University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Valerija Jerina

Creating prediction models for cervical cancer forecasting

Mathematical Statistics

Bachelor's Thesis (9 EAP)

Supervisor: PhD Raivo Kolde

Tartu 2022

CREATING PREDICTION MODELS FOR CERVICAL CANCER FORECASTING

Bachelor thesis

Valerija Jerina

Abstract

The aim of this bachelor's thesis is to create prediction models for cervical cancer (ICD-10 C53) and pre-cancerous condition (ICD-10 R87.613) forecasting. The analysis is based on health data of 10% of Estonian population that was provided by STACC OÜ. The thesis gives an overview on cervical cancer, shows which prediction models were created using different machine learning algorithms, evaluates their performance, and gives an overview on factors that might affect risk of getting the diseases.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: prediction model, risk model, machine learning, logistic regression.

ENNUSTUSMUDELITE LOOMINE EMAKAKAELA VÄHI PROGNOOSIMISEKS

Bakalaureusetöö

Valerija Jerina

Lühikokkuvõte

Bakalaureusetöö eesmärk on luua ennustusmudelid emakakaelavähi (ICD-10 C53) ja sellele eelneva oleku (ICD-10 R87.613) prognoosimiseks. Analüüs põhineb STACC OÜ terviseandmetel, mis hõlmavad 10% Eesti rahvastikust. Töö annab ülevaate emakakaelavähist ja näitab, millised ennustusmudelid on

loodud, kasutades erinevaid masinõppe algoritme. Samuti hinnatakse mudelite jõudlust ning antakse ülevaade faktoritest, mis võivad mõjutada haigestumise riski.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: ennustamismudel, riskimudel, masinõpe, logistiline regressioon.

Contents

Introduction

About 160 women are diagnosed with cervical cancer in Estonia annually, and approximately 60 of those cases have the lethal outcome (Eesti Haigekassa, 2020). This is a problem not only in Estonia — it is considered to be the most common cancer type in 23 countries and fourth most common cancer type among women all over the world. It is believed that there is a strong connection between human papillomavirus and malignant neoplasm of cervix. This is the reason for screening researches for women in numerous countries, however, those are very random, since women are invited based on age only. (World Health Organisation, 2022a)

The main purpose of this thesis is to create prediction models for cervical pre-cancerous condition (HSIL) and cervical cancer to determine which factors might affect the risk of getting those diseases. Another goal was to offer better solutions for screening invitation system.

This thesis consists of two main parts — theoretical and empirical. The theoretical part gives an insight on the cervical cancer problem and explains which methods are used for its prevention. Moreover, this part contains information on how risk models are created, what is done during training phase, which machine learning algorithms are used and how prediction models can be validated. Additionally this part contains a data overview, explaining which format of data is used and why.

Empirical of this thesis consists of different stages of model creation. It contains the base scenario used in the research, and the course of actions used for the model training and the prediction parts. Analyses results based on the methods provided in the theoretical section, factors that might affect the diagnoses of interest interpretation, and short overview of the future work are also provided in this part.

The author would like to thank this thesis supervisor Raivo Kolde for help with work structure, result interpretation, general help and advises. The author is also grateful to Marek Oja for technical support and all answered questions.

1 Problem overview and related theory

1.1 Problem overview

According to the World Health Organization, cancer is a leading cause of death worldwide. It led to nearly 10 million deaths in 2020, or nearly one in six deaths in total that year. The most common cancer types are breast, lung, colon, rectum, and prostate cancer. However, cervical cancer is considered as the most common in 23 countries, and it is the fourth most common cancer type among women worldwide (World Health Organisation, 2022a). In Estonia, cervical cancer is the second most common incidence of gynecological malignancies. Every year, an average of 160 women in Estonia are diagnosed with cervical cancer, and about 60 of them die from it annually (Eesti Haigekassa, 2020).

It is known that cancer has some triggers, such as tobacco use and high body mass index. When it comes to cervical cancer, it is known that types 16 and 18 of human papillomavirus (HPV) are responsible for approximately half of the high grade cervical pre-cancers and more than 95% of cervical cancers. Another risk factor is the human immunodeficiency virus (HIV). The researches has shown that about 5% of cervical cancer incidents are referable to HIV. So it is believed, that cervical cancer can be theoretically prevented. (World Health Organisation, 2022b)

Right now, there are three types of cervical cancer prevention [1]:

- primary prevention, or HPV vaccination;
- secondary prevention, or screening and treating precancerous lesions;
- tertiary prevention, or invasive cancer treatment and palliative care.

The primary prevention is understandable — prophylactic vaccine against HPV opens up the possibility of preventing and reducing the incidence of cervical can-

[1] Cervical knowledge repository

cer. In a situation where the incidence of cervical cancer is increasing among young women in developed countries (such as Estonia), vaccination against HPV is beneficial for everyone. As an example, some countries promote vaccinating against the HPV, especially among people from 9 to 26 years old. Tertiary prevention in also logical — those, who already got the disease, should be treated and receive palliative care. But when it comes to screenings, many questions start to appear, such as who has to be examined and how often.

1.1.1 Cervical cancer screening in Estonia

This section is based on "Emakakaela sõleuuring" (2020) manual.

Cancer screening is considered to be one of the most efficient ways to detect cervical cancer and precancerous conditions, giving the possibility of early diagnosis and treatment. According to the data of different countries, well-organized screening might help to reduce morbidity and mortality by 80%.

Every year, Estonian Cancer Screening Register identifies women in the target group for cervical cancer screening by linking data from the Population Register and the Estonian Health Insurance Fund databases. Those cohorts include women from 30 to 65 years old, with an interval of 5 years (meaning, that women who are 30, 35, 40, ... , 65 years old are a part of cohort). The target cohort is considered to be all belonging to the cohort of the respective year. The Cancer Screening Register creates a list of women who could be invited to be a part of the target group (excluding women with specific diagnoses) and in the first half of the year mails $\frac{1}{5}$ of them monthly with an invitation to the screening. If the result of HPV test was not received by august, they resend an invitation and if the woman did not appear till the end of January following the survey year, the corresponding entry in the list is made (that woman did not take part in it). In case of positive test result, a fluid-based gynecocytological test shall be performed in the laboratory on the same biomaterial. If the answer appears to be that no cancer or other abnormal

cells have been found on the surface of or in the tissue that lines the cervix, the woman is informed that the HPV-test is positive and asked to do the second test. If the test result appears to be atypical squamous cells of undetermined significance, then the woman is also informed about this and asked to do the second test in 12 months. Another result, which requires specific treatment, is a low-grade squamous intraepithelial lesion — a colposcopy must be performed.

In order to predict cervical cancer, cytological examinations are made. Currently, there are two known types of examinations — liquid-based cytological test method and Pap-test. Pap-test is a conventional cytological examination, which was invented over 80 years ago. To conduct this examination, a gynecological mirror is inserted into the vagina. Then a sample of epithelial cells is taken from the surface of the cervical canal using special instruments for that. The sample received fixated on the glass using alcohol, then it has to be dried out and researched in a labor under microscope. The information they are looking for are any cell changes which can be considered as cancerous or pre-cancerous [2]. The possible results of the test can be:

- NILM — negative for intraepithelial lesion or malignancy;
- ASCUS — atypical squamous cells of undetermined significance;
- ASC-H- — atycipal squamous cells, cannot exclude HSIL;
- AGC-FN — typical glandular cells of undeterminal significance;
- LSIL — low-grade squamous intraepithelial lesion;
- HSIL — high-grade squamous intraepithelial lesion;
- AIS — adenocarcinoma in situ, which is considered as cancer early stage.

The only normal test result is considered to be NILM. ASC-US is the most common pap-test finding, where cells do not look completely normal, but the cause of it

[2] Emakakaela vähi sõeluuring – PAP test, Tartu Ülikooli Kliinikum

can be way to different from HPV. Then comes AGUS, which means that some cells look not normal and this might be a sign of a more serious disease. LSIL means that there are low-grade changes caused by HPV. When test outcome is ASC-H, meaning that abnormal squamous cells were found and HSIL can not be excluded. HSIL itself means that there is moderate or severe amount of abnormal cells persistent and can easily become a cancer if not treated. Last but not least, AIS is an advanced lesion, can easily become cancer as well.

1.2 About risk models

A risk model is a statistical method which uses patient risk factor data for allocation of a probability of developing a future negative outcome in a given time period to a certain individual (Whittemore, 2019). These models mostly created using machine learning, time series regression, or curve and surface fitting approaches. No matter which way the model is created, every single predictive risk model has to go through steps like: data cleaning, identification of the approach (parametric or non-parametric), data transformation into the suitable form, specifying data subset to be used for model training and so on.

In order to understand what has to be done in those steps base case scenario has to be written. Scenario is a brief narrative or story which is used to hypothetically describe the situation, where the rationale behind sampling is explained for this model. The descriptive situation has to contain information such as what is going to be predicted, on whom it is done, how it is done and what is the purpose of this prediction. It is affecting the model a lot — how the result is going to be defined, which people are going to be a part of the target cohort, what model is going to be used and which features and attributes are going to be meaningful, what is the desired goal and what is the purpose of this model.

1.3 Training phase

In order to run analyses, some specific things have to be defined — target cohort, outcome cohort, time-at-risk, concept sets and lookback period. In order to create those, a web-based open source application ATLAS has been used. Important, that ATLAS requires data to be in OMOP CDM format.

As it has been shown in the Book of OHDSI (2019), Figure 1 illustrates the prediction problem that has to be solved. It states that to define a prediction problem, t=0 is defined by target cohort, the outcome of interest by outcome cohort and time-at-risk window.

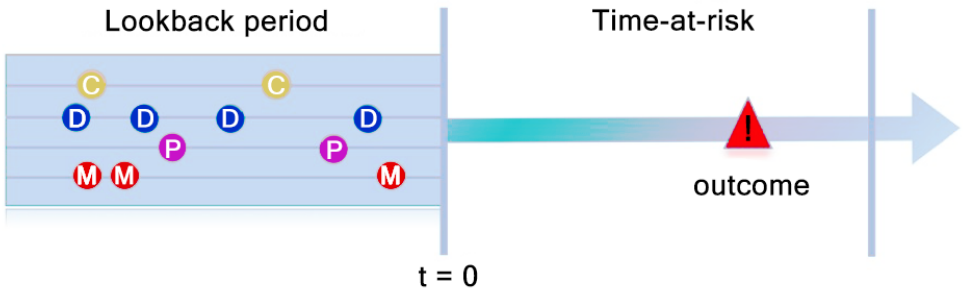

Figure 1: The prediction problem.

According to the Book of OHDSI (2019), a cohort is a specific set of persons where each patient satisfies the necessary (one or more) inclusion criteria during a certain time period. Unlike many other, in OHDSI every single cohort is defined independently, which allows using them multiple times (in the same study or in different ones). Additionally, what differs in OHDSI, is that a person can belong to the same cohort for multiple different time periods and a specific person can not belong to the same cohort multiple times during the same time period. When defining a cohort, it has to be explained how a person enters and exits the cohort.

When defining the target cohort, the set of subjects of interest is taken (for whom the prediction is going to be made). Following clarifications can be included: cohort entry events that consist of initial events and their restrictions, general cohort entry

rules. Basically, this information shows which attributes are required to be included in a cohort and what attributes make patients unsuitable. The time when the person enters the target cohort is denoted as $t = 0$ and is also known as cohort index date.

Outcome cohort is a cohort representing the relevant outcome. In other words, the outcome's explanations must be determined: it can be some disease, test results or any result of interest.

Time-at-risk is a time when the risk of the outcome is taken into consideration. Basically, time-at-risk shows in which time window relative to $t = 0$ prediction is done. It might be any period of time — from 1 day to many years.

Moreover, the models suitable for the prediction have to be defined — there might be many of them or only one. When it comes to the algorithms used, desired predictor variables are shown — these are going to be looked for in the observation window.

If features (predictor variables) corresponding to a specific topic only (as an example, pregnancy) want to be included, there is a possibility to create concept sets. Every single concept set contains concepts in itself that, taken together, describe a specific topic. All the concepts are taken from standardized vocabulary.

Minimum lookback period defines the minimum baseline period, the minimum number of days before the cohort start date that a patient has been continuously observed. The default value given by OHDSI is 365 days. It can be expanded, and will complete the picture of a patient, but the amount of patients who will meet this requirement will be smaller.

1.4 Machine learning and prediction models

This section and following subsections are based on "The Book of OHDSI" (2019).

Machine learning is a method of data analyses that automates analytical model building ("Machine Learning: What it is and why it matters"). Machine learning

gives computers an ability to learn from data on their own, without any human interactions, so it makes it possible to analyze big amount of data automatically.

A model performed will try to find a decision boundary that will optimally separate outcome classes. Different learning techniques lead to different decision boundaries, thus it is good to compare results from a few of them. Nonetheless, very complex models increase a risk of overfitting. Since not all of the data can be seen, overfitting might negatively impact the generalizability of the model, and it would not work very well to the unseen data.

In the beginning, 3 different models were taken to choose from. All the models can give different results, thus, it has to be checked which one of them is going to have the best prediction ability.

1.4.1 Constructing features

Since it was already mentioned what is a target and population cohort, it has to be known how these are created. In order to understand, which model is going to give the best result possible, features have to be properly constructed.

When running analysis in ATLAS, features are created based on covariate settings chosen. Covariate settings page consists of multiple tables with various choices, where covariates of interest can be chosen (nothing can be chosen if data does not include any subject of interesint). The first table seen consists of demographics — such as gender, age, race, ethnicity and so on. All further tables are connected to the time bound covariates. At first, time windows can be defined — long term (standard value is -365 days), medium term (standard value is -180 days), short term (standard value is -30 days), end days (standard value is 0 days), however all of these can be changed. Then, health-related covariates can be chosen: condition, drug, procedure, measurement (analysis), observation, devices. For every single one of those it has to be chosen whether that specific field is interesting in a long term, medium term, short term or any time prior. Those time windows create zero-one

vectors and check if the procedure (or measurement, or anything else) was been recorded within the desired time period. If long term is taken as -365 days and the subject of interest is procedure made, a null vector will be created, and the model will look for the procedures made during one year prior cohort index date.

Covariate settings is the place, concept sets can be included. It can be left blank if everything has to be included, or a concept set can be created in order to include only those covariates relative to the desired topic. Also, some concept sets can be excluded if needed.

1.4.2 LASSO Logistic Regression

Logistic regression appears to be a most efficient and simplest method to solve linear classification problems (Subasi, 2020).

LASSO (the least absolute shrinkage and selection operator) logistic regression is kind of a generalized linear model, which learns a linear combination of the variables and then a logistic function evaluates the linear combination to a value between 0 and 1. Cost, defined as the sum of the absolute values of the linear combination of the coefficients, is taken and then, depending on a model complexity, this cost is added to the objective function. In order to make the cost as small as possible, the model automatically selects features.

According to Hastie et al. (2009), LASSO regression is a linear regression that uses shrinkage. Shrinkage is the reduction of the effects caused by sampling variation. The vector of input variables is defined as $(x_1, ..., x_p)$ and the vector of the response as $y = (y_1, ..., y_k)$. For each $j = 1, ..., k$ the coefficients are trained by minimizing a regularized mean squared error (MSE) objective function.

Since logistic regression is a transformed form of linear regression, its entity can be

written as:

$$\begin{aligned} logit(\text{probability of getting } y_j) &= log\frac{\text{probability of getting } y_j}{1 - \text{probability of getting } y_j} \\ &= \beta_{0j} + x_1\beta_{1j} + ... + x_p\beta_{pj}, \end{aligned}$$

where logit function is the quantile function associated with the standard logistic distribution, $i = 1, ..., p$, x_i is an independent variable, $j = 1, ..., k$, y_j is referred as dependent variable and $\beta_{0j}, \beta_{1j}, ..., \beta_{pj}$ are LASSO regression estimates.

If the occasion of interest is denoted as y_j and it's probability as $P(y_j = 1)$, the estimated prediction score will have the following entity:

$$P(y_j = 1) = \frac{exp^{\beta_{0j}+x_1\beta_{1j}+...+x_p\beta_{pj}}}{1 + exp^{\beta_{0j}+x_1\beta_{1j}+...+x_p\beta_{pj}}}$$

In the training stage, $(x_{11}, ...x_{1p}, y_{11}, ..., y_{1n}), ..., (x_{n1}, ...x_{np}, y_{n1}, ..., y_{np})$ are taken as a set of training data, where the matrix X has to be of size $n \times p$ and contains observations of independent variables (x_{ij}). It has to be considered, that the estimate in LASSO is defined as

$$\hat{\beta}^{lasso} = \text{argmin}_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \text{ where } \sum_{j=1}^{p} |\beta_j| \leq t.$$

And then, for each response $(y_j, j = 1, ..., k)$, the LASSO problem in Lagrangian form is:

$$\hat{\beta}^{lasso} = \text{argmin}_\beta \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_{ij}x_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\sum_{j=1}^{p} |\beta_j|$ is L_1 LASSO penalty and λ is the amount of shrinkage.

When performing LASSO logistic regression, the starting variance of prior to distribution has to be chosen, where typical value considered to be 0.1. The starting

variance will not change model performance too much, nonetheless, choosing variance value too far from optimal will cause long time of model fitting.

1.4.3 Random Forest

The decision-tree is a flowchart that has a stricture similar to a tree, where the internal node means a test on an attribute, branches represent possible outcomes of the test that was performed, and terminal nodes hold class labels (Gupta, 2021).

Random forest (RF) is a bagging ensemble technique which combines multiple decision trees. Sometimes a problem of overfitting can be faced and in order to reduce the probability of this problem weak classifiers are used and combined into strong ones. RF achieves this by training several decision trees yet using only a subset of the variables, which differs from tree to tree.

The parameters in decision trees are the number of maximum levels in a tree (typically 4, 10, 17), number of features in a tree (typically 5, 20 and square root of total features) and the number of trees wanted to see (typically 500).

According to the Hastie et al. (2009), Random Forest Algorithm for Regression is:

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data. A bootstrap sample is a smaller sample, that is taken from a larger sample. Bootstrapping is a statistical method for studying the distribution of statistics of probability distributions, which is based on multiple sample generations using Monte-Carlo method on the original sample. Since bootstrapping is based on the law of large numbers, re-sampling is going to approximate our data to the true population data.

   (b) Grow a random-forest tree T_b to the bootstrapped data. To do this, for each terminal node the recursion has to be done, until n_min (minimal node size) is going to be reached. Recursion consists of:

i. Randomly select m variables from p variables.

ii. Pick the best one among m.

iii. Split the node into two daughter nodes.

2. The output is going to be the group of T_{b1}^{B} trees.

And finally, if prediction is wanted to be made at some point x, the following is done:

$$\hat{f}_{(rf)}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

1.4.4 Gradient boosting Machine

Boosting is an ensemble learning method, and in the framework used it combines multiple decision trees. In boosting, decision trees are added iteratively, and extra weight is added to the data-points that were misclassified by previous decision trees in the cost function when training the next tree.

Parameters such as the boost learn rate (typically 0.005, 0.01, 0.1), maximum levels in a tree (typically 4, 10, 17), minimum data points in a node (typically 2), number of trees (typically 100, 1000) and the amount of rounds after which model will be stopped if no improvement has been made (typically 25).

According to Hastie et al. (2009), the generic gradient boosting algorithm for given training set $(x_i, y_i)_{i=1}^{n}$ and number of iterations M for regression is:

1. Initialize $f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$, where $f_0(x)$ is a constant function (meaning that the output has to be the same for every input value), L is a loss function and *argmin* means the argument value, where expression reaches its minimum and y_i represents the observed value. A loss function is a function that represents values of one or more variables (or some event)

as a real number, where this real number means the intuitive cost of the corresponding event.

2. For $m = 1$ to M:

   (a) For $i = 1, 2, ..., N$ compute

$$r_{im} = -\left[\frac{\sigma L(y_i, f(x_i))}{\sigma f(x_i)}\right]_{f=f_m-1},$$

   where basically, all possible pseudo-residuals are found.

   (b) Fit a regression tree to the targets r_{im}.

   (c) Compute

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma), \text{ where } j = 1, 2, ..., J_m,$$

   which is also known as the optimization problem. γ_{jm} is optimised γ, where the Loss Function has a minimum.

   (d) Then, $f_m(x)$ is updated as:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

3. Receive the output $\hat{f}(x) = f_M(x)$

1.5 Validation and related terms

The following section has been written based mostly on Janssens and Martens (2018).

The model performance has to be evaluated to see how well a prediction model can estimate risks. In order to do so, the predictive ability and discriminative accuracy

of the model have to be known.

Models' predictive ability is specified by risk distribution, negative and positive predictive values at possible risk threshold. Discriminative accuracy, however, is designated by ROC curve, AUC, specificity and sensitivity for specific risk thresholds. When it comes to discriminative accuracy, it is needed to show how well the model can differ patients and nonpatients.

A risk distribution shows all possible risk values among our population. In order to receive higher predictive ability, more variation of predicted risk is needed.

Positive Predictive Value (PPV) shows how likely a person with risk higher than threshold value will get diseased.

Negative predictive value (NPV) shows how many people are not getting the disease among those whose predicted risk is below the threshold.

In prediction models, sensitivity is the percentage of patients whose predicted risk is above the threshold value. Specificity, on the other hand, is the percentage of nonpatients whose predicted risk is below the threshold. When threshold is lowered, typically, sensitivity is getting higher and specificity is becoming lower. All is going to depend on the model — whether sensitivity or specificity is more important. Usually, when it comes to screening tests, high sensitivity and decent specificity are wanted.

ROC-curve is used in the medicine to describe the accuracy of diagnostic methods. Basically, it shows how well the model can make a decision whether a patient has a specific disease or not. Usually it is given as a graph where x-axis is used to show 1-Specificity and y-axis to show sensitivity. This curve shows all possible combinations of sensitivity and specificity for all possible threshold values. It has to be understood what good and bad ROC-curves are. Even a bad test can have a sensitivity equal to one if all the patients are considered diseased (actually all diseased people will be found) but in this case specificity must be really small. The same can be done otherwise - specificity taken as 100%, so all the patients

are considered as healthy, but sensitivity will drop to zero. The better test is, the higher is ROC-curve, meaning that having any specificity greater sensitivity can be achieved. So to understand, whether the test is good or not, AUC has to be used.

AUC is area under the ROC-curve. The maximum possible area under the ROC-curve is 1 (achievable with a reference value of 100% sensitivity and 100% specificity). AUC equal to $\frac{1}{2}$ would show that our model is useless and AUC less than $\frac{1}{2}$ means that our data contains information about the disease, but used in a wrong way. AUC can be explained as the probability that a randomly chosen patient has higher predicted risk to get diseased than a randomly chosen non-patient. AUC is often used to describe the accuracy of a predictive model and to quantify the improvement if predictors were changed, added or removed.

2 Data overview

Data overview was based on Jäe (2021).

Database that has been used contains a random sample of approximately 10 percent of Estonian population (those with Estonian ID code). The data has been collected from Estonian Health Insurance Fund database, Recipe Center and Digilugu Patient Portal from 2012 to 2019.

Observational Medical Outcomes Partnership Common Data Model is a standard data model for analyzing health data and data from international studies that uses agreed standard terminology. It allows for the systematic analysis of disparate observational databases. Basically, all the data is divided into different tables according to its entity type. For diagnoses is used SNOMED, for drugs is used RxNorm and for laboratory measurements are used both SNOMED and international standard for laboratory tests identification called LOINC (Logical Observation Identifiers Names and Codes). As it can be noticed, they do not use terminology which is common in Estonia, however, they give a possibility to harmonize disparate coding systems with a minimal data loss.

All three databases that existed by themselves were compiled into a single dataset by scientists from University of Tartu and STACC OÜ, where all the personal data of patients has been changed into the pseudonymous form for the same person in all databases. After data has been processed, it was transferred to the PostgreSQL database by the research team according to the OMOP CDM principles. Since the presented data is given in a uniform format using common terminology and coding, it allows systematic surveys to be carried out the same way in different databases all around the world.

In order to get the amount of patients, gender distribution and other information, SQL-queries were needed. To use them in PostgreSQL database has been used DataGrip.

The dataset used includes 18 354 512 health records from 149 351 patients. Those health records include medical bills, epicrises and purchased prescriptions. Among those patients, there are 78 188 female and 70 358 male patients, and 805 are unknown.

To run analyses in R, HADES (Health Analytics Data-to-Evidence Suite) packages are used. HADES is a set of open source R packages which contains libraries for population characterization, population-level causal effect estimation, and patient-level prediction. The main package used is called PatientLevelPrediction, and it is used to build and implement patient-level predictive models (such as Regularized Logistic Regression, Gradient Boosting Machines, Random Forest, K-Nearest Neighbors, etc.) in OMOP CDM format. Another important package is Shiny, since it allows building interactive web apps straight from R — with its help, there is no need to draw graphs manually. [3]

The main data classes used are contained in tables "Person", "Observation", "Measurements", "Drug_exposure", "Condition_occurrence" and "Procedure_occurrence". Every table contains data according to its entity type. Table "Person" contains data mostly related to demographics — ID, gender, date of birth, race and ethnicity, and so on. Table "Observation" provides information on visit details, observation concept, observation type concept, etc. Table "Measurements" has data related to analyses done — value, visit details, units and analyses concepts. Information related to medicines taken is located in table "Drug_exposure", about diseases — in "Condition_occurrence", and, respectively, in "Procedure_occurrence" information about procedures made.

[3] PatientLevelPrediction by HADES

3 Model creation

3.1 Base scenario

Currently, in Estonia women from 30 to 65 years old with different health conditions (only some specific cancer and neoplasm cases are excluded) are invited to cervical cancer screenings in order to prevent as many cancer cases as possible and discover cancer an pre-cancerous conditions as soon as possible. Thus, this study aims to predict which female patients in this age range are going to have high-grade squamous intraepithelial lesion (HSIL, ASC-h or AGC-FN) test result and try to predict malignant neoplasm of cervix using LASSO Logistic Regression, Random Forest and Gradient Boosting machine. There is a possibility that no good results will be received, since the amount of cervical cancer and high-grade squamous intraepithelial lesion in our dataset is not that big. If the good results will be achieved and the model will be able to determine, which woman is most likely going to have high-grade squamous intraepithelial lesion (or malignant neoplasm of cervix), it will provide a possibility to save resources by inviting only those who appeared to be in high risk group. Moreover, this will make invitation system more precise.

3.2 Model training

While creating risk model the target and outcome cohort have to be determined.

At first, the target cohort of interest is going to be all women from 30 to 65 years old who are doing cancer screenings. Cancer developing is a really long process, it might take from couple of months to years. As an example, it is believed that for most breast and bowel cancers, the tumours begin to grow around ten years before they're detected. And for prostate cancer, tumours can be many decades old, so enough space for observation time has to be left.[4] Thus, target cohort is defined

[4] 'Science Surgery: How quickly do tumours develop?' by UK Cancer Research

using those settings:

1. The observation period is defines as any of those:
    - Starting at 01/01/2014 and ending at 31/12/2014;
    - Starting at 01/01/2015 and ending at 31/12/2015;
    - Starting at 01/01/2016 and ending at 31/12/2016;
    - Starting at 01/01/2017 and ending at 31/12/2017;
    - Starting at 01/01/2018 and ending at 31/12/2018;
    - Starting at 01/01/2019 and ending at 31/12/2019;
2. Age from 30 to 65;
3. Having any of the following occurrences:
    - A measurement of pap-test exists in our data;
    - A condition occurrence of pap-test exists in our data;
    - An observation of pap-test exists in our data.

What can be noticed is that woman can appear in our target cohort multiple times if she has occurrences connected to the pap-tests in different observation periods.

Outcome cohorts that has been chosen are:

1. Observation of high-grade squamous intraepithelial lesion, which includes HSIL, ASC-h and AGC-FN;
2. Malignant neoplasm of cervix, which includes primary malignant neoplasm of uterine cervix, exocervix and endocervix, carcinoma in situ of uterine cervix, exocervix and endocervix, and malignant neoplasm, overlapping lesion of cervix uteri.

Also, there is a possibility that a lot of covariates will need comparison and we might want to shorten them. To do this, a new concept set has been created. Concepts, that may become risk factors, have been included — anything related to pregnancy, abortion, delivery, contraception diagnoses, contraception drugs, HIV treatment diagnoses.

3.3 Prediction

The first step in creating our prediction model is to decide which machine learning model is the best to use. As the base, 3 algorithms have been used: Random forest, LASSO Logistic Regression and Gradient Boosting Machine. In order to see which one of them will give the best result possible, analyses had to be done on different covariate settings. The following covariate setting sets were chosen:

1. age groups;
2. age groups; condition, drug exposure, procedure, measurement, observation - in long term (730 days).

Also, population settings were included. Time-at-risk window start was defined as 0 days and time at risk window end as 365 days. Minimum lookback period applied to cohort was chosen as 365 days and subjects without minimum time at risk (90 days) have been removed, also patients who have observed the outcome prior to cohort entry have been removed.

All non-zero values were shown in a Table 1. First of all, it has to be mentioned that random forest did not give any results at all — it might have happened because of the small sample size or the algorithm itself — since it usually tries to treat features as equivalent, and many possibly have null-values. It can be noticed, that Gradient Boosting Machine gives slightly better results for malignant neoplasm of cervix. Although, the difference is minimal — about 1.5%. However, for observation

Table 1: Patient Level Prediction results, ages 30-65.

Model	**Covariate Settings**	**AUC**
Observation of high-grade squamous intraepithelial lesion		
LASSO Logistic Regression	2	0.6715
Gradient Boosting Machine	2	0.631
LASSO Logistic Regression	1	0.5937
Malignant neoplasm of cervix		
Gradient Boosting Machine	2	0.6587
LASSO Logistic Regression	2	0.6432

of high-grade squamous intraepithelial lesion the LASSO logistic regression gives better results — having the same covariate set the result is 4.05% better. Also, it can be noticed, that covariate setting set 1 did not give many results and the only result received was not very good. Since the difference between two models for the observation of high-grade squamous intraepithelial lesion group is greater than the difference in the second group, it was chosen to stick to the LASSO Logistic Regression.

Table 2: Patient Level Prediction results, ages 30-65, with covariate set.

Model	**AUC**
Observation of high-grade squamous intraepithelial lesion	
Lasso Logistic Regression	0.55364
Gradient Boosting Machine	0.5005
Malignant neoplasm of cervix	
LASSO Logistic Regression	0.55634

In order to see if covariate set that has been created will improve our model, analysis was done, but now using only covariate setting set 2. The results received are shown in a Table 2, all null results were hidden. Noticeable, that results with concept sets are way worse than without them. LASSO Logistic Regression gives better results than Gradient Boosting Machine. Again, no results with Random Forest have been received. Although, the amount of null-coefficients is way smaller than it was — among 7405 covariates 60 had some value, now only 6 out of 30 are

non-zero.

3.3.1 HSIL prediction

Taking a look at the Figure 2, first glance at the model performance can be done. Unfortunately, the model does not manage to predict really well, since the AUC value is below 0.7, however, the model is not useless — the AUC is not that close to the diagonal. Since AUC does not give any actual information about specificity and sensitivity at different thresholds, discriminative accuracy has to be estimated not only based on it.

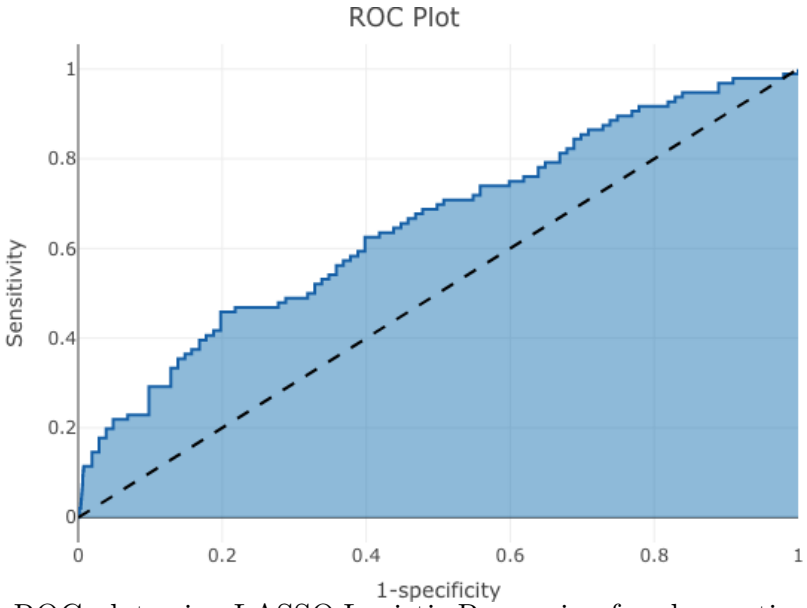

Figure 2: ROC-plot using LASSO Logistic Regression for observation of high-grade squamous intraepithelial lesion group.

Table 3 gives information about the most interesting points at Figure 2. This table contains different specificity, sensitivity and $1-$specificity values for different threshold values when predicting HSIL. According to those points, it can be stated that randomly chosen patients have higher risk to get diseased than randomly chosen nonpatients.

Tables 4, 5 and 6 contain information about different cutoff performances. They

Table 3: ROC-plot points, HSIL prediction

Risk threshold	Specificity	Sensitivity	1-Specificity
0.01157	0.8022	0.4583	0.1978
0.0086556	0.6019	0.6251	0.3981
0.00703	0.4418	0.7396	0.5582

show how many true positives, true negatives, false positives and false negatives will be obtained using different thresholds.

Besides that, Table 3 offers possible threshold values based on which people can be invited for testing. If the risk threshold will be taken as 0.01157, and information provided in Table 4 will be taken into consideration, 45.83% of diseased people will be found and only 20% of the population would have to be invited.

Table 4: Cutoff performance, threshold 0.01157, HSIL

	Ground truth positive	Ground truth negative
Predicted positive	2148	44
Predicted negative	8711	52

If risk threshold would be taken as 0.0086556, then according to the Table 5, 62.51% of patients, who actually get HSIL, will be denoted and invited 40% of the population.

Table 5: Cutoff performance, threshold 0.0086556, HSIL

	Ground truth positive	Ground truth negative
Predicted positive	4323	60
Predicted negative	6536	36

And last but not least, if risk threshold will be defined as 0.00703, then Table 6 shows that approximately 73.96% of all disease cases will be found and by observing only 56.01% of the female population. Offered values are provided for women from 30 to 65 years old.

Table 6: Cutoff performance, threshold 0.00703, HSIL

	Ground truth positive	**Ground truth negative**
Predicted positive	6065	71
Predicted negative	4794	25

According to this data, it has be told that there is a possibility to invite less people and still receive relatively good outcome. As an example, if 20% of the female population will be invited in age range from 30 to 65, 45.8% of all HSIL cases will be denoted and another 80% of the population will not be examined — meaning, that screenings will need much less resources.

Based on prediction score distribution, which is shown in Figure 3, it can be told that healthy people receive prediction score from 0 to approximately 0.02 and most of the diseased people have predicted score up to 0.025. Unfortunately, this prediction model can not distinguish between patients and nonpatients really well. However, it can be noticed that amount of diseased people who receive prediction score above 0.03 is way higher than people who are not getting diseased.

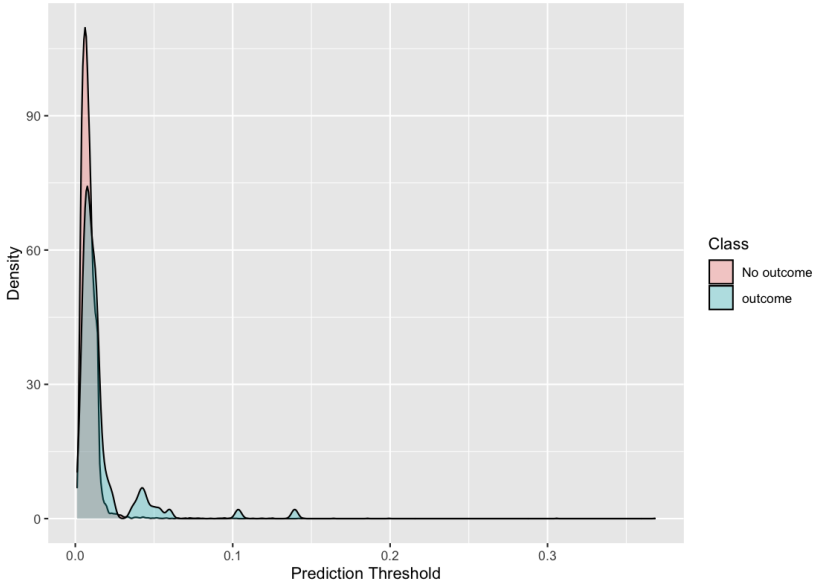

Figure 3: Prediction Threshold Score Distribution, HSIL prediction

It would be beneficial to take a look at the biggest non-null covariates, where LASSO Logistic Regression was used on covariate setting set 2, since that test gave the best results. Covariates with negative values for high-grade squamous intraepithelial lesion have been provided in Table 7. In this table covariates are a part of the input vector and give information on which drugs, procedures, observations, diseases and measurements can reduce the risk level and value is a corresponding negative β coefficient for the model. Twelve more covariates appeared to be negative for this group, however, the values were way smaller than those provided in a Table 7. Now, it is important to look at the covariates that make risk to get

Table 7: Patient level prediction, important covariates with negative value for observation of high-grade squamous intraepithelial lesion group.

Covariate group, name	**Value**
Drug exposure: Levothyroxine	−0.3779
Measurement: In-vitro immunologic test	−0.3246
Procedure occurrence: Mammography	−0.2889
Procedure occurrence: Refraction assessments	−0.2633
Procedure occurrence: Transvaginal echography	−0.2603
Procedure occurrence: Radiographic imaging procedure	−0.2447
Observation: Patient encounter procedure	−0.1988
Condition occurrence: Negative for intraepithelial lesion or malignancy	−0.187
Measurement: Glucose measurement	−0.1538
Measurement: Thyroxine (T4) free in Serum or Plasma	−0.1451
Procedure occurrence: Ultrasonography of abdomen	−0.1423
Procedure occurrence: Gynecologic examination	−0.1407
Observation: Surveillance of intrauterine device contraception	−0.1345
Drug exposure: Meloxicam	−0.1204
Measurement: Hemogram, automated	−0.1178
Observation: Admission by physician	−0.1084
Measurement: Diastolic blood pressure	−0.1043

disease higher. This data can be found in Table 8. In this table covariates have the same role as in Table 7 and values are corresponding positive β coefficients for the model. A better overview of Tables 7 and 8 is given is Section 3.3.3, where the casual link between possible covariates and HSIL is explained.

Table 8: Patient level prediction, important covariates with positive value for observation of high-grade squamous intraepithelial lesion group.

Covariate group, name	**Value**
Observation: Atypical squamous cells of undetermined significance	1.2597
Condition occurrence: Cervical intraepithelial neoplasia grade 2	0.8793
Condition occurrence: Human papilloma virus infection	0.7101
Condition occurrence: Cervical intraepithelial neoplasia grade 1	0.5908
Drug exposure: Zopiclone	0.3516
Procedure occurrence: Cervical biopsy	0.2837
Procedure occurrence: Computerized axial tomography	0.2755
Drug exposure: Glucosamine	0.1739
Measurement: Coagulation pathway screening	0.1499
Drug exposure: Diclofenac	0.1364
Measurement: Glomerular filtration rate/1.73 sq M.predicted	0.135
Drug exposure: Ethinyl estradiol	0.1223
Measurement: Neutrophil count	0.0997
Measurement: Creatinine in Serum or Plasma	0.0875
Measurement: Protein in Urine	0.0763
Measurement: Aspartate aminotransferase in Serum or Plasma	0.0625
Drug exposure: Azithromycin	0.0444

3.3.2 Malignant neoplasm of cervix prediction

The following subsection contains information on how well prediction model works for malignant neoplasm of cervix — it can be seen in Figure 4. Situation, similar to the HSIL prediction, can be noticed — model is predictive, AUC value is below 0.7 and there are some points that can give decent results using different thresholds.

All interesting points for this ROC-curve have been provided in Table 9. This table contains different specificity, sensitivity and $1-$specificity values for different threshold values, when predicting cervical cancer.

Using the information provided in Table 9 and Table 10, if risk threshold is going to be denoted as 0.00284 and people with risk higher than that will be invited (approximately 17.4% of the population), 47.4% of all diseased people will be found successfully. If denoted threshold will be 0.001483 and data provided in Table 11 taken into consideration, then model will be able to denote 73.7% of cervical

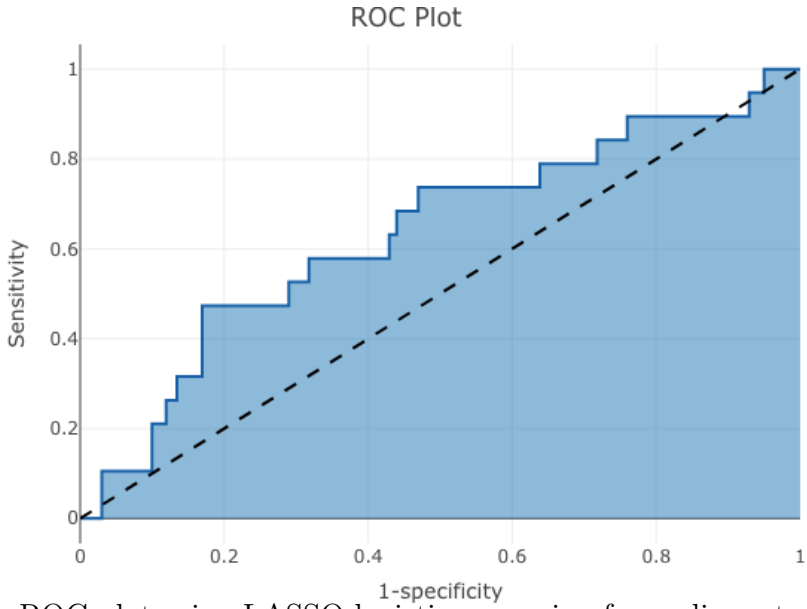

Figure 4: ROC-plot using LASSO logistic regression for malignant neoplasm of cervix.

Table 9: ROC-curve points, Malignant neoplasm of cervix prediction

Risk threshold	**Specificity**	**Sensitivity**	**1-Specificity**
0.00284	0.83	0.474	0.17
0.001483	0.53	0.737	0.47
0.0007475	0.24	0.895	0.76

cancer cases with 47.01% of the population invited. And if the threshold will be considered as 0.000747 and data provided in Table 12 will be used, about 89.5% of the malignant neoplasm of cervix cases will be found and 76% of the population will be observed. Offered values are provided for women from 30 to 65 years old. Since all of the women, who are used in this study, have visited screening research, it shows that actually inviting all the women of certain age group is not needed.

Table 10: Cutoff performance, threshold 0.00284, cervical cancer

	Ground truth positive	**Ground truth negative**
Predicted positive	1861	9
Predicted negative	9124	10

Table 11: Cutoff performance, threshold 0.001483, cervical cancer

	Ground truth positive	**Ground truth negative**
Predicted positive	5159	14
Predicted negative	5826	5

Table 12: Cutoff performance, threshold 0.0007475, cervical cancer

	Ground truth positive	**Ground truth negative**
Predicted positive	8344	17
Predicted negative	2641	2

Prediction score distribution, which is given in Figure 5, tells that healthy people receive prediction score from 0 to approximately 0.0034 and most of the diseased people have predicted score up to 0.0056.

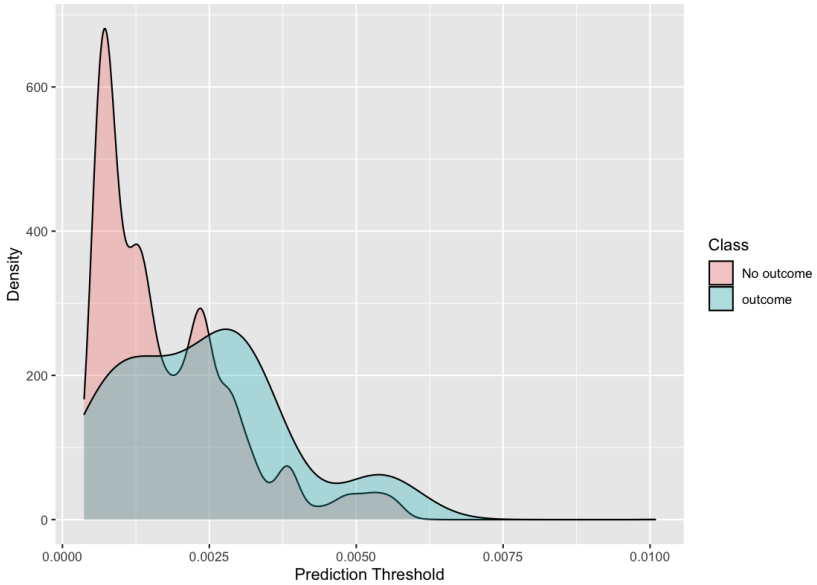

Figure 5: Prediction Threshold Score Distribution, Cervical Cancer prediction

And again, the model does not distinguish between patients and non-patients well. Although, the amount of people who have outcome and prediction score over 0.0025 is greater than those, who will not experience the outcome (will not develop cancer).

A better overview of Table 13 is given is Section 3.3.3, where the casual link between possible covariates and cervical cancer is explained.

Compared to the amount of non-null covariates received for observation of high-grade squamous intraepithelial lesion group, now there is less of them. Among 7401 covariates, only 15 appeared to be non-null. All factors that might affect predicted risk value can be found in Table 13. Covariates here are a part of the input vector, and the value is the β coefficient for the corresponding covariate.

Table 13: Patient level prediction, important covariates for malignant neoplasm of cervix.

Covariate group, name	**Value**
Procedure occurrence: Transvaginal echography	−0.653
Measurement: Microscopic cytologic examination of smear of specimen from female genital tract prepared using Papanicolaou technique	−0.5267
Observation: Admission by GP	−0.3847
Observation: Admission by physician	−0.3043
Age group: 35 - 39	−0.2051
Measurement: C-reactive protein measurement	−0.1986
Drug exposure: Ethinyl estradiol	−0.1769
Measurement: Body mass index	−0.1453
Procedure occurrence: Driver license medical examination	−0.0993
Procedure occurrence: Radiography of spine	−0.0794
Procedure occurrence: Mammography	−0.0524
Procedure occurrence: Esophagogastroduodenoscopy	0.0656
Drug exposure: diazepam	0.0648

3.3.3 Model interpretation

In order to understand, which factors might affect the risk of getting diseased, the biggest non-null covariates for both predictions are provided. Some of them will be explained, however, it is important that if a prediction model has been done using machine learning, not all of the covariates must have casual link between them and the disease. When covariate does not have a casual link with the disease of interest,

it is unknown if those are going to replicated in the other datasets as well.

Firstly, covariates with negative values for high-grade squamous intraepithelial lesion are provided in Table 7. Only 2 drug exposures are seen in this table, where the first drug is used for thyroid disease treatment (including cancer) and the second for symptomatic treatment of osteoarthritis, rheumatoid arthritis and other diseases. Secondly, the table contains a lot of measurements, such as immunologic tests, glucose and thyroxine levels, hemogram and diastolic blood pressure. Among all of them, only thyroxine level in plasma or serum is not a very common analysis, which is used to diagnose thyroid diseases. Others are more of a general tests done to determine the general health condition. When it comes to procedure occurrences, general health-related things done can be seen — such as refraction assessments and radiographic imaging. However, it is noticeable that there are procedures which are done in order to control female health — mammography, transvaginal echography, ultrasonography of abdomen and gynecologic examination. When it comes to observations, it is noticeable that there are only patient encounter procedure, surveillance of intrauterine device contraception and admission by physician. This far, it can be said that woman who generally take care of their health are less likely to get HSIL as possible outcome. What is interesting, is that those who consume tryhoid treatment drugs and get tested for tryhoid diseases are less likely to get HSIL as well. What can be noticed from other covariates that give negative result (and were not added to the table) is that age group from 45 to 49 do not get HSIL as often as others, however, the value it has was actually really small.

According to the Table 8, if the person has been observed with ASC-h (Atypical Squamous Cells, where HSIL cannot be excluded), then it means that the possibility to have the same diagnose later is really high, which is logical. Moreover, if a person has had neoplasm before (neoplasia condition occurrence) or has been diagnosed with HPV infection, it really increases risk the risk of pre-cancerous condition occurrence. However, other factors that might affect patient's condition are more interesting. Among these factors following drugs are included: zopiclone, glu-

cosamine, diclofenac and azithromycin. Diclofenac and glucosamine can be bought without using a prescription, both are often used as prophylactic medication. Diclofenac reduces inflammations and helps with joint and muscle pain, and glucosamine is used in order to reduce joint pain. Zopiclone, however, is a sedative and hypnotic drug. What unites all of these, is that they are often used by people who work hard physically. As an example those, who work as builders, often get joint and muscle pain and use drugs like diclofenac in order to reduce it, also they start taking glucosamine in order to "save" their joints. Then, those people who work hard start having sleep issues and might get prescribed hypnotic and sedative medications. Usually, those people take less care about their health since they are tired after work and consume not treating, but prophylactic medications. It might be assumed that this is how those medications are connected with HSIL occurrence. Another thing that might connect them is that some drugs have carcinogenic active substance, however, diclofenac is not considered as cancer-causing drug, and there is no evidence that glucosamine and azithromycin can be carcinogenic. According to Kripke (2018), it is believed that hypnotics can damage chromosomes, thus they can cause cancer and pre-cancerous conditions — and zopiclone is not an exception.

Taking a glance at the procedures provided in Table 8, appearance of cervical biopsy is logical. While doing a biopsy, the doctor takes a small part of body tissue in order to make a research and see if something is wrong. When it comes to cervical biopsy, it is often taken to see if and what kind of neoplasm cell present — it also might be caused by HSIL. Then, people who have done computer tomography (CT) can be in a greater risk group as well. Even though radiation exposure from CT is small, the increase in cancer and pre-cancerous condition risk from one CT scan is actually small. However, CT is also used in order to find and diagnose cancer. Thus, it is understandable, that if the patient has done tomography, there might have been tumor assumptions before. When it comes to analyses, coagulation pathway screening is done in case of blood diseases, protein in urine is used to check whether there is a problem with kidneys, aspartate aminotransferase — for liver

or/and pancreas. Other analyses are not very specific and done as a base with almost any problem — most likely there is no casual link here, however, since those analyses are actually done to denote problems and not for general check-ups, it can be assumed that those people who take less care of their health (do not treat problems, rarely do check-ups and start having more disease) are more likely to get HSIL as an outcome.

Comparing important covariates that were received for malignant neoplasm of cervix in Table 13 and the ones seen in Tables 7 and 8, it has to be mentioned that some things are actually similar — in both cases transvaginal echography and admission by physician reduce the risk of getting the disease. Noticeable, is that the second most important covariate that prevents women from getting cervical cancer is pap-test. This actually shows that if women does it, the probability of getting malignant neoplasm of cervix becomes smaller. Other factors that might reduce the risk are admissions by general practitioner, which might reduce the risk since that means that a person actually checks on his or her health and dose analyses. Noticable, that chances to get cancer for women who are 35 to 39 years old are smaller than those, who are 30 to 34 of 40 to 65. C-reactive protein in body is checked in order to check if there is an inflammation or not, kind of general analysis as well, might mean that person also takes care of the health and does checks in time. BMI check is slightly questionable. According to the Bhaskaran et al. (2014), increase in BMI index has been associated with small cervical cancer risk increase, however, it can be assumed that this is a general procedure and is often done by doctors and most of the people in Estonia do not experience obesity (20.4% adult females are obese[5]). Another factor is ethinyl estradiol exposure, which is relatively common medication that is used in birth control pills in combination with progestins. Several assumptions on why they prevent female from getting cancer can be made — one of those is that female has to visit a gynecologist in order to get the prescription. That means that most likely doctor will do the general examination and the problems

[5]Nutrition, Physical Activity and Obesity in Estonia, by WHO

will be detected on the early stage. According to Gadducci et al. (2014), the oral contraception usage increases risk of cervical cancer, but only in those, who have used it for more than 5 years and especially among HPV-positive women. However, women, who use oral contraception mostly have only one partner — thus, if they were not HPV-positive before, they will not obtain the disease and most likely the risk will not be increased. When it comes to driver license medical examination and radiography of spine, the main assumption is that both general health checks (including clinical blood tests) and an x-ray can possibly show problems on early stages. Mammography again shows that women who take care of their health are less likely to get diseased.

The last two covariates provided in Table 13 give information on what actually increases cervical cancer risk appearance. First is the procedure called esophagogastroduodenoscopy, which allows a doctor to check a person's stomach, part of small intestine and esophagus (also known as food pipe). It is hard to make a casual link between a procedure meant for digestive system disease diagnostics and cervical cancer, however, a lot of articles state that there might be a connection between them. According to the Bozdayi et al. (2019), the relationship between HPV and helicobacter pylori exists — and since it is assumed that HPV might affect the risk of getting stomach diseases, people who have done gastroscopy might have HPV — which can also lead to cervical cancer. The last covariate received is a drug called diazepam — which is also considered as sedative and hypnotic medicine. As it has been stated before, hypnotics can damage chromosomes, thus they can cause cancer and pre-cancerous conditions.

3.3.4 Future work

This bachelor's thesis gives a short overview on cervical cancer and HSIL problem and provides results on which factors might affect conditions such as high-grade squamous intraepithelial lesion and malignant neoplasm of cervix, however, received

results were not really good. Thus, the work on this topic can be continued outside of the bachelor's thesis.

In order to receive better results, different scenarios have to be simulated — as an example, they can be based on threshold values provided in Tables 3 and 9. Some specific threshold can be taken and people, whose predicted risk would be higher than chosen risk threshold, would be invited more often than those, whose risk threshold is going to be below chosen point. In order to do so, cost-effectiveness of different scenarios has to be modeled. It is going to show how big these in-between invitation periods have to be and how much money can be saved with little efficiency loss (or how to boost efficiency with the same amount spent). Based on the data received, a new model could be created, model fitting for the new data would be needed.

Another thing that can be done is applying LASSO logistic regression with same parameters on different dataset. This is a common practice in prediction models in order to avoid over-fitting of the model and control, which covariates might and might not appear outside of the dataset.

Conclusion

The aim of this bachelor's thesis was to create prediction models for high-grade squamous intraepithelial lesion and cervical cancer forecasting, using data provided by STACC, and try to suggest better solutions than inviting women once in five years (if PAP-test result has been negative) and annually, if the result has been positive. This analysis has been based on machine learning algorithms — LASSO Logistic Regression, Random Forest and Gradient Boosting Machine.

The study has showed that LASSO Logistic Regression gave the better results for predicting high-grade squamous intraepithelial lesion and slightly worse results for malignant neoplasm of cervix prediction than Gradient Boosting Machine. Random Forest, however, did not work on our dataset. Thus, LASSO Logistic Regression has been used for model creation.

Models received were not good at distinguishing between patients and nonpatients and AUC values were only tolerable, but not good. However, it has shown that people are most likely to obtain high-grade squamous intraepithelial lesion if the threshold value is above a certain point and those people can be distinguished really well. For cervical cancer outcomes, it has shown that if a person gets predicted risk above some threshold, that person will most likely get diseased, however there was still a probability that person will not experience an outcome.

The prediction has provided factors that might affect getting pre-cancerous conditions and cancer itself. Most of the covariates that reduce risk of getting those conditions are associated with regular healthcare, meaning that those women who are taking care of their health and attend doctors regularly have lower risk level. People, whose risk is higher, on the other hand take less care of their health — they have less general health-check analyses and more already disease-related check-ups. Instead of treating their diseases with needed medicines, procedures and so on, they prefer prophylactic drugs that are only useful for symptomatic help. Moreover, important information about sedative and hypnotic medicines has been received —

they increase risk of HSIL and cervical cancer.

Taking into consideration information received in this thesis, further studies on this topic can be done. It might be a big plus to run same analysis on a data from different country to make model more precise, simulate different scenarios based on evaluated cost-effectiveness and as a result offer better solutions for screening researches.

References

Bhaskaran, K., I. Douglas, and H. Forbes (2014). "Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults". DOI: https://doi.org/10.1016/S0140-6736(14)60892-8. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)60892-8/fulltext.

Bozdayi, G., B. Dinc, H. Avcikucuk, N. Turhan, A. Altay-Kocak, S. Ozkan, Y. Ozin, and B. Bostanci (2019). "Is Human Papillomavirus and Helicobacter pylori Related in Gastric Lesions?" DOI: 10.7754/Clin.Lab.2019.181244.

Gadducci, A., S. Cosio, and F. Fruzzetti (2014). "Estro-progestin Contraceptives and Risk of Cervical Cancer: A Debated Issue". DOI: https://doi.org/10.21873/anticanres.14620. (Visited on 02/06/2022).

Gupta, S. (2021). "Decision Tree". URL: https://www.geeksforgeeks.org/decision-tree/.

Hastie, T. J., R. J. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. second. Heidelberg: Springer-Verlag. URL: https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf (visited on 10/05/2022).

Janssens, A. C. J. W. and F. K. Martens (2018). *Prediction Research: An introduction*.

Jäe, R. (2021). "Ravimite täiendavate riskivähendamise meetmete rakendamise analüüsi automatiseerimine". MA thesis. University of Tartu. URL: https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=72375&year=2021.

Kripke, D. F. (2018). "Hypnotic drug risks of mortality, infection, depression, and cancer: but lack of benefit". DOI: 10.12688/f1000research.8729.3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890308/.

SAS. “Machine Learning: What it is and why it matters” (). URL: https://www.sas.com/en_us/insights/analytics/machine-learning.html.

Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*. Academic Press. DOI: https://doi.org/10.1016/C2019-0-03019-1.

Eesti Haigekassa (2020). *Emakakaela sõleuuring*.

Observational Health Data Sciences and Informatics (2019). *The Book of OHDSI*. Independently published. URL: https://ohdsi.github.io/TheBookOfOhdsi/ (visited on 04/05/2022).

World Health Organisation (2022a). “Cancer”. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 07/05/2022).

– (2022b). “Cervical Cancer”. URL: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer (visited on 07/05/2022).

Whittemore, A. S. (2019). “Evaluating Health Risk Models”. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.3991.

Non-exclusive licence to reproduce thesis and make thesis public

I, Valerija Jerina,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis
   Creating prediction models for cervical cancer forecasting,
   supervised by Raivo Kolde.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Valerija Jerina

10/05/2022