assist the predictive algorithms.An example of information about the patient that could be collected is::

Pre-existing conditions and medical history:

- Smoking history, Presence of chronic diseases (diabetes, heart disease, obesity

Symptoms:

- Fever
- Difficulties breathing
- Cough
- Loss of taste and smell

Clinical signs:

- Temperature
- Respiratory rate
- Oxygen saturation

Fig 14: Draft of the UI for the app

2.3. What's next

- Finalize testing of the scanned images through DeepChest by running predictions on the dataset of original vs scanned images. Depending on the result, a decision will need to be taken: continue with the image scanner if the loss of information is low enough for DeepChest to perform properly, or find a new way to extract the LUS images from the various ultrasound devices used.

- Develop the android app. For the back end, it is preferable if the preprocessing (image scanning, CNN, reshaping and cropping of text) happens locally on the smartphone, for privacy reasons. The CNN was developed with this in mind, indeed the keras model can be deployed on Android through TensorFlow Lite [29]. As for the DeepChest model, it would preferably run locally for the same privacy reasons. Yet, in standard practice such a model runs on a cloud server and is accessed through an API. A feasibility and trade off study need to be done for this matter.

Bibliography

[1] MAHMOUD SAID (2020); Algononymous:A Roadmap to developing a platform for algorithm-optimized anonymization in medical datasets. (semester project report)

[2] LAMBERT, R., LEUZ, C. and VERRECCHIA, R.E. (2007), Accounting Information, Disclosure, and the Cost of Capital. Journal of Accounting Research, 45: 385-420.

[3] Thijs Benschop, Cathrine Machingauta, Matthew Welch (Nov 12, 2019); Statistical Disclosure Control: A Practice Guide

[4] Pierangela Samaratiy, Latanya Sweeney (1998); Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression

[5] EMA. European Medicines Agency policy on publication of clinical data for medicinal products for human use [Internet]. https://www.ema.europa.eu/. 2018. Available from: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, (2007); l-diversity: Privacy beyond k-anonymity,"

[7] Arbuckle L, El Emam K. Chapter 3. A Practical Risk- Management Framework. Build. an Anonymization Pipeline Creat. Safe Data. O'Reilly Media, Inc; 2020.

[8] M. J. ELLIOT, A. M. MANNING and R. W. FORD (2002); A computational algorithm for handling the special unique problem.

[9] Krishnamurty Muralidhar, Rathindra Sarathy (2006); Data Shuffling — A New Masking Approach for Numerical Data

[10] Gouweleeuw, J. et al. "Post randomisation for statistical disclosure control: Theory and implementation." (1997)

[11] Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen (2014); Introduction to Statistical Disclosure Control (SDC)

[12] D. Defays and M.N. Anwar1 (1998); Masking Microdata Using Micro-Aggregation

[13] Yancey W.E., Winkler W.E., Creecy R.H. (2002) Disclosure Risk Assessment in Perturbative Microdata Protection. In: Domingo-Ferrer J. (eds) Inference Control in Statistical

Databases. Lecture Notes in Computer Science, vol 2316. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-47804-3_11

[14] Tropical medicine, Oxford University website: https://www.tropicalmedicine.ox.ac.uk/research/oxford/infectious-diseases-data-observatory

[15] L. Sweeney (2011); Patient identifiability in pharmaceutical marketing data," Harvard University, Cambridge

[16] S. Sharma, P. Gupta, and V. Bhatnagar (2012); Anonymisation in social network: A literature survey and classification," International Journal of Social Network Mining"

[17] C. A. Cassa, S. C. Wieland, and K. D. Mandl (2008); Re-identication of home addresses from spatial locations anonymized by gaussian skew," International journal of health geographics, vol. 7,no. 1, p. 45

[18] P. Dahal, Matthew Brack, Laura Merson (2020); IDDO: Assessment of statistical disclosure.

[19] UTD-AT: Hussien, A. , Darwish, N. and Hefny, H. (2015) Utility-Based Anonymization Using Generalization Boundaries to Protect Sensitive Attributes. *Journal of Information Security*, **6**, 179-196.
UTD-AT source code: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=home

[20] CAT download: https://sourceforge.net/projects/anony-toolkit/files/Documents/cat-mannual-1.0.PDF/download

[21] Amnesia: https://amnesia.openaire.eu/index.html

[22] *μ*-Argus: Franconi L., Polettini S. (2004) Individual Risk Estimation in *μ*-Argus: A Review. In: Domingo-Ferrer J., Torra V. (eds) Privacy in Statistical Databases. PSD 2004. Lecture Notes in Computer Science, vol 3050. Springer, Berlin, Heidelberg.

[23] About PARAT De-Identification Software [cited 04 Aug 2014] Privacy Analytics Inc; Available from: http://www.privacyanalytics.ca/software/parat/

[24] Arx-a comprehensive tool for anonymizing biomedical data. American Medical Informatics Association; 2014. AMIA Annual Symposium Proceedings
Arx download: https://arx.deidentifier.org/downloads/

[25] SDCmicro: Matthias Templ. 2008. Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. Trans. Data Privacy 1, 2 (August 2008), 67–85.
SDCmicro download: https://github.com/sdcTools/sdcMicro

[26] Heroku cloud application platform: https://www.heroku.com/

DeepChest:

[27] Hugo Schmutz (2020); Master thesis report; DeepChest: A neural attention model for interpretable, missingness-resilient diagnosis and risk stratification of COVID-19 from lung ultrasound images

[28] Website Screenshots Dataset:
https://public.roboflow.com/object-detection/website-screenshots

[29] To deploy ML model on android: https://www.tensorflow.org/lite/