Opportunities for Automated Liver Disease Risk Prediction in the Finnish Healthcare Environment

Viljami Männikkö, Janne Tommola, Emmi Tikkanen, Olli-Pekka Hätinen, Fredrik Åberg

Submitted to: Journal of Medical Internet Research
on: June 07, 2024

Table of Contents

Opportunities for Automated Liver Disease Risk Prediction in the Finnish Healthcare Environment

Viljami Männikkö[1, 2*]; Janne Tommola[2*] MSc; Emmi Tikkanen[3*] PhD; Olli-Pekka Hätinen[3*] PhD; Fredrik Åberg[4, 5*] MD, PhD

[1]Faculty of Medicine and Health Technology Tampere University (TUNI) Tampere FI
[2]Atostek Oy Tampere FI
[3]Pfizer Oy Helsinki FI
[4]Transplantation and Liver Surgery, Helsinki University Hospital Helsinki FI
[5]Helsinki University Helsinki FI
[*]these authors contributed equally

Corresponding Author:
Viljami Männikkö
Faculty of Medicine and Health Technology
Tampere University (TUNI)
Arvo Ylpön katu 34
Tampere
FI

Abstract

Background: Chronic liver disease incidence and mortality have been rising worldwide. In many cases, liver disease is detected late in the symptomatic stage, while the earlier detection would be crucial for early initiation of preventative actions. “The Chronic Liver Disease score”, CLivD, risk detection model has been developed with Finnish healthcare data and it predicts a person's risk of getting the disease in future years.

Objective: We had two main objectives: 1) to evaluate feasibility to implement automatic CLivD score with current Kanta platform, 2) to identify and suggest the improvements for Kanta that would enable accurate automatic risk detection.

Methods: In this study, real-world data repository (Kanta) was used as a data source for “The ClivD score” risk calculation model. Our dataset consisted of 96 200 individuals whole medical history from Kanta. For real-world data utilization we designed process to handle missing input in calculation process.

Results: We found that Kanta currently lacks many CLivD risk model input parameters in the structured format required to calculate precise risk scores. However, the risk scores can be improved by utilizing the unstructured text in patient reports and by approximating variables by utilizing other health data like diagnosis information. With only utilizing structured data we were able to identify only 33 persons out of 51 275 persons to “Low risk” category and under 1% to “moderate risk” category. By adding the diagnosis information approximation and free text utilization we were able to identify 37% of persons to “Low risk” category and 4% to “moderate risk” category. In both cases we were not able to identify any persons to “high-risk” category because of the missing waist-hip ratio measurement. We evaluated three scenarios to improve the coverage of waist-hip ratio data in Kanta and these yielded the most substantial improvement in prediction accuracy.

Conclusions: We conclude that the current structured Kanta data is not enough for precise risk calculation for CLivD or other diseases where obesity, smoking and alcohol use are important risk factors. Our simulations show up to 14% improvement in risk detection when additional data sources are considered. Kanta shows potential for implementing nation-wide automated risk detection models that could result in improved disease prevention and public health. Clinical Trial: This study didn’t have any trial registration. All data utilized in this study were retrieved through Finnish authorities Findata and Kela with all required permissions.

(JMIR Preprints 07/06/2024:62978)
DOI: https://doi.org/10.2196/preprints.62978

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Viljami Männikkö, Janne Tommola, Emmi Tikkanen, Olli-Pekka Hätinen and Fredrik Åberg

Viljami Männikkö is with the Tampere University (TUNI), Tampere, Arvo Ylpon katu 34, Finland and Atostek Oy, Tampere, Hermiankatu 3, Finland (e-mail: viljami.mannikko@atostek.com)
Janne Tommola is with the Atostek Oy, Tampere, Hermiankatu 3, Finland (e-mail: janne.tommola@atostek.com).
Emmi Tikkanen and Olli-Pekka Hätinen are with Pfizer Oy, Tietokuja 4, 00330 Helsinki
Fredrik Åberg is with Transplantation and Liver Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Opportunities for Automated Liver Disease Risk Prediction in the Finnish Healthcare Environment

Background: Chronic liver disease incidence and mortality have been rising worldwide. In many cases, liver disease is detected late in the symptomatic stage, while the earlier detection would be crucial for early initiation of preventative actions. “The Chronic Liver Disease score”, CLivD, risk detection model has been developed with Finnish healthcare data and it predicts a person's risk of getting the disease in future years.
Objective: We had two main objectives: 1) to evaluate feasibility to implement automatic CLivD score with current Kanta platform, 2) to identify and suggest the improvements for Kanta that would enable accurate automatic risk detection.
Methods: In this study, real-world data repository (Kanta) was used as a data source for “The ClivD score” risk calculation model. Our dataset consisted of 96 200 individuals whole medical history from Kanta. For real-world data utilization we designed process to handle missing input in calculation process.
Results: We found that Kanta currently lacks many CLivD risk model input parameters in the structured format required to calculate precise risk scores. However, the risk scores can be improved by utilizing the unstructured text in patient reports and by approximating variables by utilizing other health data like diagnosis information. With only utilizing structured data we were able to identify only 33 persons out of 51 275 persons to “Low risk” category and under 1% to “moderate risk” category. By adding the diagnosis information approximation and free text utilization we were able to identify 37% of persons to “Low risk” category and 4% to “moderate risk” category. In both cases we were not able to identify any persons to “high-risk” category because of the missing waist-hip ratio measurement. We evaluated three scenarios to improve the coverage of waist-hip ratio data in Kanta and these yielded the most substantial improvement in prediction accuracy.
Conclusions: We conclude that the current structured Kanta data is not enough for precise risk calculation for CLivD or other diseases where obesity, smoking and alcohol use are important risk factors. Our simulations show up to 14% improvement in risk detection when additional data sources are considered. Kanta shows potential for implementing nation-wide automated risk detection models that could result in improved disease prevention and public health.

Trial Registration: This study didn’t have any trial registration. All data utilized in this study were retrieved through Finnish authorities Findata and Kela with all required permissions.

Keywords: Kanta archive, National Patient Data Repository, Real World Data, Risk prediction, Chronic Liver Disease

Introduction

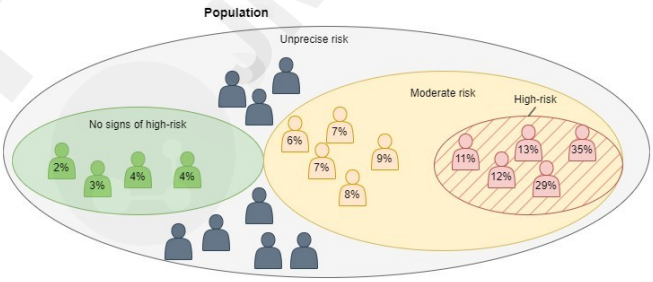
Even though healthcare risk models have been developed very long time and have been implemented to be available for individuals, healthcare still lacks automated health risk analysis because of limited real-world data (RWD). Burden of the liver disease increases yearly in Finland because the Finnish population age average grows and obesity and overweight are more common problems in the Finnish population. [8] There are an average 1 000 deaths caused by alcoholic liver disease yearly. [1] For the early detection of individuals from the general population at high risk for future severe liver disease, “The CLivD Score” was developed. It can be used to predict severe liver disease incidence in 10 years. The model itself was developed by linking data from Finnish population-based health-examination surveys (FINRISK and Health 2000) with Finnish healthcare registries. The model has been validated in data from the UK, Denmark, USA and China. [8][17][18][19]

Risk model application to real world data

If inputs are missing for risk model it will often prevent calculating the exact risk value in the way it is originally defined [8]. To be able to calculate exact risk value in many cases, would require additions to the RWD sources but also changes to healthcare professional’s work so that professionals measure correct things from the patients regularly. One problem is also the negation information which is in some cases completely missing from healthcare RWD, like a person does not smoke and person does not use alcohol. This leads to a problem where we cannot know if a person for example does not smoke if the person does not have any record of it. Despite these problems, what we can do with the healthcare RWD, and which is valuable for healthcare professionals in treatment planning, is to detect early high-risk and high-risk person based on current characteristics in RWD. Exact risk value would be easy to understand for everyday people but for healthcare professionals in preventive work and in treatment planning as important information is also to detect the potential high-risk as early as possible and detect if the person is high-risk patient. For this kind of usage, we introduce the process, which helps to detect the high-risk and potential high-risk persons. The following figure describes that kind of division in population level.

FIGURE I

RISK CATEGORIZATION BASED ON A RISK VALUE



Finnish National Electronic health record system

Kanta Services is the Finnish National Electronic health record system where data is recorded from almost all Finnish healthcare providers including the public and private sector and primary and special care. [10] Kanta Services entered production in Finland in 2010 and it has been used for over 10 years in Finland overall. [10] However, Kanta has developed in stages so different datatypes have different availabilities in Kanta. Kanta Services in general consists of four different parts: social healthcare part, prescription center, patient data repository (PDR), and personal health records. [11]

In this study, we only utilize data from Kanta PDR.

In Finland, the 2019 Act on Secondary Law made it possible to utilize healthcare data in research usage. [12] Before that, the data was available only for its primary use which is patient care. Kanta is an exceptional RWD source because it covers almost the whole Finnish population and has recordings from over 90% of all Finnish healthcare service providers. [10] This makes it possible to analyze previously developed healthcare risk models and develop new risk models by utilizing big enough data sources to represent the whole national population. Because data is produced all the time in Finnish day to day healthcare and all Finnish healthcare systems are integrated into the Kanta it is also possible to utilize developed and tested risk models in actual patient care. Popular well-being service providers have implemented automated health risk analysis based on data asked from the person regularly or by utilizing data from wearable devices like smartwatches. [20] These kinds of data still lack healthcare professionals' validation. Validation is a key element to being able to utilize automated risk analysis results in everyday patient care and as a tool for diagnosing. Kanta PDR contains only data which is validated by the healthcare professional because the data is recorded to the Kanta only by the healthcare professionals. [13] Kanta Services has also a personal health records data repository which contains data recorded by the persons themselves, but it is not utilized in this research because it is still under development and in this study, we aim to analyze only data validated by the healthcare professionals. [21]

Methods

In this study, we aim to research The CLivD score risk model automation possibilities with Kanta PDR. We use Kanta PDR as the only data source for the risk model to have an overall picture of data availability status. We consider four different scenarios of data usage possibilities, at the first we test the risk model results with the available structured data, the next we aim to utilize other structured healthcare information to approximate missing information, after that we analyze free text utilization possibilities and last we analyze completely missing input variables. For this research we have two main objectives:

1. to evaluate feasibility to implement automatic CLivD score with current Kanta platform.
2. to identify and suggest the improvements for Kanta that would enable accurate automatic risk detection.

Dataset

In this study, we utilized the dataset which consists of 192 400 persons’ medical documents archived in the Kanta PDR between 2014 and mid-2022. Data recorded in Kanta is recorded by utilizing the Clinical Document Architecture Release 2 (CDA R2) format defined by the Health Level 7 (HL7). [14] CDA R2 documents are extensible markup language (XML) documents that follow the defined format. [15] In the Finnish Healthcare environment, the local Finnish version definition from global HL7 CDA R2 is defined by Health Level 7 Finland, Kela, and the Finnish Institute of Health (THL). [14]

The dataset was picked by Kela from Kanta and the study cohort was randomly selected across the whole of Finnish population without any limitations to specific healthcare providers or locations. The dataset included all documents that were recorded to Kanta PDR after a person turned 18 years old. Documents were pseudonymized by Findata and delivered in original CDA R2 XML format to secure the environment Kapseli. Before the analysis was done the dataset was split randomly into development and validation datasets. Data was split in half equally so that development and validation datasets both contained 96,200 patients. Data was split due to the plans of usage machine learning and other methods requiring the validation in later projects, but in this

project we only used the first half of data. After data split CDA R2 XML documents must be processed so that all relevant data were parsed for the analysis. For the data processing, the separate data process library and data model were designed. Data parsing consisted of structured laboratory measurements, structured diagnosis data, structured physiological measurements, patient ongoing treatment reports free text section and some basic information about the document and patient. We did not obtain access to death records, because it is not recorded to the Kanta and would have needed separate permission and pick from Digital and population data services agency.

Risk model implementation

The original research of CLivD Score introduced two risk calculation models: *Model*lab and *Model*non-lab. Both models predict the risk of chronic liver disease for age of 40+ people. Difference between models is that *Model*lab also considers person's Gamma glutamyl transference (GGT) laboratory test result. Both risk calculators have four different exclusion criteria. [8] People who match one of the criteria there is a chance that risk calculator does not work expected way. Each exclusion criteria is introduced in Table 1.

TABLE I
RISK FUNCTION EXCLUSION CRITERIA DESCRIPTIONS

Exclusion criteria	Description
Age	Age should be above 40 and under 71.
Liver disease diagnosis	ICD-10: K70-K77, C22.0; ICD8/9: 570-573, 155.0
Chronic viral hepatitis diagnosis	ICD-10: B18
Current alcohol abstainer	Previous alcohol use, then stopped. Probably can be identified with ICD-10 -codes: F10.20 and other F10.2X

For these risk models and use cases we introduce the model for categorizing patients to four different categories: "Low risk", "Moderate risk", "High-risk" and "Unprecise risk". Categorization is based on two risk values "minimum risk" and "maximum risk". The risk calculation process itself is introduced little bit later. This approach is good also from the Kanta development point of view because when Kanta is developed further and in the future more information is added to Kanta, this risk categorizing can apply that additional data easily. The more data we get from the person the more reliable the model becomes, and the "Unprecise" -category becomes much smaller. Table 3 introduces cut-off values for each category.

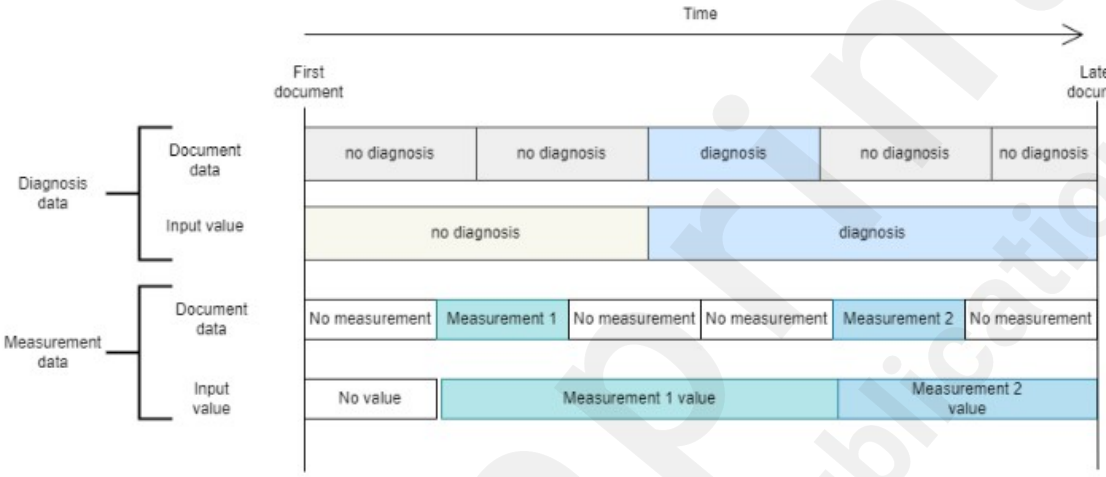
TABLE II
RISK CATEGORY CUT-OFF VALUES

	Low risk	Moderate risk	High-risk	Unprecise risk
Minimum risk	<= 5%	>5% and <10%	>=10%	<5%
Maximum risk	<= 5%	>5%	>=10%	>10%

The CLivD score risk function was developed and validated utilizing cohort studies, where the population is sampled and measured at a certain time and then followed for outcomes. [8] Kanta patient data is distributed over time, with measurements done at different points in time.

We utilize parameter lifecycles or lifetimes to do this, where we prefix the length of time a measurement or diagnosis is valid for, before and after it appears in the medical record. In this work, we used 2 different lifecycles: one-year lifecycles where measurements are valid for 1 year after measurement, and infinite lifecycles where they stay valid until the next measurement, or until the end of document history. For both lifecycle types, we used infinite validity time for diagnoses. An example of parameter lifecycles is shown in Figure 2 with finite (e.g., 1-year) lifetime for measurements and infinite lifetime for diagnoses.

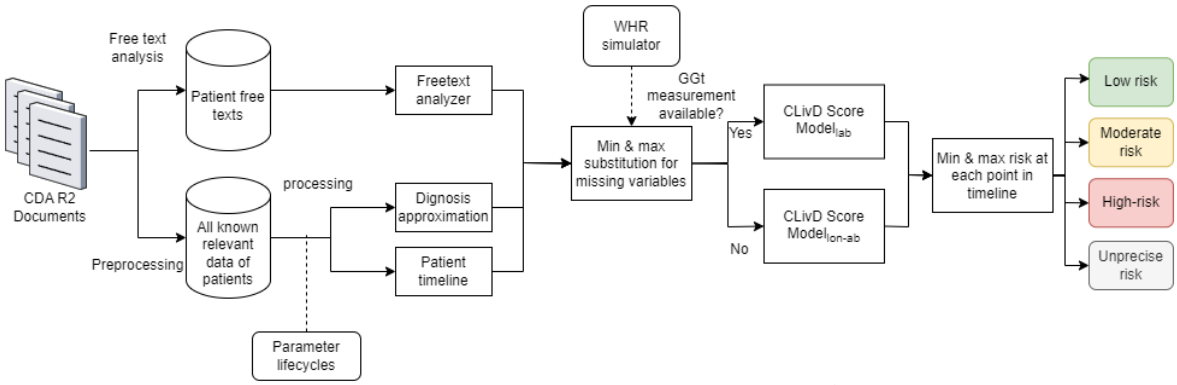
FIGURE II
EXAMPLE OF FINITE PARAMETER LIFECYCLE



We implemented the CLivD risk function in Kapseli using Python. The architecture is depicted in Figure 3. The timeline is formed by utilizing all known relevant patient data and applying the previously discussed parameter lifecycles to it. Because of missing input variables, we calculate 2 risk values: the minimum and maximum risk. The minimum risk is obtained by substituting the missing variables with their lowest value and likewise for maximum risk with their highest value.

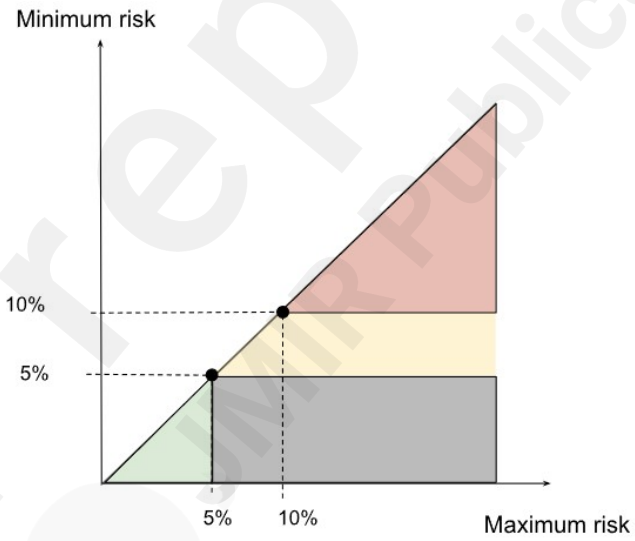
By calculating the minimum risk and maximum risk we have as a result risk range. A smaller risk range directly implies a more accurate risk value. Missing input parameters increases the risk value range. The goal is to get the risk value range narrowed towards the actual risk value in the long term. Possible ways to narrow down the risk range are for example getting the exact missing value from other data sources, asking for information from the patient, or determining the magnitude of the missing information from free text evaluation.

FIGURE III
RISK CATEGORIZATION ARCHITECTURE



In Figure 3, an illustration and example goal for the risk value ranges are shown. The line represents an ideal case where all input variables are known, and an exact risk value with equal minimum and maximum risk can be calculated. The lower turquoise triangle is an area where the maximum risk is below 5%, and the patient can be considered as low risk. The upper red triangle is an area where the minimum risk is above 10%, and the patient is at high risk. The yellow are the moderate risk area where patients' risk value is higher than 5% but not high enough to determine person as a high-risk patient. The Grey colored area represents the "unprecise risk" category where the risk range between minimum risk and maximum risk is too wide for identifying correct category.

FIGURE III
RISK CATEGORIES IN MINIMUM AND MAXIMUM RISK PLANE



In this work, we target these colored areas as a goal, as an exact risk value is not feasible with the missing data and a risk range is still helpful for determining the course of action. The risk thresholds can be configured for different use cases.

Results

Analysis were made in 4 iterations. At the first we analyzed the structural data and its occurrences. After that we tried to improve model with diagnosis utilization to determine magnitude of missing input variables. After that we utilized a free text for tobacco and alcohol information and at the last, we simulated waist-hip ratio (WHR) in cases where it would have been available from different data source.

Initial person number for the study is 96 200. When applying the Table 1 exclusion criteria to

that group of persons we get N = 51 275 suitable persons for the risk model. 44 925 persons were left outside of this analysis because they did not match to original criteria. Almost 50% of original number of persons were excluded where the age exclusion criteria where the most common reason for exclusion.

Structural data

At the first we analyze the CLivD score risk calculation results when only Kanta PDR structured data is utilized as a precise input parameters without any approximation of parameter magnitude. In Table 3, we describe how many times the relevant variables appear in a patient's medical history in a structured format. The occurrence of 1 means that the variable has been measured just once during the patient's medical history according to the Kanta data. Age and gender are registered in every document and are always available. The table shows that 0 measurements are the most common, and over 5 measurements is very rare case among risk model input variables.

WHR and alcohol use are not shown in the table as they are not present in Kanta data at all in structured format. Besides age and gender, the only input variables with at least 1% availability are GGT and fasting glucose, with around 5% and 13% total occurrence, respectively. As fasting glucose (of over 7.0 mmol/l) is only used as an alternative to a diabetes diagnosis, it is not a required input variable. We consider that a missing measurement means that the patient does not have diabetes if the other criteria are not fulfilled. BMI and waist circumference are not input variables for the risk model but were investigated as possible alternatives for missing WHR. Waist circumference measurements are very rare and so it is not considered further, while BMI has moderate availability and will be discussed later.

TABLE III

Parameter occurrence in patient Kanta medical history

Occurrence during the patient history, times	Fasting glucose	Height	Weight	Body mass index	Waist circum-ference	Gamma glutamyl transferance	Smoking
0	87%	79%	81%	83%	99%	95%	99%
1	10%	11%	9%	9%	<1%	3%	<1%
2-5	2%	8%	8%	6%	<1%	1%	0%
6-10	<1%	1%	1%	1%	0%	<1%	0%
11-15	<1%	<1%	<1%	<1%	0%	<1%	0%
>15	<1%	<1%	<1%	<1%	0%	<1%	0%

Diagnosis utilization

Because the Kanta PDR is missing WHR and alcohol usage data and tobacco usage information is quite rarely found in structured format we need to use other healthcare information found from the Kanta PDR. For the alcohol usage and tobacco usage there are some ICD-10 and ICPC-2 diagnosis codes which can be utilized in risk calculation. With the smoking diagnose codes we get the information that person is smoking and it is also exact value for the CLivD score model because it doesn't take into account how many cigarettes person smokes per week.

In alcohol usage the CLivD score model uses servings number as an input and because of that with diagnose codes we aim to have magnitude of alcohol usage and narrow down risk range. For the alcohol usage we use diagnosis codes which relates to heavy usage of alcohol, by those diagnose codes we are able to say that person have used alcohol above average for some time before getting the diagnose. The biggest ICD-10 group which will be utilized is F10 “Mental and behavioral disorders due to use of alcohol” and from the ICPC-2 codes we utilize P15 “Chronic alcohol abuse” and P16 “Acute alcohol abuse”. For high-risk alcohol users we utilize 23 alcohol servings per week approximation for men and 12 alcohol servings per week approximation for women. [16] In the table 4 we introduce statistics from diagnosis occurences in our dataset.

TABLE IV

Alcohol and smoking related diagnosis occurences in dataset

	Diagnose code	**Display name**	**Number of diagnosis between 2014 - 2021**
Alcohol related	F10.09	Unspecified alcohol intoxication	2 158
	F10.1	Alcohol abuse	18 141
	F10.20	Alcohol dependence uncomplicated	3 697
	F10.24	Alcohol dependence with alcohol-induced mood disorder	3 251
	F10.25	Alcohol dependence with alcohol-induced psychotic disorder	3 921
	F10.26	Alcohol dependence with alcohol-induced persisting amnestic disorder	3 328
	F10.29	Alcohol dependence with unspecified alcohol-induced disorder	3 268
	F10.39	Alcohol withdrawal symptoms unspecified	1 690
	P15	Alcohol abuse, chronic	4 494
	P16	Alcohol abuse, acute	2 417
Smoking related	Z72.0	Tobacco use	1102
	P17	Tobacco abuse	465

Free text analysis

Kanta PDR contains lot of free text in patient ongoing treatment report which describes the patient overall health status and living habits. Because of that the free text can be utilized for finding information concerning about alcohol usage habits and tobacco usage. Because of the limitations on available resources like lack of graphics card in secure Kapseli environment we had limited options in free text analysis. We were not able to utilize advanced machine learning models or any generic artificial intelligence for text analysis. Instead of those methods we utilized simple regex-based keyword finding and converted found text phrases in to usage categories. With this analysis our aim was to have understanding how often tobacco or alcohol related texts occur in ongoing treatment report and how it would improve the CLivD score model. For more precise free text analysis, the more advanced tools should be utilized to gain more reliable results.

Tobacco usage is simpler case because we can find texts which refers that person is smoker or

non-smoker. There are cases where there can be texts like “person has smoked 10 years ago”, where we can’t exactly say that person currently doesn’t smoke without other text context. But in the most cases we can get quite good results by finding sentences related to smoking status.

Alchol usage is different case than tobacco usage. Because alcohol servings are defined in weekly servings format at CLivD Score model we have to define alcohol usage categories also for free text analysis. Based on analysis we converted phrases in to alcohol usage categories. Category mappings are defined in table 6 ant they are based on THL alcohol risk table. [16] We noticed that alcohol usage was described in free text with multiple different ways and words. We also noticed that amount descriptions were in many cases abstract and conversion to alcohol usage category was not reliable. However, we managed to find clear and reliable cases for alcohol usage from the free text managed to utilize them. For the text finding we utilized keywords like “käytä alkohol”, “ei\s*\w*\ s*alkohol”, “runsa\w*\s*alkohol” and “vieroitus|vierotus|päihtymys|putki|katkaisu”.

TABLE V
Alcohol usage categories definitions

Risk category	Gender	Alcohol servings per week range
Absolutist	Male	0
	Female	0
Light to moderate use	Male	1-14
	Female	1-7
Moderate/Heavy use	Male	14-49
	Female	7-49
Risk user	Male	23-49
	Female	12-49

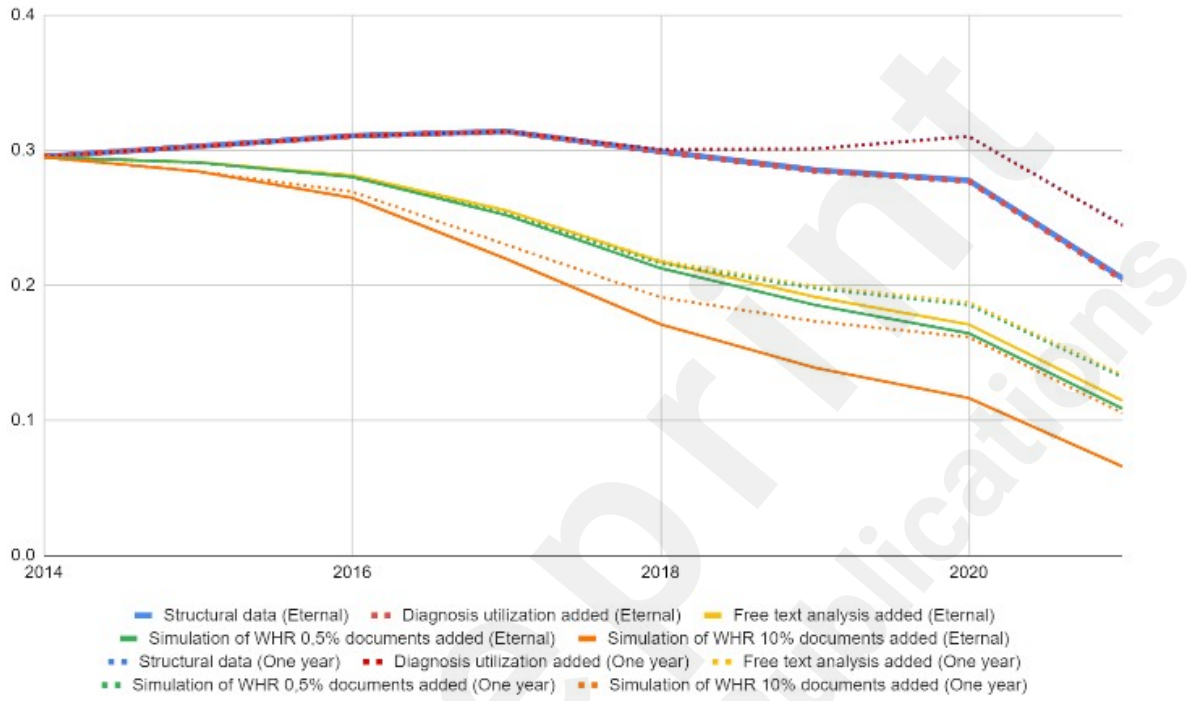
Risk calculation results

We apply the risk function to the patient timelines and attempt to assign their risks into the colored areas shown in Figure 4. These risk areas would be enough to classify patients into low- and high-risk patients, with the uncolored area being a sort of grey area where no definite conclusions can be made.

The results of our test runs risk difference average can be seen in Figure 5, with full lines representing cases where infinite timeline were used and dotted lines representing the case where one-year life time was used. Risk difference is calculated by substituting minimum risk value from maximum risk value. Smaller differences imply a more precise risk value. As we can see from the figure as the data have been developed in Kanta

the average risk difference has reduced. If we look the case where structural data, diagnosis and free text all have been utilized the risk difference has reduced almost 20 percentage points. We can also see that biggest impact for risk difference have made the free text utilization and the high availability simulation of WHR.

FIGURE V
AVERAGE RISK DIFFERENCE DEVELOPMENT BETWEEN 2014 AND 2021



In the test run we calculated for each person the most precise risk value throughout the history in Kanta and categorized the person by that risk value. Results can be seen in Table 7.

TABLE VII
RISK CATEGORIZATION WITH INFINITE LIFE CYCLE. N = 51 275

Risk category	Structural data	Diagnosis utilization added	Free text analysis added	Simulation of Waist-Hip -ratio 0,5% documents added	Simulation of Waist-Hip -ratio 10% documents added
Low risk	33	33	18 895	19 341	33 760
Moderate risk	308	310	2 125	2 351	4 495
High-risk	0	0	0	6	24
Unprecise	50 934	50 932	30 255	29 577	22 420

As we can see based on the results with only structural data, we can only categorize under 1% of persons to any category. By adding diagnosis utilization for approximation, we don't achieve notable improvement for categorization. When adding the free text analysis, we were able to categorize 41% of people. Even though we didn't manage to categorize any persons into the high-risk category we managed to identify persons who are not in high risk. After the high availability of WHR information we can achieve 75% categorization. Small number of persons in high-risk category at WHR simulation can be caused by wrong kind of WHR value for persons overall health status at simulation but it won't affect to categorization percentage. At the original CLivD score research <2% of persons were identified as a high-risk persons from 25 760 persons cohort. Approximately 3% were identified as a moderate risk in original research and in our categorization around 4% were categorized in to that category. [8] Based on these results we can say that results match because the ClivD Score development dataset were from before 2012 and our dataset were after 2014 and risk for chronic liver disease have increased in overall population in Finland.

WHR data

The waist-hip ratio (WHR) is among the most important variables for the CLivD risk score. Unfortunately, it cannot currently be obtained from Kanta data as there is no support for it in structured format and we were unable to find it in texts as well. For the analysis purposes we simulated WHR data affect to the CLivD score risk model results in few different scenarios. The first scenario was that the exact WHR data would be found from the Kanta PDR, the second scenario was that WHR data was asked from the person itself by utilizing the WHR groups and the last scenario was that person measures exact WHR. These three scenarios were created because they serve use case scenarios of the risk model differently.
Simulation was done by populating the timelines with WHR data so that in the first case 0,5% of all documents recorded to the Kanta contains the WHR data and in the
second case 10% of documents recorded to the Kanta would have contained the WHR data. The first case represents the case where WHR data would have same kind of availability than all other basic physiological measurements currently at the Kanta. The second case represents the case where the WHR data would be highly available for persons, for example from other data source. For the simulation WHR values were generated by utilizing normal distribution so that for men WHR mean was 0.96 and for women WHR mean was 0.84. For the variance the 0.07 were used for both cases. These values are found to be representative for the Finnish population based on research. [9]

WHR as a measurement is little bit complicated to measure by the persons themselves and that rises bar for the measurement and CLivD score risk model utilization. Because of that we tested the cases if the simpler WHR categorization would give good results for risk categorization. WHR categorization for the person could be easily implemented by showing images of the different body types using these types as a WHR range. For the person it would be easy to answer which image match the best for the person's own body. In this simulation populated exact WHR values were changed to corresponding WHR category. After that risk difference averages for the population were calculated for the cases; 3 categories and 5 categories. WHR categorization definitions can be found from Table 6.

TABLE VI
WHR categorization definitions

Results of the test run with WHR simulation can be found from Figure 4. As we can see from the graph, there is no big change between 3 categorical WHR data simulation and exact WHR data simulation. If the WHR were asked from the person, it would be simpler to ask body type with 3 categorical questionnaires than ask exact WHR value.

FIGURE IV
WAIST-HIP-RATIO (WHR) DIFFERENCE DEVELOPMENT

Waist-Hip -ratio category	3 categories, Waist-Hip -ratio range	5 categories, Waist-Hip -ratio range
1	<0.9	<0.82
2	0.9-1.1	0.82 - 0.94
3	>1.1	0.95 - 1.06
4	-	1.07 - 1.18
5	-	>1.19

CATEGORIZATION IMPACT TO RISK BETWEEN 2014 AND 2021

Discussion

As a result of this study, it is not possible to calculate precise risk scores with CLivD risk model based on a current data in Kanta PDR. However, risk categorization enables possibility to utilize CLivD risk model with Kanta PDR data so that it takes into account missing input parameters and enables future data development in Kanta. We noticed that the risk categorization improved when input parameters magnitude were approximated with diagnosis information and free text in patient ongoing treatment report were utilized for input parameter parsing. When using the structured data from Kanta PDR as an input we were able to identify only 33 persons out of 51 275 persons to “Low risk” category and under 1% to “moderate risk” category. When diagnosis utilization and free text analysis were added to model we were able to identify 37% of persons to “Low risk” category and 4% to “moderate risk” category. In both cases we were not able to identify any persons to “high-risk” category, because of the missing WHR data. When we added the WHR simulation to risk model we started to identify “high-risk” persons also. We evaluated three scenarios to improve the coverage of waist-hip ratio data in Kanta and these yielded the most substantial improvement in prediction accuracy.

Approximation of smoking status and alcohol usage can be extracted from free text or based on a different ICD-10 or ICPC-2 diagnosis codes. These are still approximations and have error and

especially for alcohol usage it is unusual to have exact value from free text, because alcohol usage can be described with multiple different words. Waist-hip ratio is more problematic because we don't have any signs of its magnitude from free-text or other health information.

To significantly increase risk model results precision waist-hip ratio should be got somewhere. As an alternative, we considered BMI, waist circumference, or hip circumference to predict the WHR. Of these, only BMI has meaningful availability in Kanta, so waist and hip circumferences were not considered further. BMI has been measured for around 17% of the patients.

While associations between BMI and WHR can be found in the literature [2][3], the suitability of BMI as a predictor for severe liver disease has recently been disputed [4][5][6][7], with its suitability affected by gender and potentially other factors. Due to the seemingly complex and unclear nature of BMI and WHR interaction, we decided to not pursue WHR prediction for now.

There are a few possible ways to get that information and their goodness depends on a use-case. For example, Kanta PDR could have a new measurement type and waist-hip ratio measurement could be added to general measurements in healthcare visits. It would take multiple years to get bigger coverage overpopulation to Kanta because the data is only recorded when a person uses healthcare services. The other possible way would be the ask for information from the person and utilize that in the risk model. There are a few ways to ask value from the person; exact value, categorical question, or camera measurement. The exact value asking is beneficial for the risk model because it will give the best results but the measurement is quite hard for individuals to do by themselves. Categorical questioning is easier to answer for the person and it can be implemented by showing photos of different body types and the person selecting the closest one. This method would have the most error, but it would be the easiest way for individuals. The last option would be the utilize a smartphone to measure waist-hip ratio through the camera. The body can be identified by utilizing artificial intelligence and the ratio could be calculated quite exactly based on an image. This requires a little bit of effort from the person but gives precise risk value as a result.

As an alternative, we could show the WHR ranges assigning a patient to the low- or high-risk category, or we may consider developing a new risk function utilizing BMI instead of WHR. As a potential future development, we may consider utilizing Kanta PHR where patients themselves could record their WHR or ask the categorized question from the patient about the WHR.

In future work, the time dimension could be further utilized to track a patient's risk over their whole patient history by plotting a graph of their minimum and maximum risk at each point in their timeline. This could help determine e.g., whether a change in their risk warrants an intervention or to highlight a period of missing data in their history where the risk value is more uncertain.

Conclusions

Our conclusion is that the current Kanta PDR data is not enough for precise risk calculation for CLivD Score or other risk models where obesity, smoking and alcohol information are important risk factors. Our simulations show up to 14% improvement in risk detection when additional data sources are considered for obesity. Kanta shows excellent potential for implementing nation -wide automated risk detection models that could result in improved disease prevention and public health.

Acknowledgements

This study was sponsored by Pfizer. All authors have made a substantial, direct, intellectual contribution to this study.

Conflicts of Interest

Emmi Tikkanen and Olli-Pekka Hätinen work for Pfizer Oy and own Pfizer Inc. stocks. Viljami Männikkö and Janne Tommola work for Atostek Oy. Atostek Oy provides an eHealth API Gateway service which enables healthcare and social welfare providers a fast and easy connection to Finland's national Kanta services. The publication is related to quality of data stored in the Kanta archive and improving the quality of the data could have positive impact for the Atostek's services. We do not consider this to be a conflict of interest as the improving the quality of the data in the national Kanta archive is possible only through constructive criticism of the current situation.

Abbreviations

CLivD Score: Shronic Liver Disease Score
Kanta PDR: Kanta Patient Data Repository
WHR = Waist-hip ratio
BMI = Body-mass index
ICD-10 = International classification disease version 10
ICPC-2 = International classification of primary care version 2
RWD = Real-world data

References

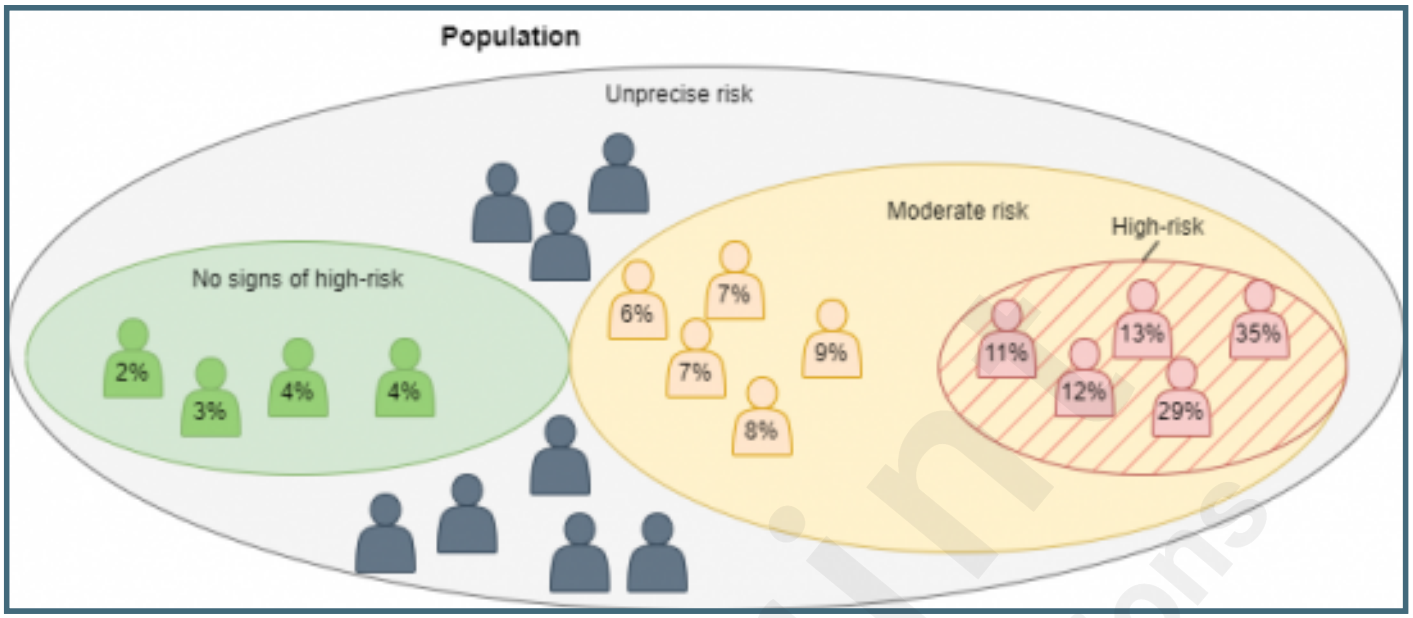
1. Fin Stat, Kuolemansyyt - Kuolleet tilaston peruskuolemansyyn (ICD-10, 3-merkkitaso), Source: https://pxdata.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin__ksyyt/statfin_ksyyt_pxt_11be.px/table/tableViewLayout1/, Accessed: 29.1.2024
2. Lahti-Koski, Marjaana, et al. "Trends in waist-to-hip ratio and its determinants in adults in Finland from 1987 to 1997." The American journal of clinical nutrition 72.6 (2000): 1436-1444.R. Fardel, M. Nagel, F. Nuesch, T. Lippert, and A. Wokaun, “Fabrication of organic light emitting diode pixels by laser-assisted forward transfer,” Appl. Phys. Lett., vol. 91, no. 6, Aug. 2007, Art. no. 061103.
3. World Health Organization. "Waist circumference and waist-hip ratio: report of a WHO expert consultation, Geneva, 8-11 December 2008." (2011).
4. Andreasson, Anna, et al. "Waist/hip ratio better predicts development of severe liver disease within 20 years than body mass index: a population-based cohort study." Clinical Gastroenterology and Hepatology 15.8 (2017): 1294-1301.
5. Schult, Andreas, et al. "Waist-to-hip ratio but not body mass index predicts liver cirrhosis in women." Scandinavian Journal of Gastroenterology 53.2 (2018): 212-217.
6. Åberg, Fredrik, and Antti Jula. "The sagittal abdominal diameter: Role in predicting severe liver disease in the general population." Obesity Research & Clinical Practice 12.4 (2018): 394-396.
7. Sahlman, Perttu, et al. "Genetic and lifestyle risk factors for advanced liver disease among men and women." Journal of gastroenterology and hepatology 35.2 (2020): 291-298
8. Åberg, Fredrik, et al. "Development and validation of a model to predict incident chronic liver disease in the general population: The CLivD score." Journal of Hepatology 77.2 (2022): 302-311.
9. Åberg F, Färkkilä M, Salomaa V, Jula A, Männistö S, Perola M, Lundqvist A, Männistö V. Waist-hip ratio is superior to BMI in predicting liver-related outcomes and synergizes with harmful alcohol use. Commun Med (Lond). 2023 Sep 6;3(1):119. doi: 10.1038/s43856-023-00353-2. PMID: 37674006; PMCID: PMC10482890.
10. Kanta, Statistics, Webpage: https://www.kanta.fi/en/statistics, Accessed 3.5.2024
11. Kanta, What are Kanta service, Webpage: https://www.kanta.fi/en/professionals/what-are-

kanta-services, Accessed 3.5.2024
12. Finlex, Act on the Secondary Use of Health and Social Data, Webpage: https://www.finlex.fi/fi/laki/alkup/2019/20190552, Accessed: 3.5.2024
13. Kanta, The data saved in Kanta are shown in MyKanta, Webpage: https://www.kanta.fi/en/data-in-kanta, Accessed: 3.5.2024
14. HL7 Finland. (2023). ”Kanta-earkiston kertomus ja lomakkeet CDA R2.”. https://www.hl7.fi/hl7-rajapintakartta/kanta-earkiston-kertomus-ja-lomakkeet-cda-r2/, Accessed 29.9.2023.
15. Dolin, Robert H. and Alschuler, Liora and Boyer, Sandy and Beebe, Calvin and Behlen, Fred M. and Biron, Paul V. and Shabo (Shvo), Amnon, HL7 Clinical Document Architecture, Release 2, Journal of the American Medical Informatics Association,vol. 13, no. 1, pp. 30–39, Jan. 2006, doi:. 10.1197/jamia.M1888.
16. Finnish institute for Health and Welfare, Alkoholin käytön puheeksiotto ja mini-interventio, Webpage: https://thl.fi/aiheet/alkoholi-tupakka-ja-riippuvuudet/ehkaiseva-paihdetyo/puheeksiotto-ja-mini-interventio/alkoholin-kayton-puheeksiotto-ja-mini-interventio, Accessed: 3.5.2024
17. Pang Y, Åberg F, Chen Z, Li L, Kartsonaki C; China Kadoorie Biobank Collaborative Group. Predicting risk of chronic liver disease in Chinese adults: External validation of the CLivD score. J Hepatol. 2024 Jun;80(6):e264-e266. doi: 10.1016/j.jhep.2023.12.022. Epub 2024 Jan 4. PMID: 38181826.
18. Åberg F, Asteljoki J, Männistö V, Luukkonen PK. Combined use of the CLivD score and FIB-4 for prediction of liver-related outcomes in the population. Hepatology. 2023 Dec 19. doi: 10.1097/HEP.0000000000000707. Epub ahead of print. PMID: 38112489.
19. Song J, Jiang ZG. A good step toward low-cost prognostication of liver-related outcome awaits more validation. J Hepatol. 2022 Sep;77(3):887-889. doi: 10.1016/j.jhep.2022.04.008. Epub 2022 Apr 20. PMID: 35460724.
20. S. T. Himi, N. T. Monalisa, M. Whaiduzzaman, A. Barros and M. S. Uddin, "MedAi: A Smartwatch-Based Application Framework for the Prediction of Common Diseases Using Machine Learning," in IEEE Access, vol. 11, pp. 12342-12359, 2023, doi: 10.1109/ACCESS.2023.3236002. keywords: {Diseases;Sensors;Machine learning;Wearable Health Monitoring Systems;Medical services;Prototypes;Machine learning algorithms;Smart devices;Healthcare;disease prediction;machine learning;smartwatch;mobile application},
21. Kanta, Kanta PHR, Webpage: https://www.kanta.fi/en/system-developers/kanta-phr , Accessed 23.5.2024

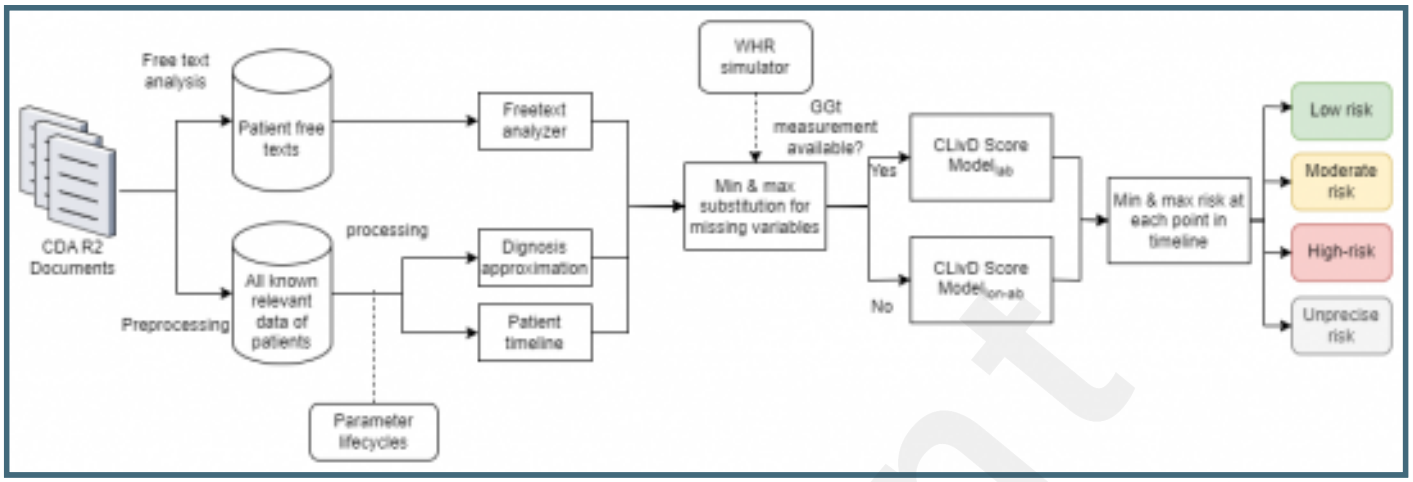
Supplementary Files

Figures

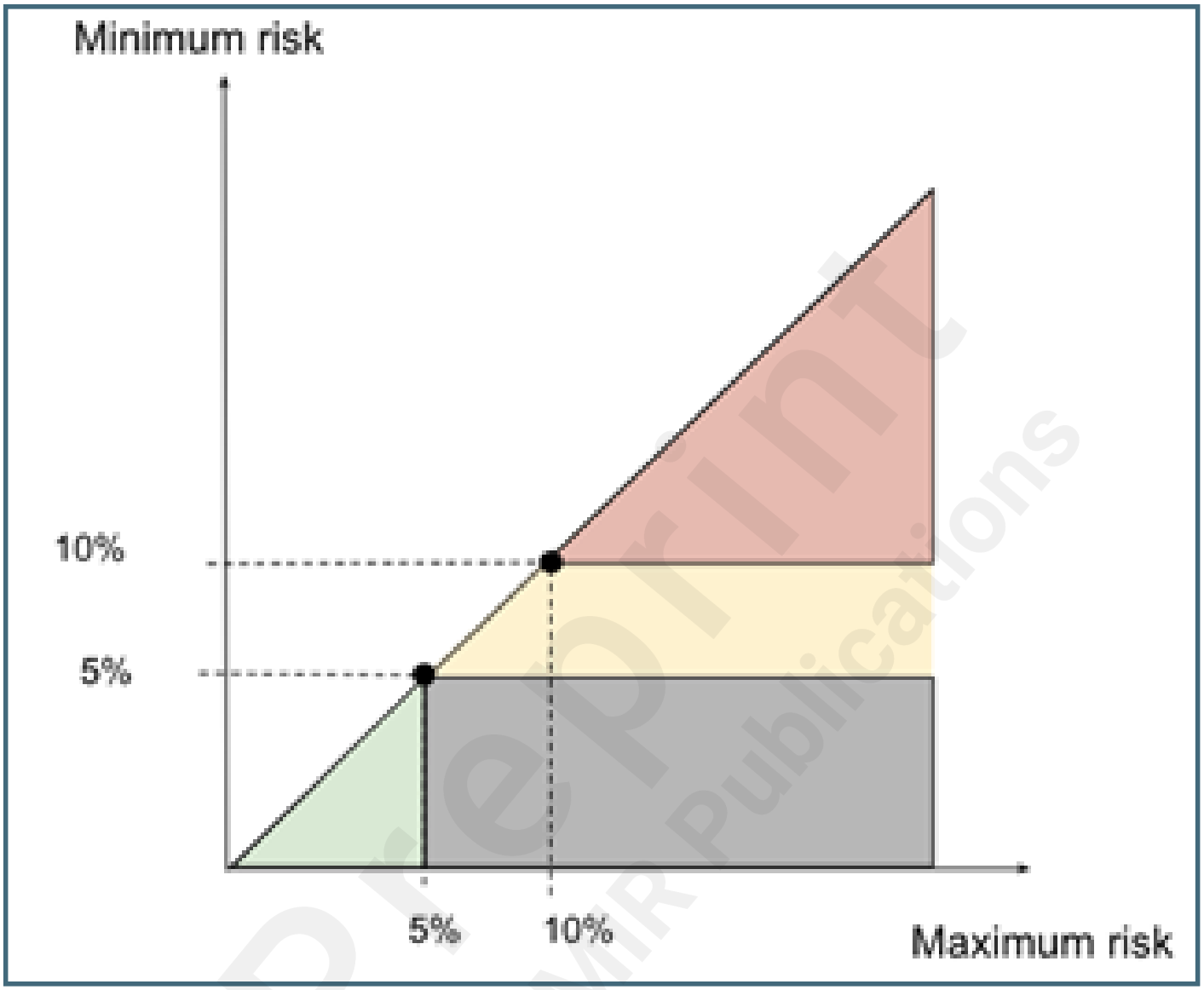
Risk categorization based on a risk value.



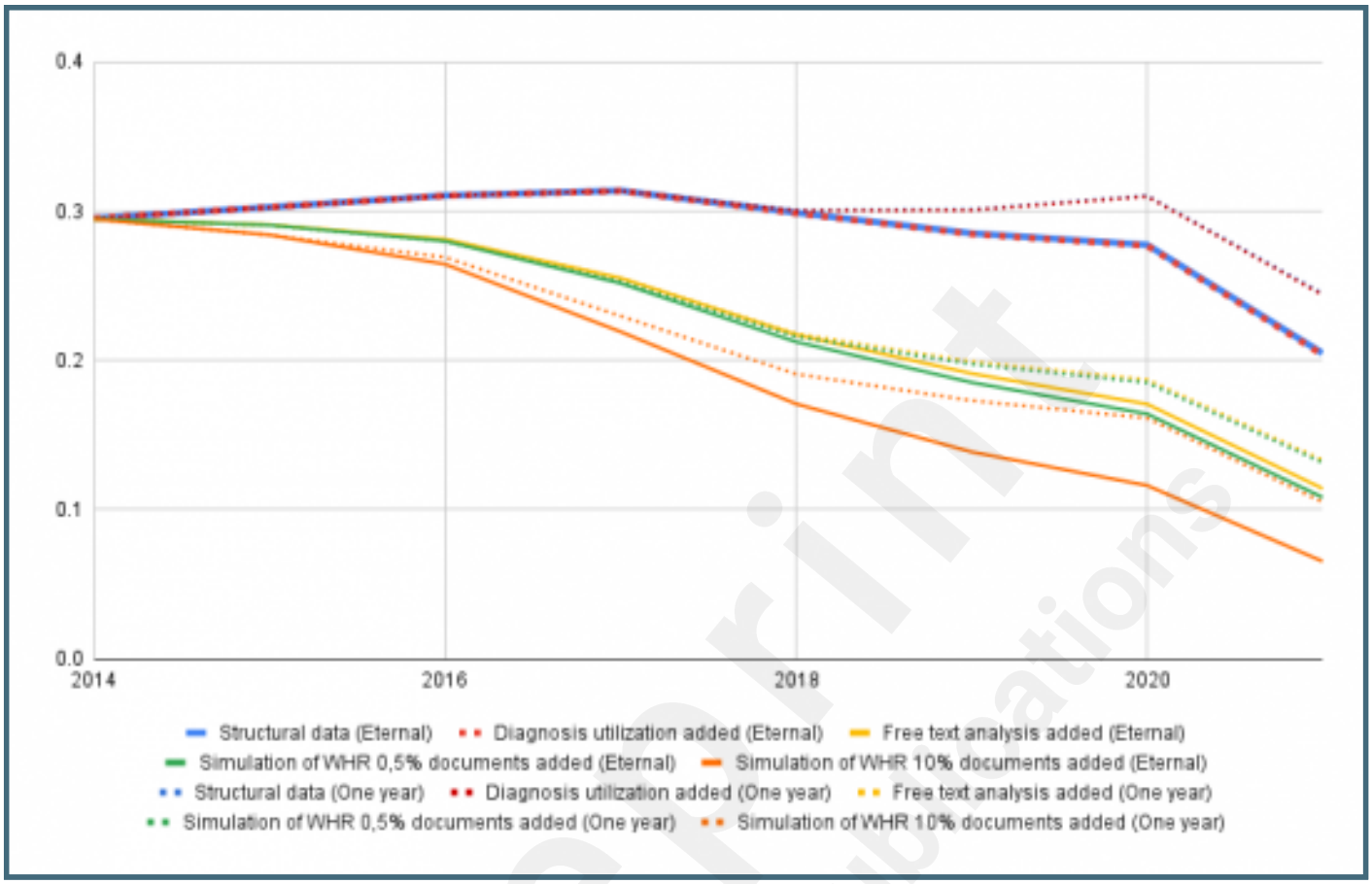
Risk categorization architecture.



Risk categories in minimum and maximum risk plane.



Average risk difference development between 2014 and 2021.



Wait-hip-ratio (WHR) categorization impact to risk difference development between 2014 and 2021.