CAS Applied Data Science, University of Bern, S. Haug

**Assignment Work Instructions for Module 1**

The assignment work for Module 1 is a Conceptual Design Report (CDR) for a Data Science project you would like to perform. Maybe you already know what will be your final CAS project, however, that doesn’t have to be? Please use or orient yourself according to the template below. Remove this page for the submission version. Key information for the report is the following.

**Language:** English (exceptionally German)

**Deadline:** To be defined in class

**Expected effort and length:** About 30 hours, minimum 5 pages. You may work in teams (max 3).

**Further formal quality requirements:**

● All references must be listed in corresponding Reference section at the end of the CDR and cited with numbers in text

● All tables and figures should have numbered legends with short explanations (tables above, figures below) and be referenced in text (Figure 1: blabla, Table 1: blablabla).

● Figures to be as self explanatory as possible, e.g. plots with at least axis labeling including units.

**Gaëlle Marti Simon Rime Lenja Flütsch**

Your addresses Gesellschaftsstrasse 31, 3012 Bern

Your emails lenja.fluetsch@students.unibe.ch

**Data Science Project**

**Deciphering C-Section Rate Disparities: Analyzing Swiss Hospital Traits for Insights**

**Conceptual Design Report**

**30 October 2023**

# **Abstract**

One important health care trend that is continuously monitored is the C-Section Rate in a given region or institution. The C-Section Rate describes the number of c-sections (operative procedure to delivery a baby) performed compared to the number of vaginal deliveries. It is well known in the medical community that the C-Section Rate in Switzerland and many other countries is too high and that there is significant variation between regions and hospitals. However, the reasons underlying the high variation between institutions are unclear. In our project, we will use publicly available healthcare data and data on the infrastructure and staffing of different hospitals and apply supervised and unsupervised machine learning algorithms to extract patterns that could potentially account for the differences in c-section rates observed across different institutions.

# 

# **Table of Contents**

**Abstract 1**

**Table of Contents 2**

**1 Project Objectives 2**

**2 Methods 2**

**3 Data 3**

**4 Metadata 3**

**5 Data Quality 3**

**6 Data Flow 3**

**7 Data Model 3**

**8 Documentation 4**

**9 Risks 4**

**10 Preliminary Studies 4**

**11 Conclusions 4**

**Acknowledgements 4**

**Statement 4**

**Appendix X 5**

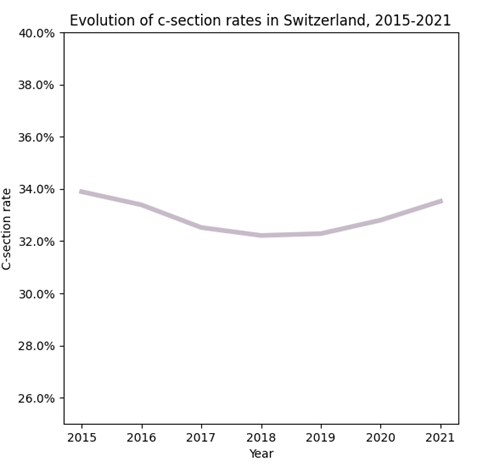
**References and Bibliography 5**

# **1 Project Objectives**

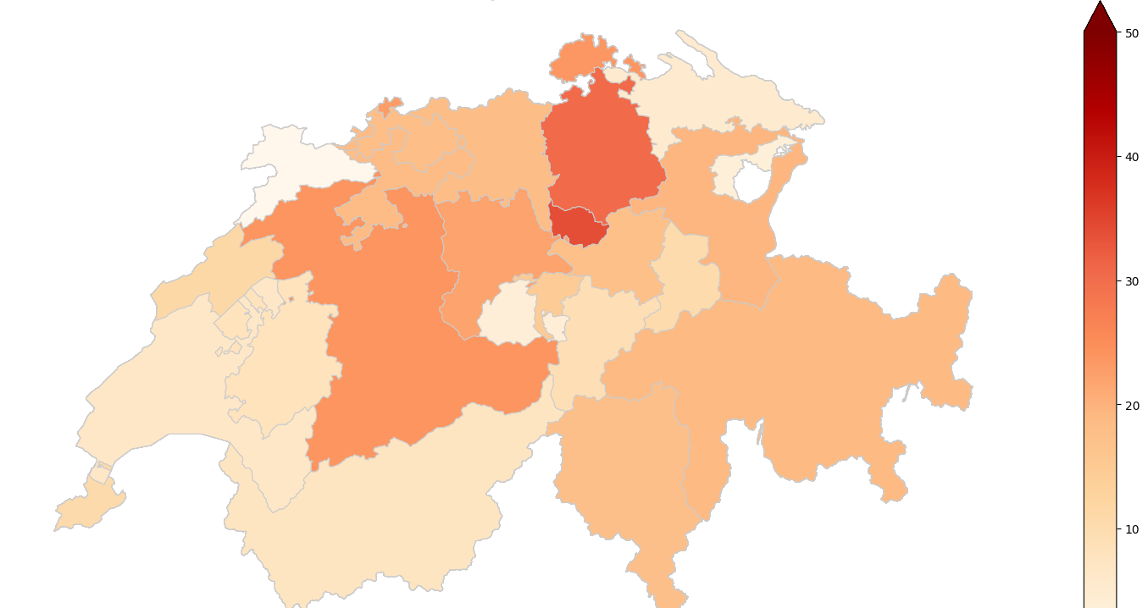
A C-Section (Cesarean Section) is an operative procedure performed by trained medical professionals and is conducted in situations where a normal vaginal delivery is not possible or deemed too risky.

Over the years, the c-section rate has been increasing worldwide. Indeed, studies found that the average c-section rate increased by 19 percentage points from 1900 to 2018, to reach 21.1% (global average of c-section in the world) and all regions have experienced such upward trend.[1.1]

Since 2015 in Switzerland and as illustrated in figure 1, the rate of c-sections seems to have the same as a “flatten U” (unfortunately, we weren’t able to obtain data prior to 2015 from the FOPH). It looks like a decrease was observed until 2018 and since then, the rate has started to increase to reach 33.53%. Indeed, in 2021, 86’810 children were born in Switzerland out of which 28’544 were delivered via a C-Section [1], thus one in three babies is born via C-section.



However C-Sections themselves are, like any other invasive operation, not without complications. Furthermore, a WHO working group has found that there is a reduction in maternal and newborn deaths when C-Section rates rise towards 10%. However, once C-Section rates rise above 10% there is no further improvement in mortality rates [2]. This illustrates that while C-Sections are lifesaving in some situations, such an operation is often performed unnecessarily. As a result, the overuse of c-section procedures may 1) expose the mother and her child to avoidable complications from the operation and 2) waste sparse healthcare resources. In 2021 the C-Section Rate in Switzerland was 32.8%, hence significantly higher than the 10-15% deemed ideal by the WHO Expert group [1,2]. Whilst there are scientific guidelines that clinicians use to decide whether a C-Section is necessary, the final decision is often based on the individual clinicians interpretation of the situation and may vary significantly between institutions and individuals. This is also observed in Switzerland, where C-Section Rates differ drastically between regions and institutions, with the difference between the German and Latin speaking cantons reaching statistical significance (see section 10 of the present document).

*Figure 1: Inpatient C-Section Rates per Canton 2021*

*Source: own dataset based on FOPH’s data and own visualization*

In our project we aim to analyze hospital traits, such as infrastructural characteristics and staff, as potential explanatory variables for the stark differences in C-Section Rates observed across Swiss hospitals. Our project is centered around determining patterns in hospital characteristics that can account for the variations in c-section rates among different healthcare institutions.

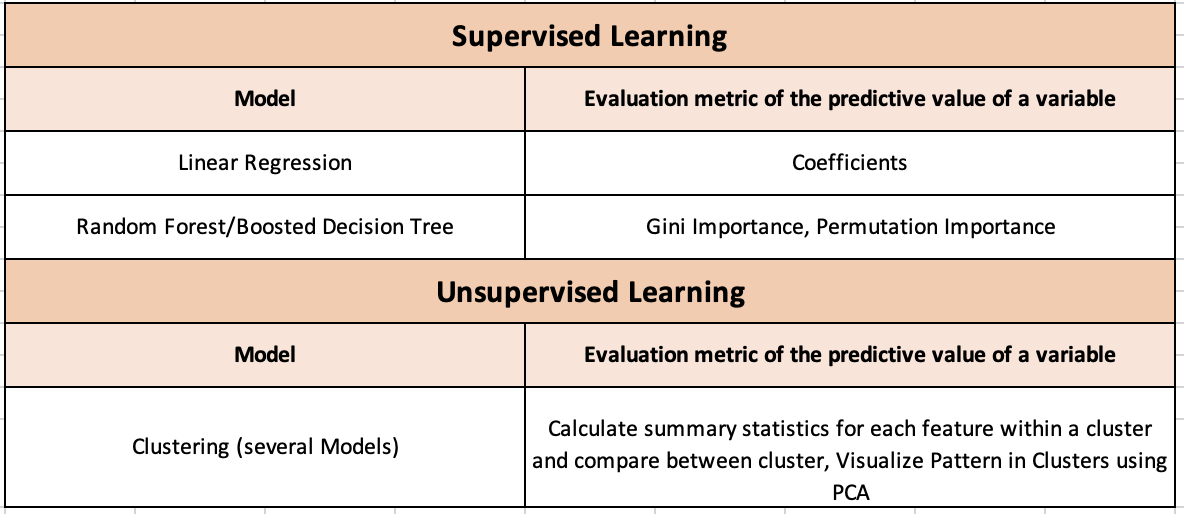
**2 Methods**

In order to be able to work on the project as a group, we are using Google Colaboratory and the files containing the code are stored on a shared google drive together with the csv-files containing the raw data and the final dataset.

The data regarding case numbers and hospital infrastructure are publicly available on the Website of the Federal Office of Public Health (FOPH). Data regarding the case number (C-Sections, Delivery) per hospital are only available online from 2015 onward, whereas the data regarding hospital infrastructure and staffing are available from 2008 onward. In order to acquire the data from 2008 to 2015 for the case numbers the FOPH Data management and statistics section was contacted.

The data was imported into a pandas dataframe and the data frames were merged on the hospital names and the year. Analysis will be carried out on one final data frame containing all the variables to be analyzed.

In order to generate robust results, several different supervised and unsupervised models will be used to uncover patterns that could potentially explain the differences in c-section rate between institutions. First, a linear regression model will be fitted and the coefficients will be analyzed. Since linear regression models assume a linear relationship between the feature and the c-section rate, we might miss more complex relationships. Hence, in a second step, a random forest model and a boosted decision tree will be used. These models may be more appropriate to uncover non-linear relationships in our data. Feature importance in those models will be analyzed by looking at the Gini Importance (a measure of how much impurity decreases when our feature of interest is used for splitting) and Permutation Importance (evaluating overall model performance when one feature is changed). Lastly, unsupervised clustering models will be generated and the clusters will be analyzed by calculating summary statistics for each feature in a cluster and then comparing these across different clusters. Furthermore, using PCA we can try to extract patterns underlying the clusters.

*Table1: Overview of Models used for Analysis*

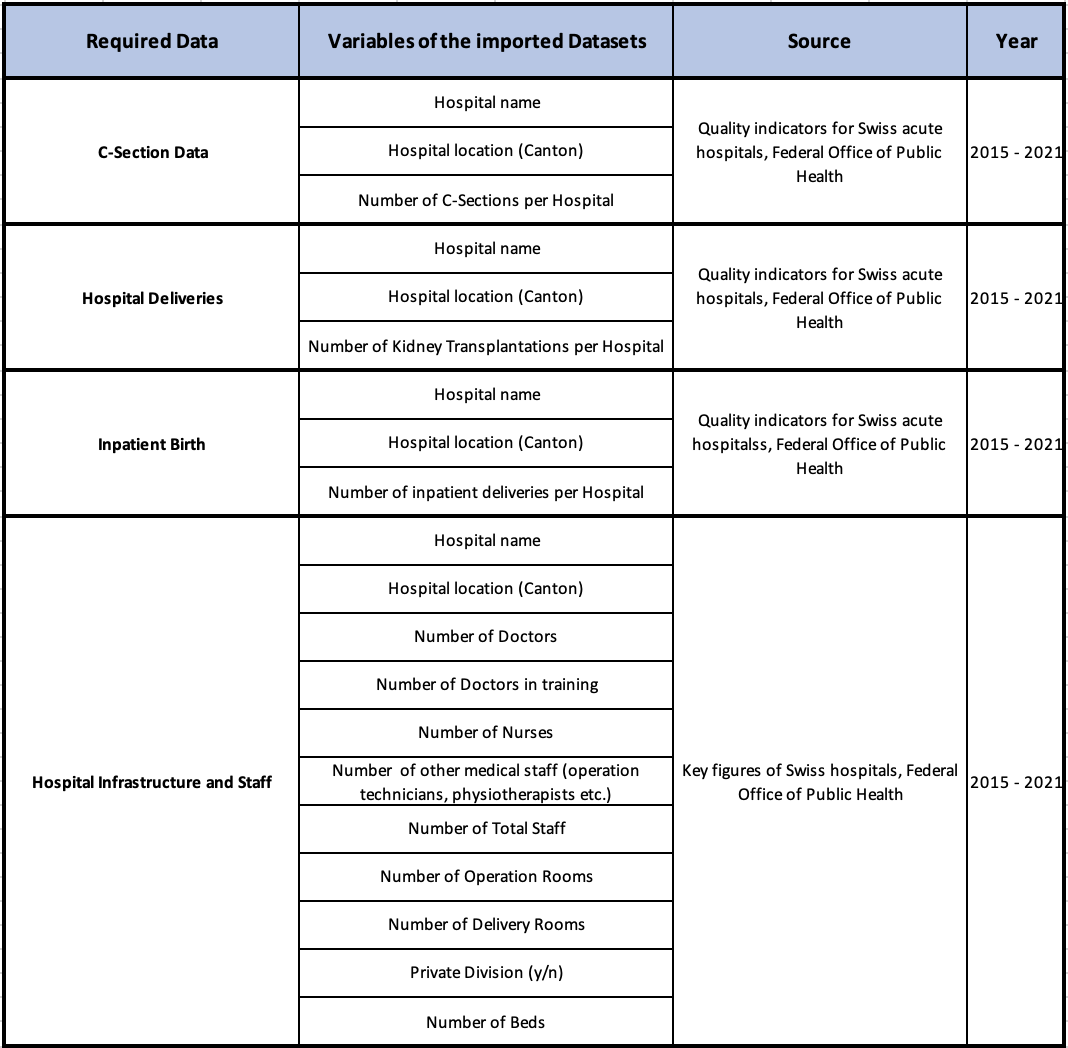
In order to run the different models, the main python libraries will be used:

* Pandas: data manipulation (https://pandas.pydata.org/)
* Matplotlib: plots and visualization (https://matplotlib.org/)
* statsmodels.api: statistical test and tools (https://www.statsmodels.org/stable/api.html)
* SciKitLearn or sklearn: machine learning (https://scikit-learn.org/stable/)
* Numpy: mathematical operations on arrays (https://numpy.org/)

# **3 Data**

In order to perform our analysis, we import relevant data regarding Swiss hospitals. Switzerland has 276 hospital (“Spitalbetriebe”), with 38% being general care hospitals and 62% being specialized clinics. The number of hospitals is decreasing, in part due to the merge/grouping of hospitals. Additionally, 32% of hospitals are spread across multiple locations [3].

All the data regarding the infrastructure, staffing and the case numbers handled at these hospitals is publicly available on the website of the Federal Office of Public Health (FOPH) [4, 5]. Since the data is publicly available, no security issues arise and no special measures need to be taken. From the FOPH Website, four different datasets were imported (see Table 2). The data from 2008-2015 for the case numbers of c-sections, deliveries and kidney transplant were requested from the responsible office and are not available on the website.

*Table 2: Overview of the imported Datasets [4, 5]*

Several operations were performed on the imported dataset (see Table 3) and the newly created variables were stored in a new column added to the dataframe. One variable was added for university hospital, as university hospitals usually follow high-risk pregnancies and hence might have higher c-section rates. We also added a variable for the language spoken in the cantons, as it might be interesting to see whether the formation of the doctors and the culture differences between these cantons might impact the c-section rate. Lastly, we calculate the c-section rate by using the case numbers obtained for c-sections and deliveries.

*Table 3: Newly created variables and operations performed*

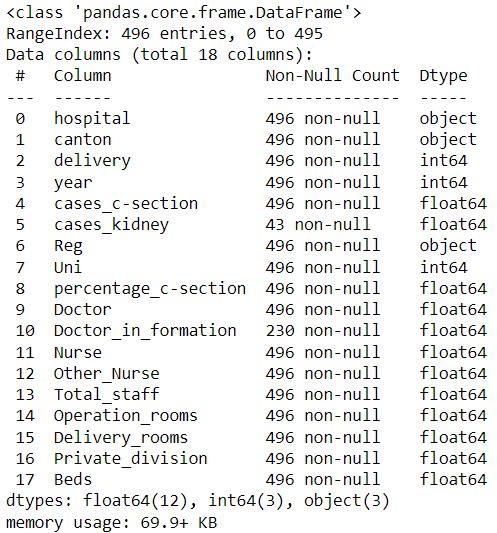
# 

Cleaning of the dataset involved several steps:

1. Removing all the ‘Geburtshäuser’/’Maison de naissance’ as they do not perform c-sections and are hence not relevant for our analysis
2. Removing all the hospitals in a given year that did not perform any c-section at all because they only had 1 delivery
3. Removing lines where data on the number of beds was missing, as number of beds will be needed in order to standardize our variables
4. The variable ‘doctor in formation’ was removed as many values prior to 2019 were missing

Our final dataframe contains the following columns: 'hospital', 'canton', 'delivery', 'year', 'cases\_c-section', 'cases\_kidney', 'Region', 'Uni', 'percentage\_c-section', 'Doctor', ‘Doctor in formation’, 'Nurse', 'Other\_Nurse', 'Total\_staff', 'Operation\_rooms', 'Delivery\_rooms', 'Private\_division', 'Beds'. Figure X below gives succinct information about the dataframe

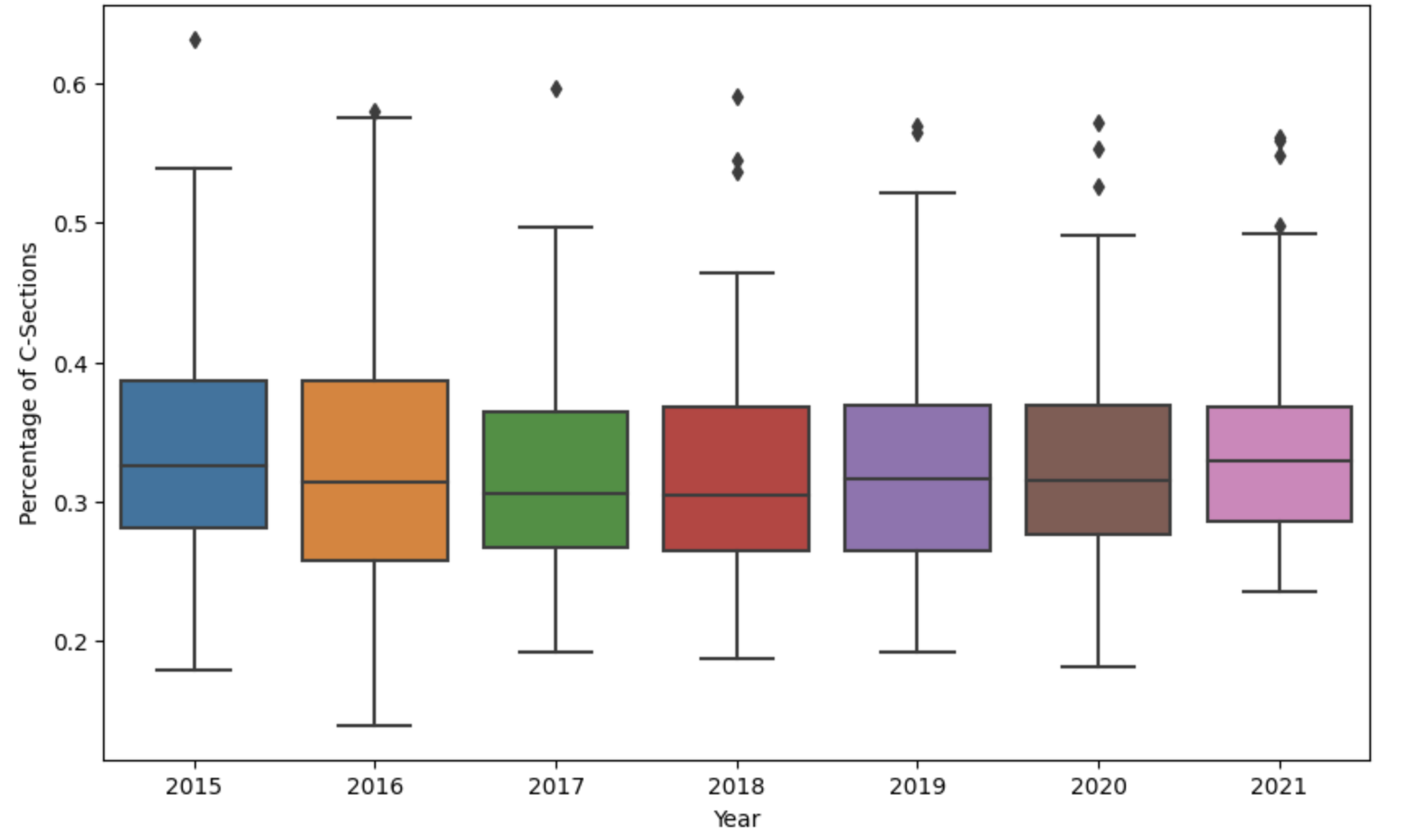
Figure X: information about the dataframe



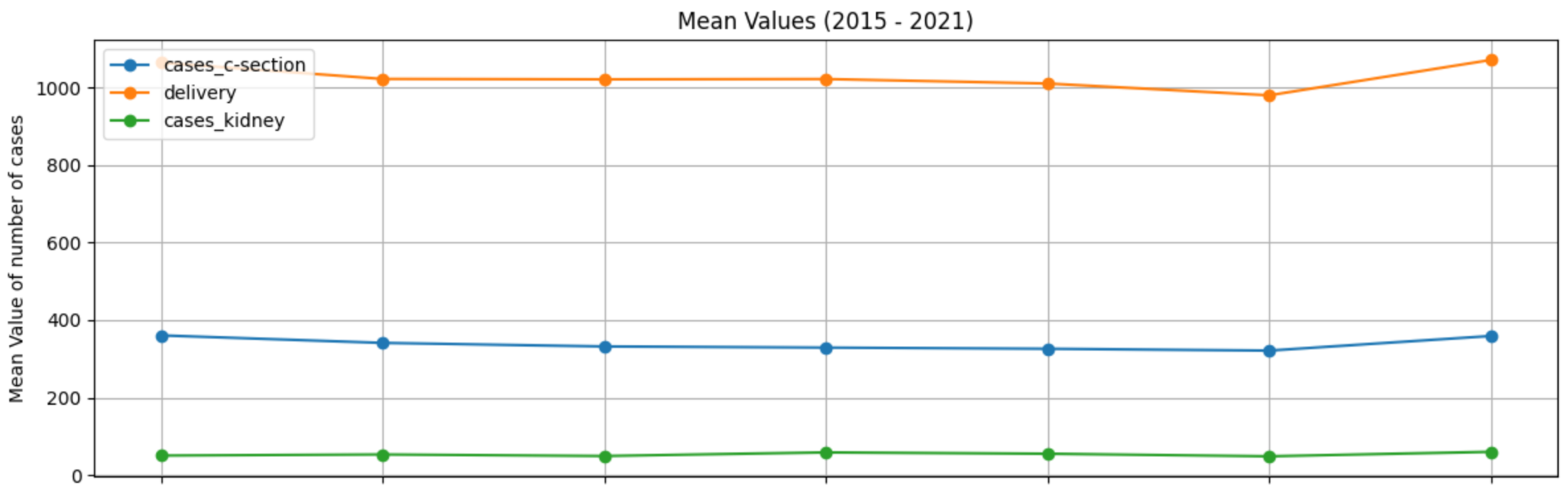
For the analysis, we will train our models to predict the variable ‘percentage\_c-section’. Moreover, we will normalize some of our numerical variables to the number of beds, in order to account for differences in hospital sizes. The variables normalized are the number of doctors (becoming number of doctors per bed), number of nurses, total staff, number of operations and delivery rooms.

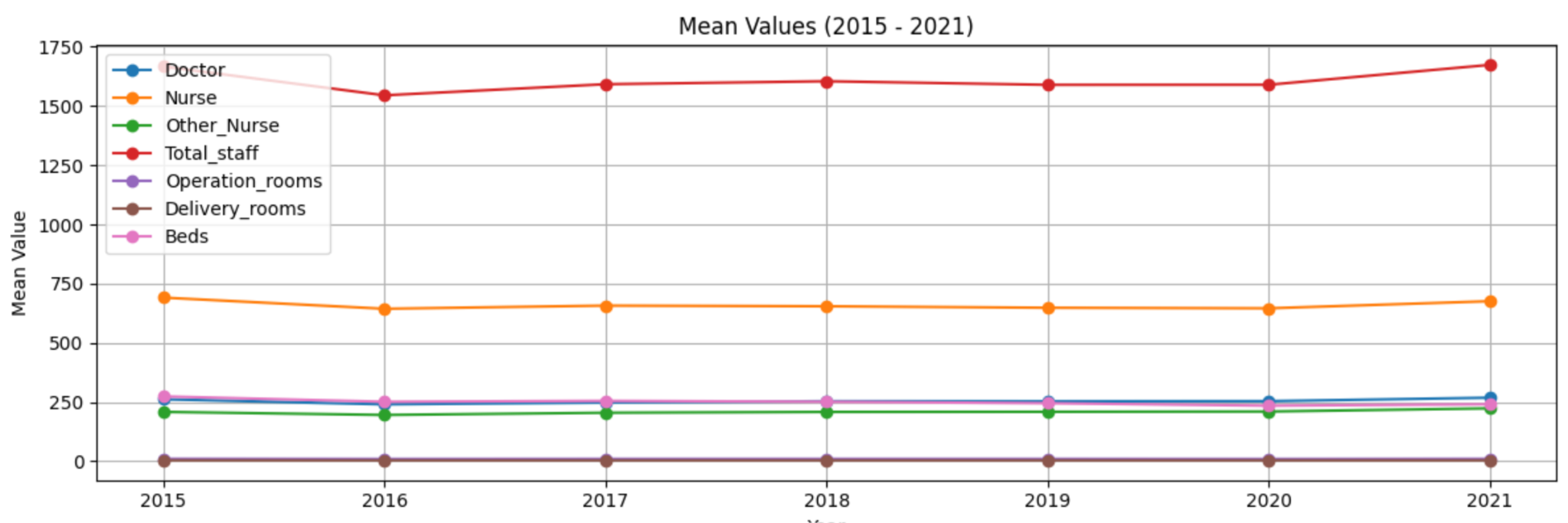
The C-Section rates calculated using our approach vary only slightly over the years and the median over the last year is around 30% (see Figure 2), which is what we would expect considering the rate of 32.8% that the Swiss Health Observatory calculated for 2021 [1]. Similarly, the mean of the values of the other numerical variables in our dataset has also remained stable over the years (see Figure 3). Regarding the categorical variables in our dataset, we have more hospitals in the german speaking region compared to the latin speaking regions (see Figure 4). The same applies for the distinction University vs Non-University hospitals (see Figure 4), where we only have 5 University hospitals nationwide and hence lesser data points.

*Figure 2: Boxplots showing the distribution of c-section rates across different hospitals in Switzerland for the year 2015-2021*

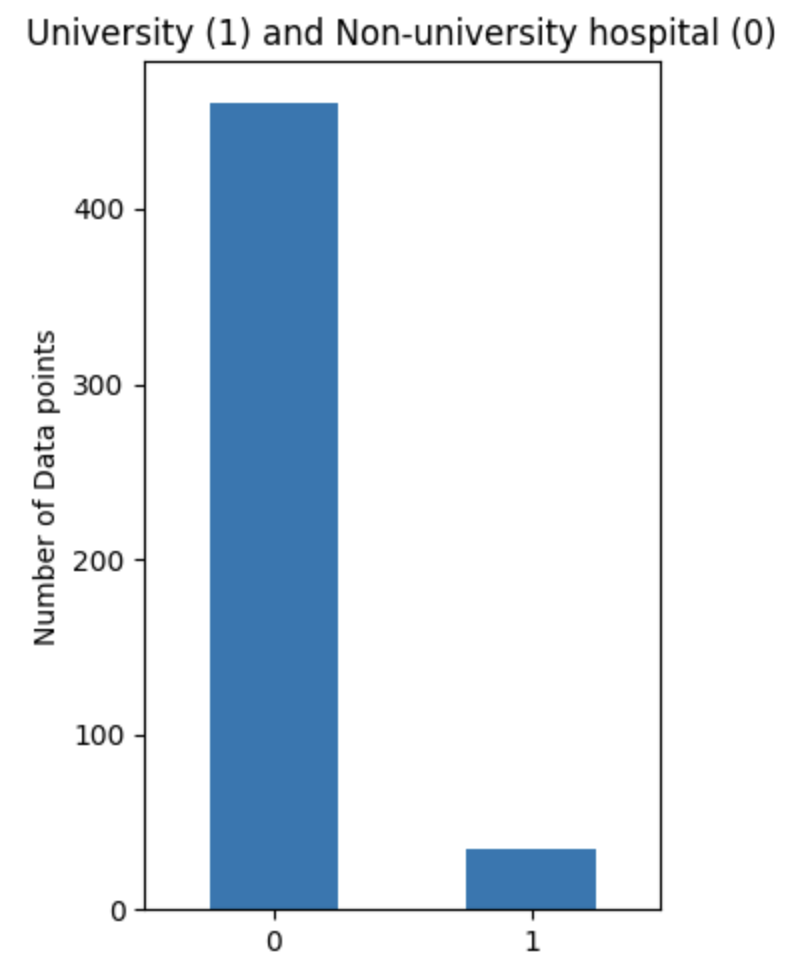
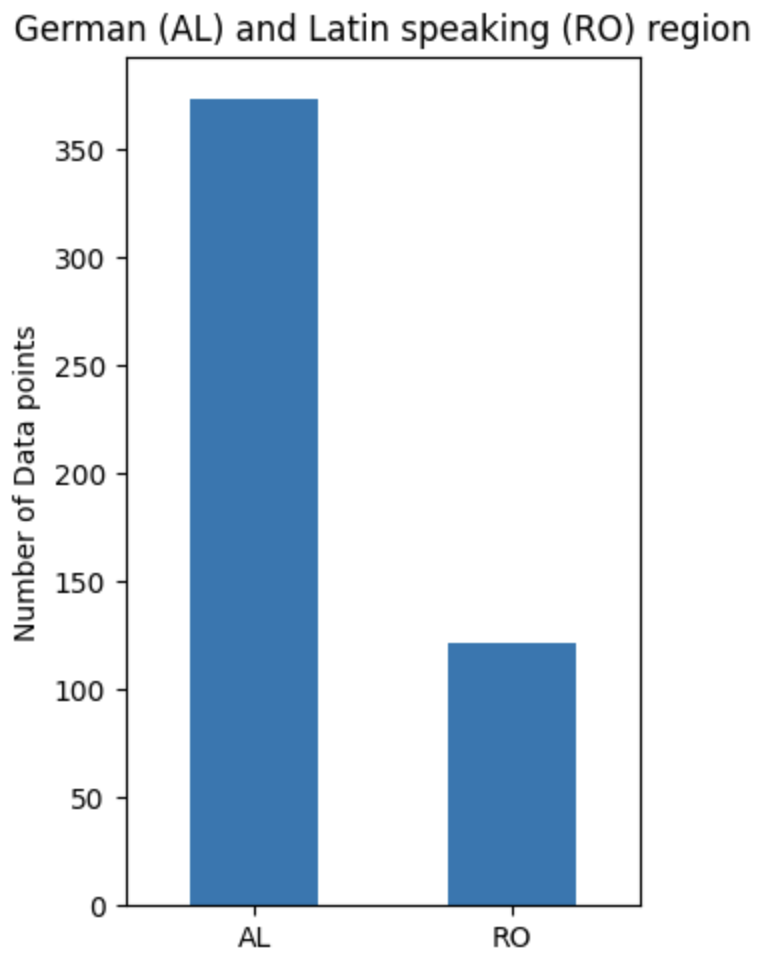


*Figure 3: Trends in case numbers, infrastructure and staffing in our dataset*





*Figure 4: Distribution of the data from 2015-2021 for the two categorical variables Region and University Hospital*

**

# 

# **4 Metadata**

The main source of metadata is the information stored on the pages of the Federal office of public health’s website of Switzerland (FOPH). We have two different sources of data:

1) Quality indicators: the indicators are created based on the Hospital Medical Statistics (MS) from the Swiss Federal office of statistics (FOS). In this statistic, the FOS collects every year patients’ data (socio-demographic information, such as age and region of residence, etc.), administrative data (e.g. type of insurance), and medical data (such as diagnoses and treatments).[1] Main and secondary diagnoses are coded according to the International Classification of Diseases (ICD-10), while treatments and procedures are coded according to the Swiss Classification of Surgical Procedures (CHOP).[2] Therefore, cases of c-section, deliveries and kidney transplants are coded according to CHOP’s classification. The latest year available is 2021.

Each of the files is obtained by webscapping (one csv file per year and per type of procedure, i.e. c-section, delivery, kidney transplant) , then merged together and saved in csv in the shared google drive.

2) Key figures of Swiss hospitals: we use the file “Zeitreihe der «Kennzahlen der Schweizer Spitäler» ab 2008 (XLSX, de fr it)” and the file is saved on the shared google drive. The key figures are computed by the FOS every year and combine data on hospital statistics (KS) and on the Hospital Medical Statistics (MS). Each year, the data are made available to the FOPH by the FOS.[3] The latest year available is 2021.

The two datasets are publicly available and are updated on a yearly basis (the FOS is legally obliged to publish operating figures and medical quality indicators from Swiss hospitals)[4]. The last update was on November 11th 2022.

[1] https://www.bfs.admin.ch/bfs/fr/home/statistiques/sante/enquetes/ms.assetdetail.7373.html

[2]<https://www.versorgungsatlas.ch/fr/p/glossar#ms>

[3]<https://spitalstatistik.bagapps.ch/data/download/kzp21_description_fr.pdf?v=1678279036>

[4]<https://www.bfs.admin.ch/bfs/fr/home/statistiques/sante/systeme-sante/hopitaux.html#:~:text=La%20Suisse%20compte%20276%20%C3%A9tablissements,sur%20plus%20d'un%20site>.

# **5 Data Quality**

What are the quality requirements you need in order to meet your project objectives (data size, data precision missing values, ….)?

Are they met? If not, do you expect a significant impact on your result?

Any measures to improve the data quality?

To meet our project objectives of analyzing C-Section rates across Swiss hospitals and identifying explanatory variables, several data quality requirements are essential.

In order to have a sufficient data volume to ensure statistical significance and robustness of our analysis, we have collected data for multiple years (2015 to 2021) for all hospitals performing C-sections in Switzerland.

Since the data originates from the FOPH, it is expected to possess the requisite level of accuracy for our analysis. However, problems of data consistency were observed. For instance, some of the hospitals' key figures have changed names since a certain year. For some others, the unit of value has changed (for example, the percentage of private rooms was written as “12” in 2015 and as “0.12” for the following years).

In order to avoid consistency problems in the time series and to improve data quality, we implemented a series of data cleaning steps to prepare the data for the analysis. The key measures were the following:

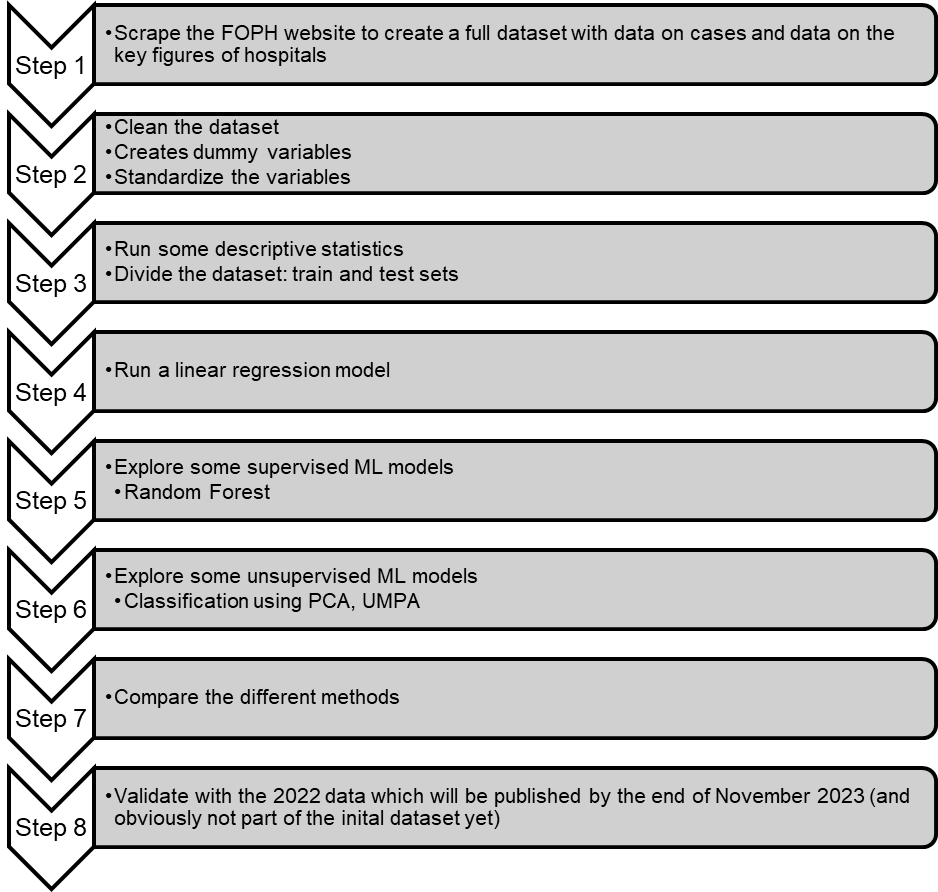
* One common issue encountered in the raw data was the presence of spaces used as thousands separators in numeric values (the value was stored as string). To address this, we systematically removed these to create integer variables.
* The raw data included percentages represented as values between 0 and 100. To make these values compatible with our analysis, we converted them to the standard float format between 0 and 1.
* In our analysis of C-Section rates across Swiss hospitals, it was essential to focus on healthcare institutions where C-sections are a relevant metric. As a result, we excluded "Maison de Naissance", "Geburtshaus" and “Geburtshus” facilities from our dataset. By definition, these locations do not perform C-sections.
* Another crucial step in data cleaning involved filtering out hospitals that did not perform any C-sections during the study period. Through analysis, we identified a group of seven hospitals that had minimal delivery activity, each reporting only one birth over a seven-year period. These hospitals, primarily health centers, were deemed exceptions in our dataset and were subsequently removed.
* The names of the hospitals have been processed to match those appearing in both publications 'Key Figures of Swiss Hospitals' and 'Indicators of Quality for Swiss Acute Care Hospitals' over the years.

# **6 Data Flow**

The procedure for data processing and analysis is described in figure 2.

Add Figure

1. Scrape data from the FOPH website
2. Clean the data and merge the different dataframe
3. Transform the data in order to use machine learning algorithms (create dummy, transform data to Numpy array)
4. Predict C-section rate using a linear regression method
5. Analyze of the linear model results (feature importance, R2)
6. Clustering of hospitals ???
7. Predict C-section rate using random forest
8. Train the model to obtain a score above XX%
9. Analysis of the results (plot results, score, feature of importance, …)



# **7 Data Model**

At the conceptual level, we are building a dataset of hospital’s characteristics that could explain patterns in hospitals performing c-sections in Switzerland.

At the logical level, we intend to apply three different models (linear regression, random forest, unsupervised models such as PCA) to explain c-section rates across Swiss hospitals. Using three models will help us analyze whether the use of more complex methods can enhance the explanation of the observed rate differences. If possible, we might want to validate our models with the 2022 data which will be published by the end of the year.

At the physical level, the dataset is stored in a CSV file on a shared Google drive (accessible only by the team member) and we assume that computations will not require any specific infrastructure.

In the appendix: Table Form

* Table Name: HospitalData
* Attributes:
  + Hospital (Text)
  + Canton (Text)
  + Delivery (Numeric)
  + Year (Numeric)
  + Cases\_C-Section (Numeric)
  + Cases\_Kidney (Numeric)
  + Reg (Text)
  + Uni (Categorical)
  + Percentage\_C-Section (Numeric)
  + Doctor (Numeric)
  + Doctor\_in\_formation (Numeric)
  + Nurse (Numeric)
  + Other\_Nurse (Numeric)
  + Total\_staff (Numeric)
  + Operation\_rooms (Numeric)
  + Delivery\_rooms (Numeric)
  + Private\_Division (Numeric)
  + Beds (Numeric)
  + Log\_Y (Numeric)
  + Kidney\_Dum (Numeric)
  + Region\_Dum (Numeric)
  + Doc\_Bed (Numeric)
  + Nurse\_Bed (Numeric)
  + Staff\_Bed (Numeric)
  + Operat\_Bed (Numeric)
  + Deliv\_Bed (Numeric)

The physical level is the following.

* File Location: The dataset is stored on Google Drive, and the file location is accessible to the team members.
* File Format: The dataset is in CSV (Comma-Separated Values) format

# **8 Documentation**

How will the project be documented?

Given that we are three people working on the project, we are very meticulous about the exchange of information. As already explained, we are working with a shared Google drive and we are making sure to include a clear description of all our codes in the different notebooks that we produce. Therefore, the explanations are useful for all team members and it is also a way to keep track of the methodology used and the assumptions made.

Moreover, we also like to work with one notebook per theme of the project, i.e. we have a notebook to create the dataset, a notebook for the descriptive statistics, etc. The project is hence documented per theme and in each notebook with detailed explanations.

# **9 Risks**

What can go wrong?

When this and that goes wrong, what counter measures do you have?

What will be the impact on the quality of the aimed output, project time schedule, project cost ?

We have grouped the risk into 3 categories.

The first category of risks identified are risks related to the availability of the data. The risk that the FOPH will no longer publish updated numbers is considered low. Indeed, they are legally obliged to produce indicators to monitor the healthcare system in our country. However, the FOPH could decide to modify the data publically reported (for instance, no longer publish information about the number of beds per hospital). Such a change of methodology would greatly affect our model which would ultimately need to be redefined.

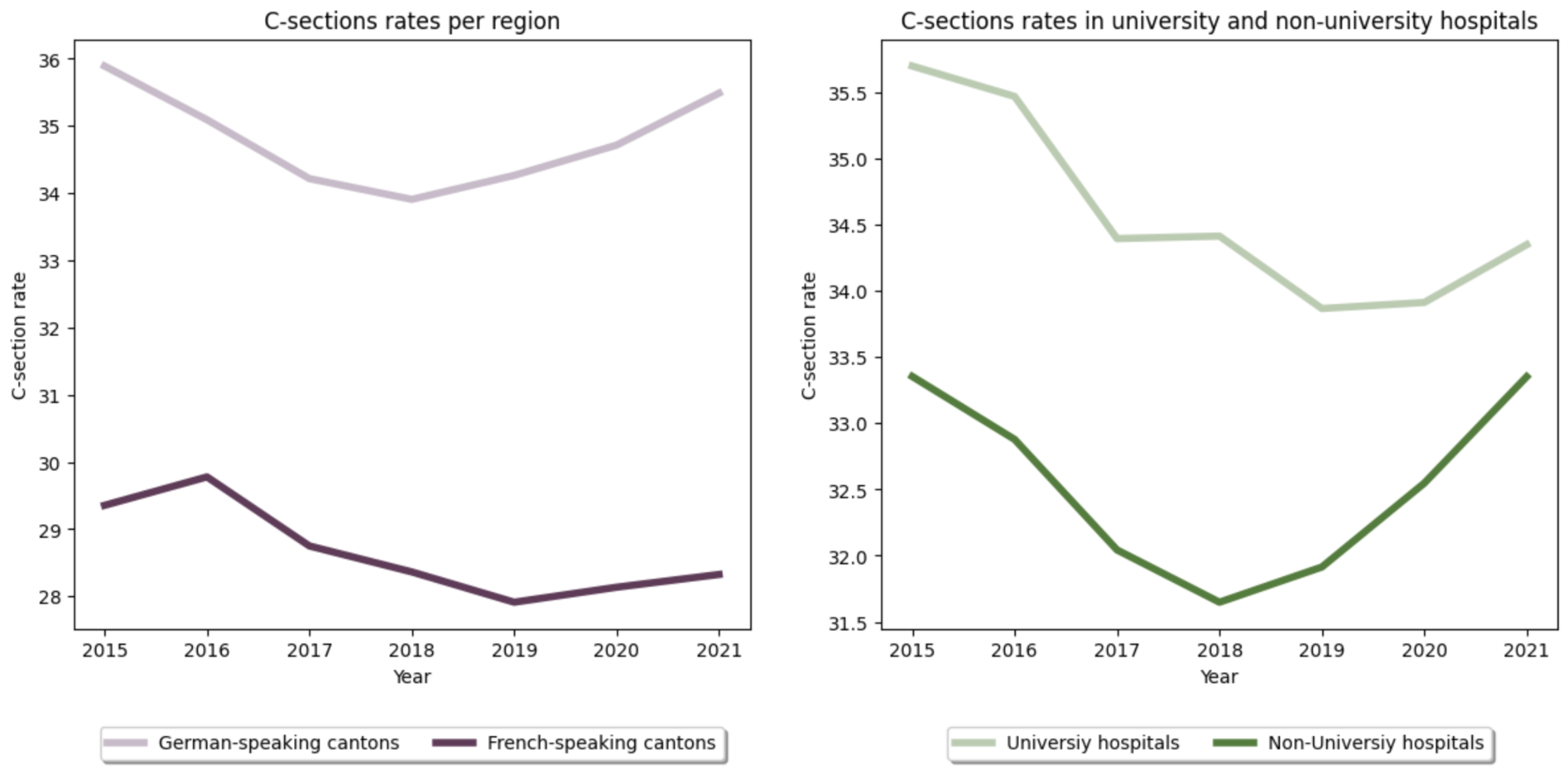
The second category of risks that we have identified is the risk that an external factor will impact the number of c-sections (a factor that would be not controlled for in our model). One important risk could come from changes in the healthcare insurance system. We could imagine important changes in healthcare reimbursement schemes which could impact the c-section procedure and therefore make our analytical model obsolete. For instance, if c-sections are no longer part of the list of interventions reimbursed by mandatory health insurance, the rate would drastically drop and our model would need to be fully reconstructed with probably other variables. Another example along the same lines would be if the criterions to perform c-section get much more restrictive. Such risk of a change in the insurance system is considered possible but in the long term.

Finally, the third category of risks related to the simplification of the model. Indeed, one of the drawbacks of our model is the fact that we don’t have any information on the patient. The health conditions of the pregnant woman are probably the main determinant of a c-section; however, we are not able to account for this. The risk is that our model is not able to capture the complexity of the inputs in the decisions to perform or not a c-section. To overcome such risk, we should try to obtain information on pregnant women based on hospital’s reports. This would result in a very costly and time consuming study.

# **10 Preliminary Studies**

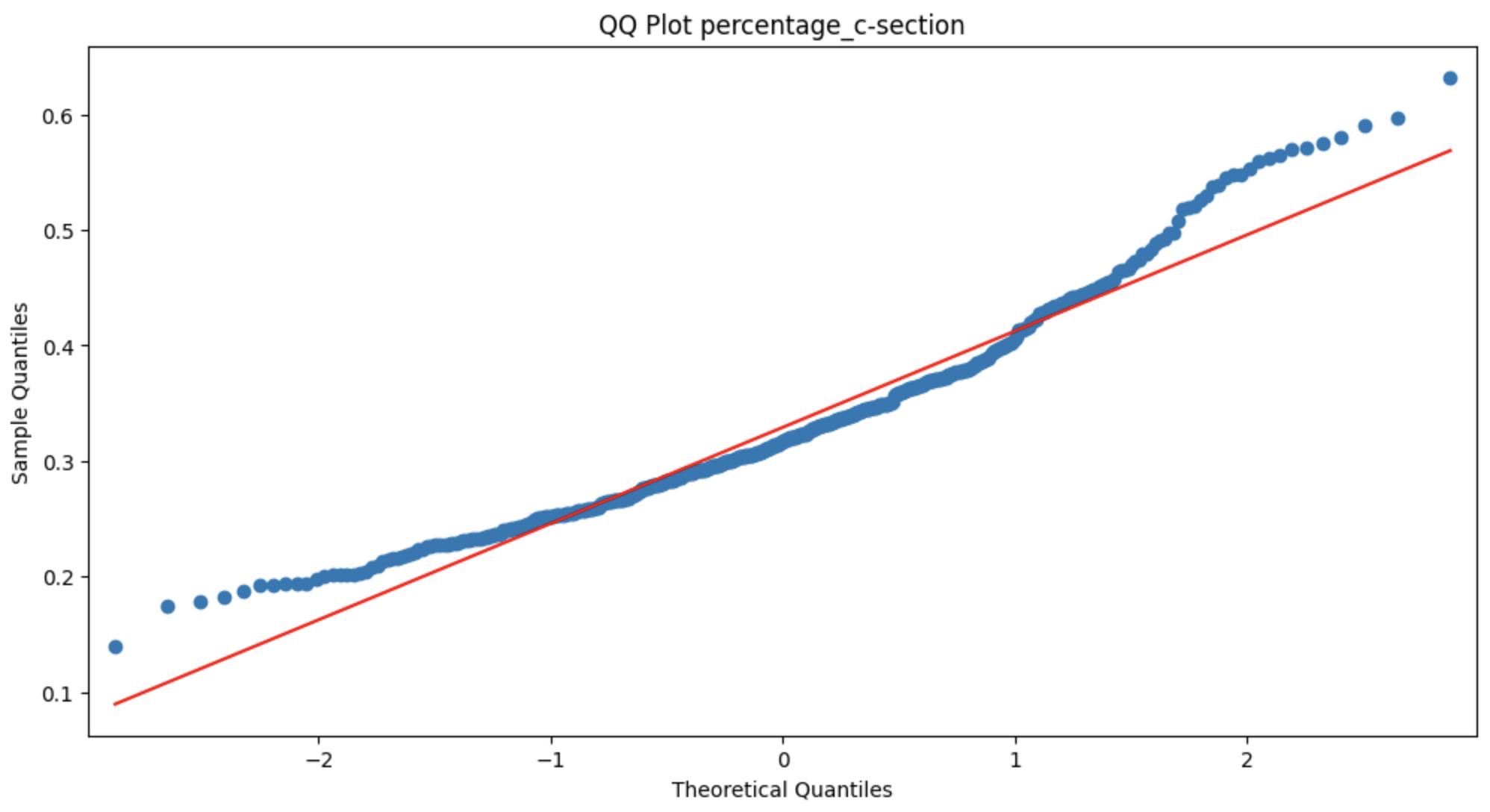
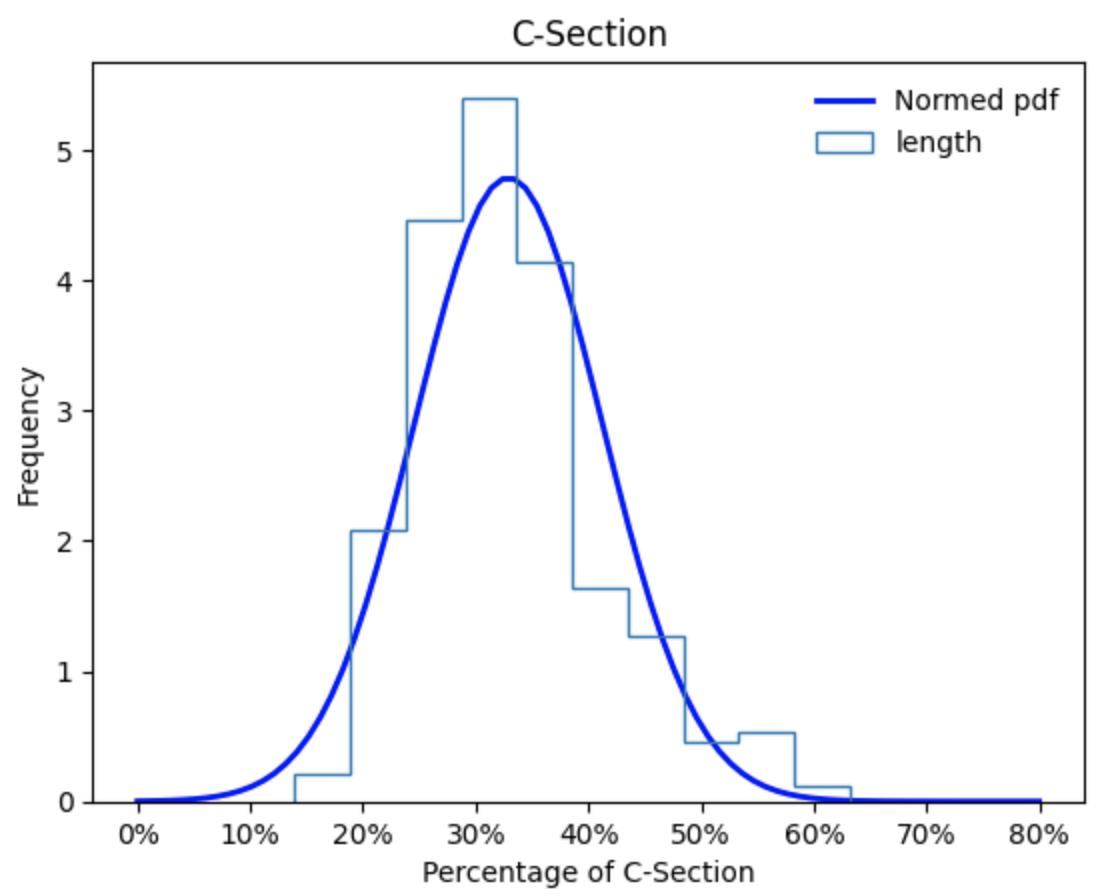
While running preliminary descriptive studies on our dataset, some factors that influence c-section rate have already become evident. There is a visible difference in c-section rates between the latin speaking and german speaking cantons as well as between university and non-university hospitals (see Figure 5). With a chosen significance level of 0.05, these differences reach statistical significance in our preliminary analysis.

*Figure 5: C-Section rates per region and type of hospital from 2015-2021*



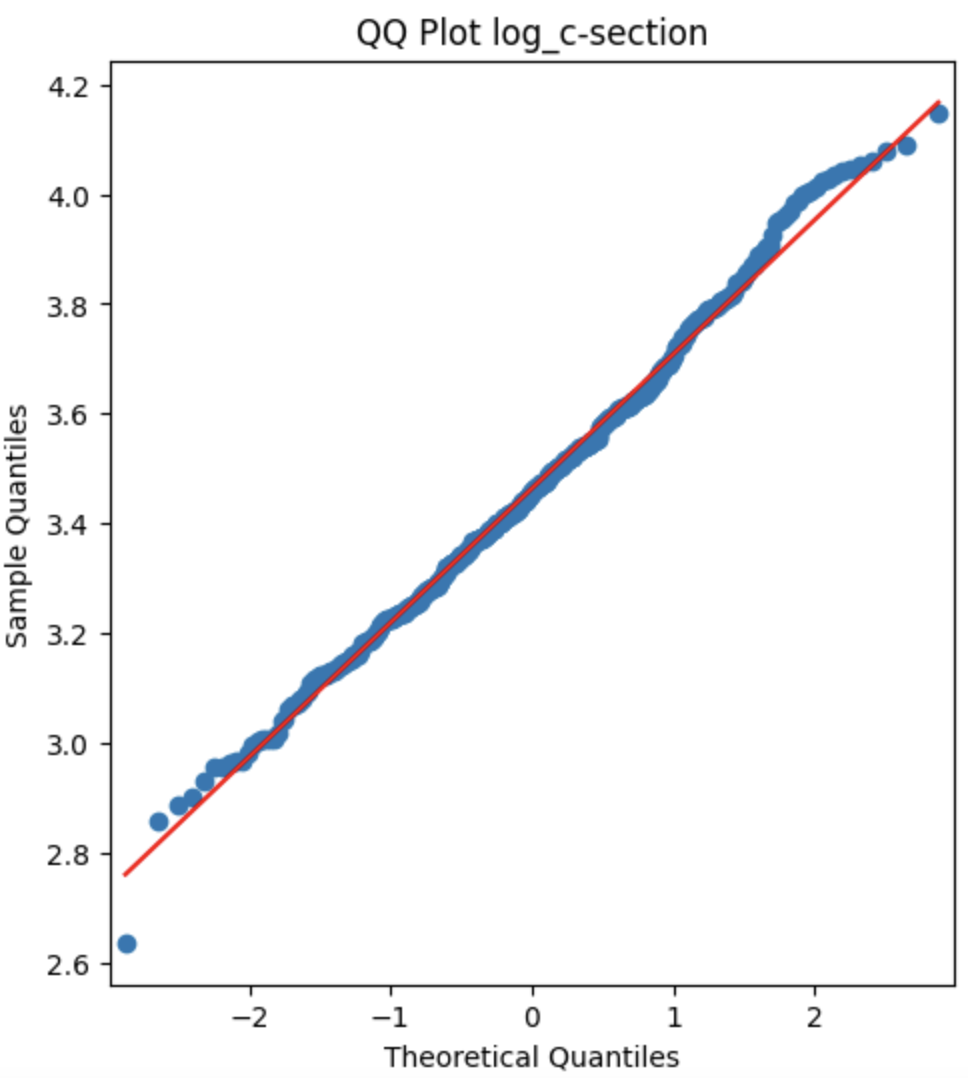
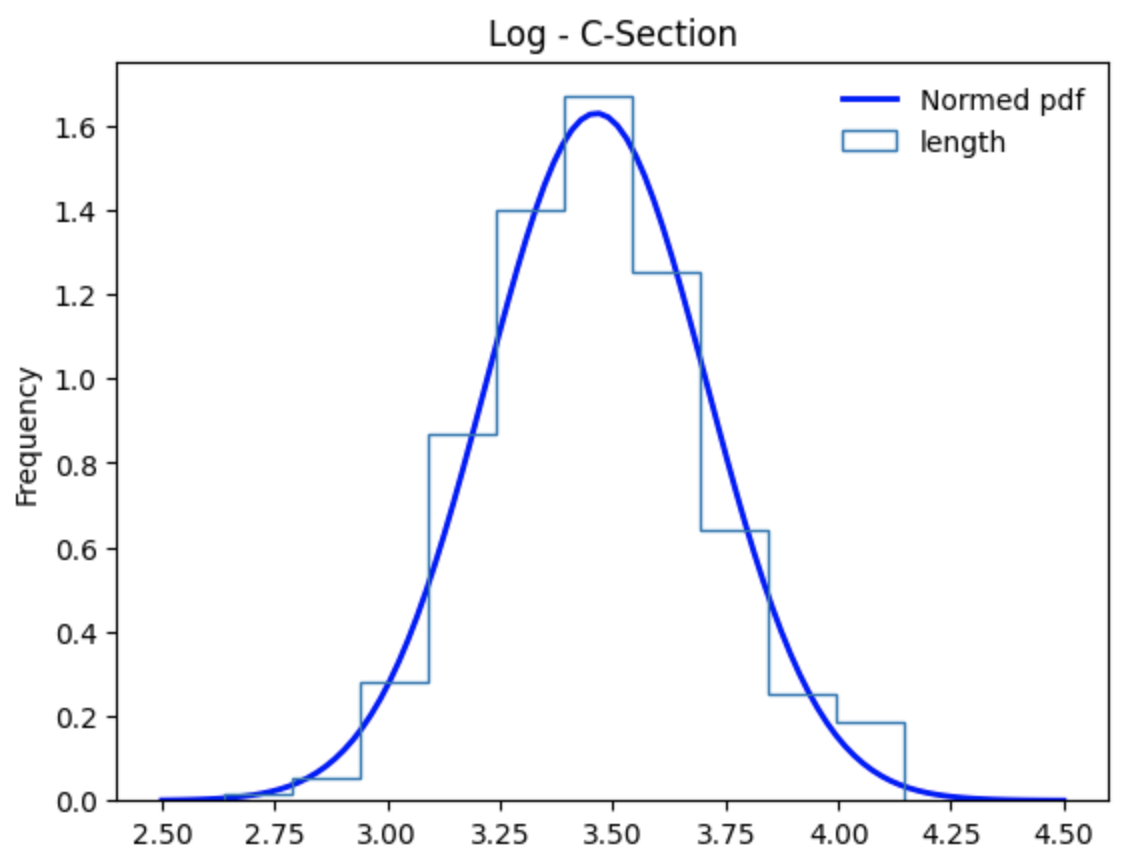
We also analyzed the distribution of our c-section rate data, checking for normality. The QQ-Plot of our data suggests that our data does not follow a normal distribution, (see Figure 6) which is supported by a d’Agostino Pearson Coefficient of 2.92E-13.

*Figure 6: Distribution and QQ-Plot of c-section rates across swiss hospitals from 2015-2021*



In order to obtain normally distributed data, we evaluated the distribution of our c-section rates when a logarithmic transformation was applied. The distribution and QQ-Plot of these values seems to be normally distributed (see Figure 7), which was confirmed by the d’Agostino Pearson coefficient of 0.33.

*Figure 7:* *Distribution and QQ-Plot of the c-section rates after logarithmic transformation across swiss hospitals from 2015-2021*



Considering these results, we may choose to use the logarithmically transformed values for some of our analysis.

**11 Conclusions**

….

# **Acknowledgements**

Acknowledge persons or institutions that helped you with the CDR here.

# **Statement**

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: Signature(s):

# **Appendix X**

If you have something to attach to your report, do it here.

# **References and Bibliography**

[1]: Kaiserschnitt (2023) Schweizerisches Gesundheitsobservatorium. Available at: <https://www.versorgungsatlas.ch/fr/indicator/_173> (Accessed: 05 October 2023).

[1.1] Betran AP, Ye J, Moller AB, Souza JP, Zhang J. Trends and projections of caesarean section rates: global and regional estimates. BMJ Glob Health. 2021 Jun;6(6):e005671. doi: 10.1136/bmjgh-2021-005671. PMID: 34130991; PMCID: PMC8208001.

[2] Betran AP, Torloni MR, Zhang JJ, Gülmezoglu AM; WHO Working Group on Caesarean Section. WHO Statement on Caesarean Section Rates. BJOG. 2016 Apr;123(5):667-70. doi: 10.1111/1471-0528.13526. Epub 2015 Jul 22. PMID: 26681211; PMCID: PMC5034743.

[3] Office fédéral de la statistique (2022) Hôpitaux, Office fédéral de la statistique. Available at: <https://www.bfs.admin.ch/bfs/fr/home/statistiques/sante/systeme-sante/hopitaux.html> (Accessed: 16 October 2023).

[4] BAG, B. für G. (2022) Kennzahlen der Schweizer Spitäler, Bundesamt für Gesundheit BAG. Available at: <https://www.bag.admin.ch/bag/de/home/zahlen-und-statistiken/zahlen-fakten-zu-spitaelern/kennzahlen-der-schweizer-spitaeler.html> (Accessed: 16 October 2023).

[5] BAG, B. für G. (2023) Qualitätsindikatoren Fallzahl, Bundesamt für Gesundheit BAG. Available at: <https://www.bag.admin.ch/bag/de/home/zahlen-und-statistiken/zahlen-fakten-zu-spitaelern/qualitaetsindikatoren-der-schweizer-akutspitaeler/qualitaetsindikatoren-fallzahl> (Accessed: 16 October 2023).