Received: 5 May 2021 | Revised: 27 November 2021 | Accepted: 25 January 2022

DOI: 10.1049/htl2.12021

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Predicting cumulative effect of lifestyle risk factors for complex disease

Emmanuel Effiok | Enjie Liu | Jon Hitchcock

School of Computer Science and Technology, University of Bedfordshire, Luton, UK

Correspondence
Enjie Liu, School of Computer Science and Technology, University of Bedfordshire, Luton, UK.
Email: enjie.liu@beds.ac.uk

Abstract

In medical domain, risk factors are often used to model disease predictions. In order to make the most use of the predictive models, linking the model with real patient data generates personalized disease progression and predictions. However, the risk factors are fragmented all over medical literature, certain risks can be accumulated for a disease and the aggregated probability may increase or decrease the occurrence of a disease. In this paper, a risk predictive framework which forms a base for a complete risk prediction model that can be used for various health applications is proposed.

1 | INTRODUCTION

Intensive research has been conducted to gather individual risk factors either lifestyle related or otherwise that affect the prevention or progression of a specific disease. However, a combination of several risk factors could affect disease outcome in different ways depending on the disease and risk factors involved. The impact of more than one lifestyle risk factors and the new probabilities presented by aggregating those lifestyle risk factors in the context of disease prevention and occurrence needs to be investigated.

Lifestyle-related diseases are defined as non-communicable diseases caused by non-physiological lifestyle factors. It includes unhealthy diet, physical inactivity, exposure to toxins such as smoking, excessive use of alcohol, environmental determinants (for example carcinogens) and psychosocial factors (for example chronic stress and depression). An example of such a scenario could be, a circumstance in which the fruit carrot has been proven to prevent prostate cancer occurrence and exposure to cadmium has been proven to support prostate cancer occurrence [1], with each risk factor having different probabilities which affects the disease in opposing ways, if an individual consumes carrots and is exposed to cadmium, what is the probability of the individual having or not having prostate cancer based on the lifestyle risk factor combinations?

Risk factors used in this research are those factors collected from various domain researches such as individual work done by researchers on specific risk factors and their relevance to diseases. Finding a way to combine those individual risk factors would yield a new probability of disease occurrence or prevention as risks across different sources can be accumulated for a single disease. This also makes it harder to predict disease occurrence or prevention as a consequence of combining multiple risk factors will not just affect the disease but may also introduce various comorbidities as well, this concept is consistent with the term aggregate risk in the Food Quality Protection Act of 1996 (FQPA) [2].

In this paper, we propose an approach for risk factor selection, classification, sorting and aggregation for a specific disease based on available publications. Our research focuses on life style related risk factors. Results from the predictive framework can be further analysed with the aim of discovering changes in disease outcome for patients with similar risk factors, as some patients could produce different probability outcomes regardless of possessing similar risk factors [3], discovering how and why each person reacts differently could help another patient in a similar circumstance, this therefore provides a slight advantage of further in-sight into understanding complex disease progression for individual patients in different scenarios and also better conditions to prescribe the best suitable care plan and treatment [4].

Risk factors for prostate cancer are used as case study for the prediction framework. A thorough understanding and summary of the risks of prostate cancer from analytical perspective could provide estimations of risk factor aggregation, and further provide and leads for medical research in terms of combinations of risk factors. The research provides both doctors and researchers with additional insight and a possible new way of predicting disease occurrence and prevention.

The rest of the paper is organised as follows: Section 2 gives a brief description of background with definitions of terminologies used in the paper, followed by literature review in Section 3. Section 4 describes risk predictive framework (RPF); in Section 5, we use prostate cancer as a case study of the proposed approach. In Section 6, test and evaluation is provided, and finally, in Section 7, we summarize the work and identify future works.

2 | BACKGROUND

2.1 | Lifestyle related risk factors

Risk factors, by definition, are factors considered to promote diseases or prevent them; same theory applies to prostate cancer lifestyle factors. Defined in [5], disease risk factors are attributes, characteristics or pre-exposures of an individual, which increase or decrease the likelihood of developing a disease or an injury. Most risk factors use different statistical rules of probability and units to express the degree of influence each risk factor has on a specific disease. Risk factors can be classified into modifiable and non-modifiable risks [6]. Modifiable risk factors have been defined as those risk factors that could be prevented, for example, smoking and heavy drinking; and non-modifiable risk factors are those that cannot be prevented, for example, age.

In this paper, prostate cancer is used as an example for a complex disease, lifestyle risk factors for prostate cancer were identified through extensive research of literature and open source data gathered from various data repositories, websites and publications via basic python scripting or direct access download. Having more data increases the chances of having a less biased result as the data used comes from different sources. Most of the data used in the case studies were retrieved from SEER, which is a recognised online data repository for cancer, including cancer research UK. Many lifestyle risk factors have been theorized and few proven to have an effect on prostate cancer, but due to the large number of proposed risk factors, there is not yet an accurate method of discovering which of those risk factors and their combinations have the highest probability of affecting the disease outcome.

2.2 | Preventive, permissive and core risk factors

Lifestyle risk factors may aid in promoting the occurrence or prevention of a complex disease. In this paper, we categorise the risk factors into three categories namely preventive, permissive and core lifestyle factors.

- Preventive risk factors are factors are those factors that have either been theorized or proven to aid in reducing the chances of disease occurrence.
- Permissive risk factors are those factors that have either been theorized or proven to increase the chances of disease occurrence.
- Core risk factors are factors are those that have been mostly proven to provide the highest chances of increasing disease occurrence if an individual possesses them. A large amount of individuals who possess these core factors end up having the disease, but the rest either live a healthy life or exhibit extremely minor disease symptoms which may or may not end up developing into the disease.

Preventive lifestyle risk factors can also be called modifiable risk factors, they have been researched by experts in their fields and has been discovered that to a plausible extent they sometimes either completely prevent or slow the occurrence of a disease, while permissive lifestyle risk factors accelerate the disease occurrence either slowly or quicker, examples of such preventive and permissive lifestyle risk factors in respect to prostate cancer are soy and calcium respectively. Normal lifestyle risk factors also known as modifiable lifestyle risk factors as mentioned in the manuscript are a combination of lifestyle risk factors without the core risk factors. The purpose of the preventive and permissive charts was to evaluate how the same combination of risk factors can affect disease outcome in both positive and negative ways hence permissive and preventive.

3 | LITERATURE REVIEW ON RISK PREDICTION MODELLING

Conventional risk estimation systems take into account only a limited sub-set of the known risk factors, but stressed by the authors of current clinical guidelines, there are needs to consider the likely impact of all risk factors before making clinical management decisions [7, 8].

One of the earliest and most recognisable prediction models is the Framingham study [9], this was an epidemiological study that focused on diseases characterised by the thickening and loss of elasticity in the walls of the arteries especially of the heart, also known as arteriosclerotic diseases. This study is used for the estimation and detailing of incidence rates for individuals who develop a certain condition within a given time period. QRISK3 is a risk prediction algorithm for cardiovascular disease (CVD) and is currently recommended by NICE to be used in UK. Research was made to identify risk factors which influenced the development of the disease [10] in order to test the efficiency of various diagnostic procedures. Disease prediction helps with disease prevention as it was believed that the diseases did not each have a single cause but rather an aggregation of multiple causes which influence disease progression. The study made use of a whole town which provided enough individuals with various socioeconomic and ethnic groups thus introducing a contrast in groups for the study [11], exams were also given in a two year cycle in the form of a specific sample size consisting of volunteers in the town with the aim of producing statistically stable data of individuals who developed the disease and those free of the disease. To avoid data bias, the scientists obtained all samples via random selections. However, the final cohort was not random and it provided statistics that led to the conclusion that it was healthier than the general population [9]. This led the investigators to the realisation that the introduction of volunteers may have introduced biases beyond the ones noticed [8].

Authors in [12] also believed that multiple genetic variants contribute significantly to disease progression, but most risk factor always combine with environmental and lifestyle factors such as smoking, obesity etc. to increase disease occurrence chances. An example of health prediction model is the IBM Watson Health, which is a generalised analytics program that combines knowledge sources such as journal papers and books with data sources such as patient record data and longitudinal records to aid health prediction [13].

The QCancer application is based on the QResearch database [14] over 35 million patients use this data from over 565 practices. QCancer developed the scores system and some of these data sets had been used for validation. It is estimated that in the derivation cohort, there are cases of 4.96 million patients aged 25–84 years old, and in the validation cohort, there are 1.64 million patients and all patients were free of the relevant cancer at baseline. Authors in [15] applied Cox proportional hazards models on GP mortality, or hospital cancer records to derive and validate a set of clinical risk prediction algorithm with the aim of estimating a 10 year risk of common cancers in men and women, to identify patients with high risk of cancers for prevention or further assessment.

Authors in [16] conducted a comparison study of three machine learning techniques for predicting breast cancer recurrence. The data used for study were from patients who registered in the Iranian centre for breast cancer (ICBC) program from 1997 to 2008.

4 | RISK PREDICTIVE FRAMEWORK (RPF)

In this research, we propose a complex disease prediction framework to estimate disease occurrence and prevention probabilities through the unique relationship between lifestyle risk factors and complicated diseases. The risk predictive framework (RPF) adapts the steps of the general predictive modelling methodology and comprises the 4 steps as shown in Figure 1.

4.1 | Risk factor gathering

Risk factors had been extensively researched to find its effect on the disease, a combination of several of those factors would provide a new probability of disease occurrence. But the question is, what will the aggregated probability be defined as, is it an increment in disease occurrence or a decrement? Most of the data were retrieved from data repositories via basic web crawling/scraping with python or direct access download as all data were open source, this data was then stored on a local database to be further analysed later. To reduce data complexity for better understanding and also better statistical computation, as all data retrieved were expressed differently in terms of units; they were all converted to the percentile probability format for future calculation.

All risk factors were classified into different categories such as preventive, permissive and core and were all considered to be completely independent of each other. The lifestyle risk factors were then combined into itemsets to prevent repetition, this is done by using the combination formula of 2^n-1, where $n =$ number of factors to combine. The final dataset after gathering, selecting and sorting were split into an 80:20 ratio for training and testing respectively for the hold-out cross validation method. An issue with this cross-validation method is that it produces a high level of variance and results could differ significantly depending on the division of the training and testing dataset as it is done randomly, but at the same time this completely reduces or eliminates the chances of algorithm bias. All the computation was done using the Anaconda distribution for data analytics which has various tools for data manipulation.

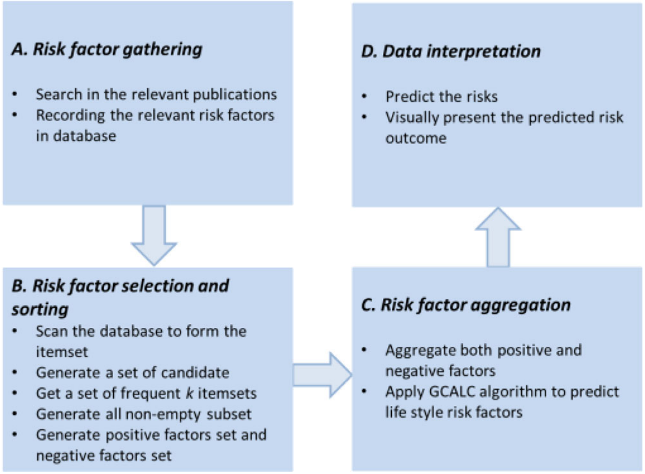

A. Risk factor gathering
• Search in the relevant publications
• Recording the relevant risk factors in database
D. Data interpretation
• Predict the risks
• Visually present the predicted risk outcome
B. Risk factor selection and sorting
• Scan the database to form the itemset
• Generate a set of candidate
• Get a set of frequent k itemsets
• Generate all non-empty subset
• Generate positive factors set and negative factors set
C. Risk factor aggregation
• Aggregate both positive and negative factors
• Apply GCALC algorithm to predict life style risk factors

FIGURE 1 Life style related risk prediction process

4.2 | Risk factor selection and sorting/classification

The required lifestyle risk factors were chosen via the positive and negative selection algorithm (PNS), this was done to select the relevant risk factors by comparing the frequencies of how each risk factor has been theorized or researched more than others while also reducing the size of the dataset, as a lot of risk factors have been proposed to affect various complex diseases. The PNS algorithm [17], is illustrated

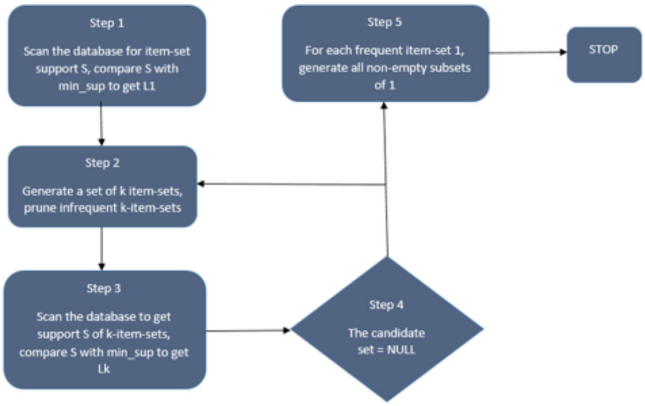
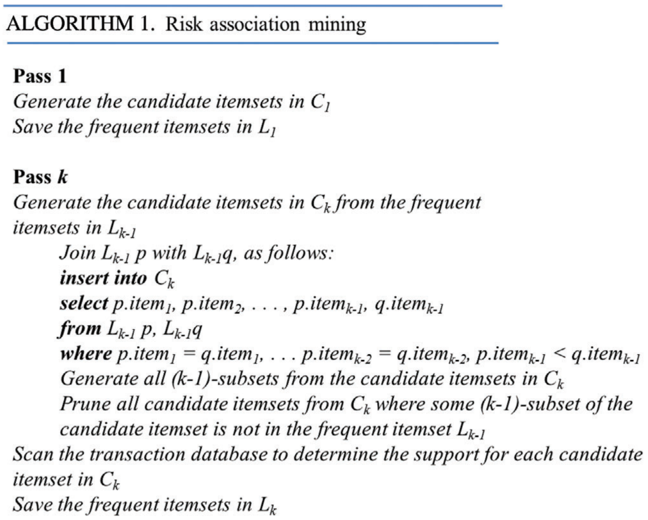

Step 1
Scan the database for item-set support S, compare S with min_sup to get L1
Step 5
For each frequent item-set 1, generate all non-empty subsets of 1
STOP
Step 2
Generate a set of k item-sets, prune infrequent k-item-sets
Step 3
Scan the database to get support S of k-item-sets, compare S with min_sup to get Lk
Step 4
The candidate set = NULL

FIGURE 2 Positive and negative selection algorithm (PNS)

in Figure 2 and is based on the Apriori algorithm, which is an algorithm that utilizes a bottom-up search method with a horizontal layout that enumerates all frequent itemsets. Lifestyle risk factor data probabilities retrieved for this research were all converted into their natural logarithms of their odds ratio, this method was chosen as a means of exception handling for scenarios where lifestyle risk factor data aggregation provided negative results. This was then converted back to probabilities for result interpretation as the odds ratio result is not a numerical measurement of how likely the disease will or will not occur, but rather a ratio of favourable to unfavourable outcomes.

To further select the relevant factors, frequency tables were created for the proposed preventive and permissive factors. The process estimates relevant risk factor selection by using either support or confidence. For this research, the lifestyle risk factor re-occurrence frequency was used as the support for choosing the factor, the re-occurrence frequency is basically the amount of times a certain risk factor has been theorised or researched to have an effect on the disease, therefore hinting at its importance based on the continuous research of the factor.

The algorithm is described in the following pseudo code.



ALGORITHM 1. Risk association mining

```
Pass 1
Generate the candidate itemsets in $C_1$
Save the frequent itemsets in $L_1$

Pass k
Generate the candidate itemsets in $C_k$ from the frequent
itemsets in $L_{k-1}$
    Join $L_{k-1}$ p with $L_{k-1}$q, as follows:
    insert into $C_k$
    select p.item$_1$, p.item$_2$, . . . , p.item$_{k-1}$, q.item$_{k-1}$
    from $L_{k-1}$ p, $L_{k-1}$q
    where p.item$_1$ = q.item$_1$, . . . p.item$_{k-2}$ = q.item$_{k-2}$, p.item$_{k-1}$ < q.item$_{k-1}$
    Generate all (k-1)-subsets from the candidate itemsets in $C_k$
    Prune all candidate itemsets from $C_k$ where some (k-1)-subset of the
    candidate itemset is not in the frequent itemset $L_{k-1}$
Scan the transaction database to determine the support for each candidate
itemset in $C_k$
Save the frequent itemsets in $L_k$
```

4.3 | Risk factor aggregation

In [18], cumulative risk assessment is defined as an analysis, characterization and possible quantification of the combined risks to health from multiple agents or stressors. These exposures or risks can be accumulated over time, with the risk factors causing similar or a variety of effects.

In this paper we proposed a geometric calculator (GCALC) [19] that used to calculate the accumulated risk factors. The renormalization factor of the pooling function intuitively introduces a holistic element enabling final aggregated results to slightly vary most times from expected results. To solve this, we introduce another stage to the geometric pooling method to create geometric calculator. The initial aggregation result of the probabilities is treated as a generalized result that be used to derive both chances of the disease occurring and not occurring:

To aggregate the result, geometric pooling method [20] is used, where:

$$P_{P_1}, \dots, P_{P_n}(\omega) = K[P_1(\omega)]^{\omega_1} \dots \left[P_n(\omega)\right]^{\omega_n} \tag{1}$$

For every ω in Ω, where $\omega_1 \dots\dots \omega_n$ are chosen as non-negative weights with a sum total 1 and k is a normalization factor, given by:

$$k = \frac{1}{\sum_{\omega' \int \Omega} [P_1(\omega')]^{\omega_1} \dots \left[P_n(\omega')\right]^{\omega_n}} \tag{2}$$

$$\pi_\omega(\alpha + \beta) = t(\omega) \prod_{j=1}^{k} \left(\pi_j(\theta)\right)^{\omega_j} \tag{3}$$

where α is positive factor, β is a negative factor and θ is the aggregation of positive and negative factors. Calculating the chances of a disease occurring can be expressed as:

$$\pi_\omega(\alpha) = \left(t(\omega) \prod_{j=1}^{k} \left(\pi_j(\theta)\right)^{\omega_j}\right) - \left(t(\omega) \prod_{j=1}^{k} \left(\pi_j(\beta)\right)^{\omega_j}\right) \tag{4}$$

where $\left(t(\omega) \prod_{j=1}^{k} (\pi_j(\theta))^{\omega_j}\right)$ is the cumulative probability and $\left(t(\omega) \prod_{j=1}^{k} (\pi_j(\beta))^{\omega_j}\right)$ is the aggregation of the negative factors.

Calculating the chances of a disease not occurring can be expressed as:

$$\pi_\omega(\beta) = \left(t(\omega) \prod_{j=1}^{k} \left(\pi_j(\theta)\right)^{\omega_j}\right) - \left(t(\omega) \prod_{j=1}^{k} \left(\pi_j(\alpha)\right)^{\omega_j}\right) \tag{5}$$

where $(t(\omega)\prod_{j=1}^{k}(\pi_j(\theta))^{\omega_j})$ is the cumulative and $(t(\omega)\prod_{j=1}^{k}(\pi_j(\alpha))^{\omega_j})$ is the aggregation of the positive factors. This algorithm was executed using the Python language and the Spyder application in Anaconda.

4.4 | Data interpretation

This stage of the framework visualises the resulting data from GCALC for further analysis or interpretation. The scatter plot was the chosen visualisation method and it can be done using any mathematical software, for example, MatLab, the final data from GCALC are all expressed in probabilities and require no conversion hence making it suitable for any graph plotting software.

5 | PREDICTING PROSTATE CANCER WITH RPF

In this section, we use prostate cancer as an example, to explain the process of risk predictive framework.

The description of RPF in this section will be focused on step B (risk factor selecting and sorting) and C (risk factor aggregation) of Figure 1.

5.1 | Risk factor selecting and sorting with RPF

Figure 3 illustrates the process of risk factor selection. The process follows the steps detailed in Figure 2.

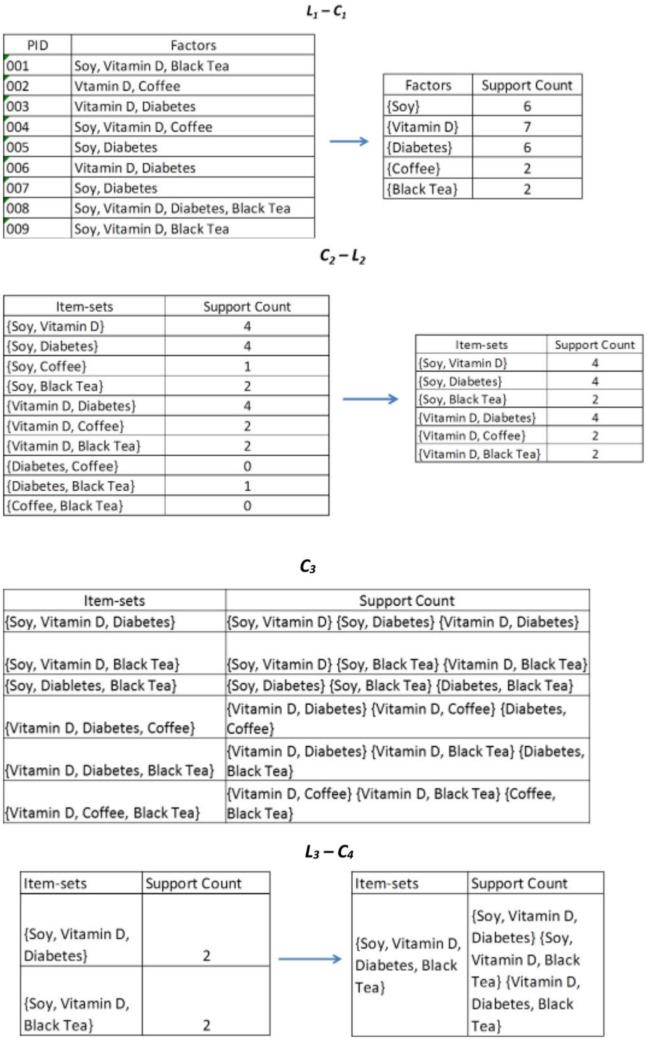
5.2 | Risk factor aggregation with RPF

Prostate cancer was used as a complex disease for prediction. Two case studies were then performed; both case studies had both permissive and preventive lifestyle risk factors with the difference being that one case study includes a core lifestyle risk factor while the other does not.

Positive and negative LRF combinations (Case study 1): This case study used only the permissive and preventive lifestyle risk factors, and was subdivided into permissive and preventive charts, were the permissive chart showed the probability of disease occurrence based on the risk factor combination, while the preventive chart showed the probability of risk factor prevention based on the same risk factor combination.

Positive, negative and core LRF combinations (Case study 2): This case study also used the permissive and preventive lifestyle risk factors but also included a core lifestyle risk factor, for this experiment age was used as it was core risk factor that was present in most of the data gathered. It was also subdivided into the preventive and permissive charts.

Initial aggregation of both positive and negative lifestyle risk factors produced results that represent the combination or aggregation of both lifestyle risk factors and it cannot be interpreted as either a cumulative probability of the disease occurring or not occurring. This was solved by distinguishing both positive and negative probabilities using concepts derived from basic mathematical problems of positive integers: using Equations (3), (4), (5), and where α = positive factors, β = negative factors and θ = geometric pooling result.



$L_1 - C_1$

PID	Factors
001	Soy, Vitamin D, Black Tea
002	Vtamin D, Coffee
003	Vitamin D, Diabetes
004	Soy, Vitamin D, Coffee
005	Soy, Diabetes
006	Vitamin D, Diabetes
007	Soy, Diabetes
008	Soy, Vitamin D, Diabetes, Black Tea
009	Soy, Vitamin D, Black Tea

Factors	Support Count
{Soy}	6
{Vitamin D}	7
{Diabetes}	6
{Coffee}	2
{Black Tea}	2

$C_2 - L_2$

Item-sets	Support Count
{Soy, Vitamin D}	4
{Soy, Diabetes}	4
{Soy, Coffee}	1
{Soy, Black Tea}	2
{Vitamin D, Diabetes}	4
{Vitamin D, Coffee}	2
{Vitamin D, Black Tea}	2
{Diabetes, Coffee}	0
{Diabetes, Black Tea}	1
{Coffee, Black Tea}	0

Item-sets	Support Count
{Soy, Vitamin D}	4
{Soy, Diabetes}	4
{Soy, Black Tea}	2
{Vitamin D, Diabetes}	4
{Vitamin D, Coffee}	2
{Vitamin D, Black Tea}	2

C_3

Item-sets	Support Count
{Soy, Vitamin D, Diabetes}	{Soy, Vitamin D} {Soy, Diabetes} {Vitamin D, Diabetes}
{Soy, Vitamin D, Black Tea}	{Soy, Vitamin D} {Soy, Black Tea} {Vitamin D, Black Tea}
{Soy, Diabletes, Black Tea}	{Soy, Diabetes} {Soy, Black Tea} {Diabetes, Black Tea}
{Vitamin D, Diabetes, Coffee}	{Vitamin D, Diabetes} {Vitamin D, Coffee} {Diabetes, Coffee}
{Vitamin D, Diabetes, Black Tea}	{Vitamin D, Diabetes} {Vitamin D, Black Tea} {Diabetes, Black Tea}
{Vitamin D, Coffee, Black Tea}	{Vitamin D, Coffee} {Vitamin D, Black Tea} {Coffee, Black Tea}

$L_3 - C_4$

Item-sets	Support Count
{Soy, Vitamin D, Diabetes}	2
{Soy, Vitamin D, Black Tea}	2

Item-sets	Support Count
{Soy, Vitamin D, Diabetes, Black Tea}	{Soy, Vitamin D, Diabetes} {Soy, Vitamin D, Black Tea} {Vitamin D, Diabetes, Black Tea}

FIGURE 3 Lifestyle risk factor selection

5.2.1 | Case study 1: Positive and negative LRF combinations

The preventive chart result for case study 1 in Figure 4 show moderate chances of disease prevention based on the combine lifestyle risk factors, with probabilities ranging between 0 and 80% and most data points ranging from 0% to 40%. This case study also showed higher chances of disease occurrence with most lifestyle risk factor combinations.

Also seen in Figure 5 for the permissive chart, those combinations produced probabilities ranging from 0–60% with some as high as a 99% chance of disease occurrence.

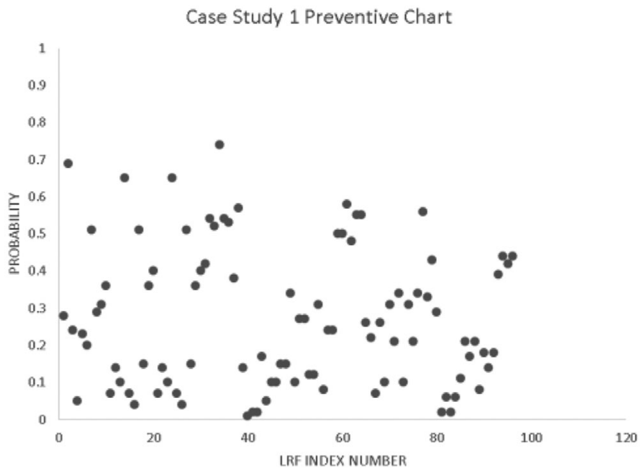


FIGURE 4 Positive and negative LRF combinations—Preventive probability chart

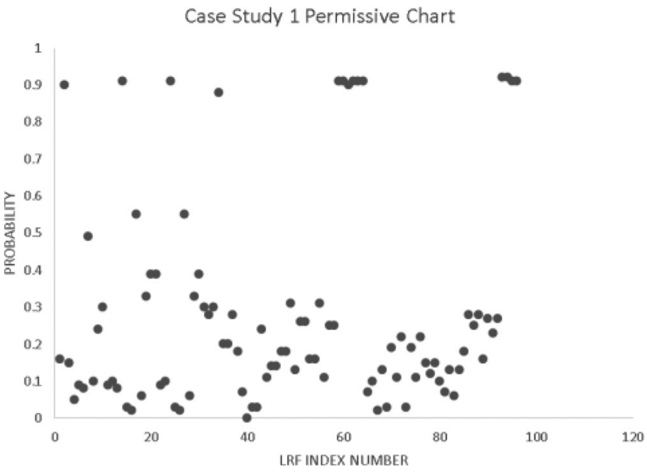


FIGURE 5 Positive and negative LRF combinations—Permissive probability chart

5.2.2 | Case study 2: Positive, negative and core LRF combinations

Case study 2 aims to discover how a core lifestyle risk factor will influence the lifestyle risk factors in case study 1. The core risk factor was included and aggregated as a permissive lifestyle risk factor, as core risk factors are lifestyle risk factors discovered to increase risk of disease occurrence.

Results from the preventive chart for case study 2, illustrated in Figure 6, showed that the inclusion of the core lifestyle risk factors provided a higher chance of disease prevention but only with specific lifestyle risk factor combinations, with probabilities ranging from 50% to 99%. While the permissive chart in Figure 7 shows a direct relationship between the lifestyle risk factors and core lifestyle risk factors as more individuals had a definite probability of having and not having the disease, which means that individuals naturally at risk of disease occurrence can reduce their risk by adapting to specific lifestyles for disease prevention.

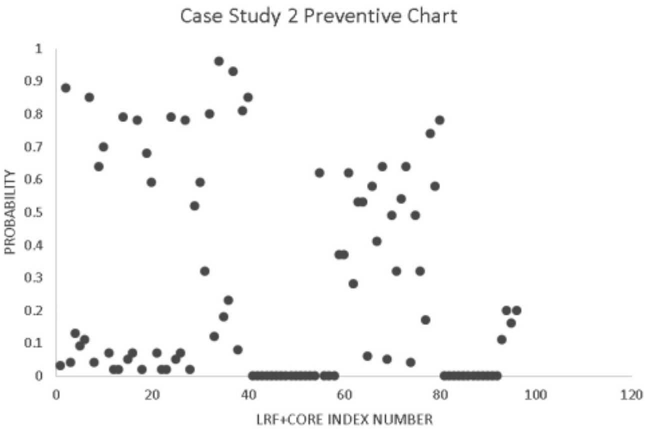


FIGURE 6 Positive, negative and core LRF combinations—Preventive probability chart

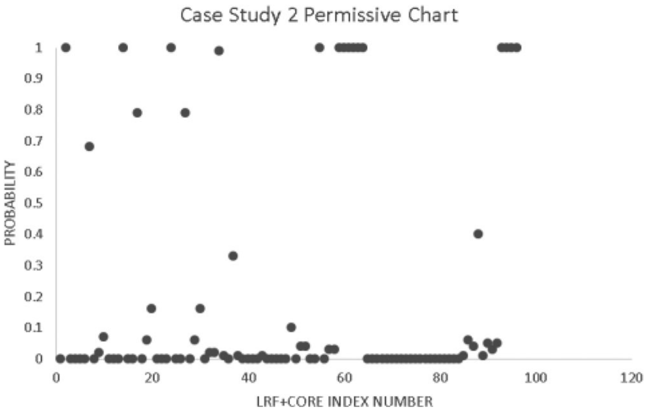


FIGURE 7 Positive, negative and core LRF combinations—Permissive probability chart

6 | TEST AND COMPARISON OF THE MODELLING

An evaluation was made between RPF and several other predictive models summarising the differences between RPF and those predictive models.

Then a validation test using the test data from the data gathering stage was made between the risk factor predictive algorithms GCALC and SWOP [21]. This validation method was chosen due to current lack of similar applications or methodologies that could be used to properly evaluate the success or failure rates of using such a prediction methodology, as such existing applications would already have produced real world results, using mathematical estimations for validation without real world data testing and comparison would only produce derived and not applied results. There is also the absence of evidence and use of the application on real patient data as such data is not open source or available to the public, hence the comparison test

TABLE 1 Comparison between PREDICT and GCALC

PREDICT	GCALC
Estimates prediction based on only the histopathology of the cancer.	Estimates prediction based on an aggregation of opposite lifestyle risk factors.
Designed for better disease prognosis only.	Designed to estimate the probability of the disease occurring and not occurring through cumulative lifestyle risk factors.
Data calculation is based on a large set of tested researched data results from age group trials and cohort studies using specific parameters.	Data calculation is based on theorized and proven data of lifestyle risk factors and disease inter-relationships.
The predictive model is designed for breast cancer.	The predictive model is designed for prostate cancer.

TABLE 2 Comparison between QCancer and GCALC

QCancer	GCALC
Estimates prediction based on more medical risk factors.	Estimates prediction based on positive and negative lifestyle risk factors.
Disease occurring and not occurring risk probability totals 1.	Disease occurring and not occurring probability does not equal 1 as they are calculated independent of each other.
Data calculation is based on researched data results from age group trials and cohort studies using specific parameters, and data routinely collected from various GPs across the UK.	Data calculation is based on theorized and proven data of lifestyle risk factors and disease inter-relationships.
It substitutes missing values for estimated values in the algorithm.	It currently does not substitute missing values for estimated ones.

using retrieved data against an application similar to GCALC that currently exists using the same data, was chosen as the validation method for this experiment.

6.1 | Comparison of the existing prediction algorithms

A few predictive models are similar to GCALC, some of them being:

- PREDICT used by the NHS: This is an online mathematical model designed for both patients and doctors to help them find the best course of treatment following breast cancer surgery. The differences between PREDICT and GCALC is shown in the following Table 1.
- QCancer by QResearch: This is a web based prediction model designed for doctors, nurses and academics. It calculates the risk of prostate cancer occurrence as yet undiagnosed, taking into account various medical risk factors and current symptoms. The differences between GCALC and Qresearch algorithms are shown in Table 2.

6.2 | Test/comparison of risk factor predictive algorithms

For algorithm validation, a cross validation method known as the holdout method was implemented. Each test data contained normalized multivariate patient data with the required risk factors for both GCALC and SWOP disease prediction. SWOP is an application that also calculates the risk of prostate cancer using various questionnaire type format of gathering data input for calculation. The same data inputted in SWOP was also added to the GCALC dataset from the gathering stage, this was done for consistency. SWOP is an application of The European Randomized Study of Screening for Prostate Cancer.

This evaluation approach was undertaken only as a means of comparison to aid analysis of how both predictive models evaluate risk factor data for prediction using different risk factor data requirements but also the same data source, as each prediction model/risk calculator use very different risk factors and calculation methods, only results produced by both calculators based on data from the same individual can be analysed and evaluated, this is also particularly useful as comparing various predictive models that perform similar tasks can lead to new knowledge discovery for critical analysis and possibly an improvement of one or few of the models compared. A similar method was executed by [22], where the authors presented a non-parametric approach for the analysis of areas correlated ROC curves, it was discovered that when two or more empirical curves were designed based on the same individual test data, a statistical analysis on the differences between the curves must take into account the correlated nature of the data.

The evaluation sequence consists of 5 patients all within the age range of 70–74 years and no family history of prostate cancer, but only 3 patient results will be used for this paper. Each patient has been afflicted with one or more urinary issues and have also been exposed to lifestyle risk factors known to affect prostate cancer, the evaluation data can be retrieved from the Appendix section. SWOP utilises urinary issues data in the form of a questionnaire for prediction, data retrieved

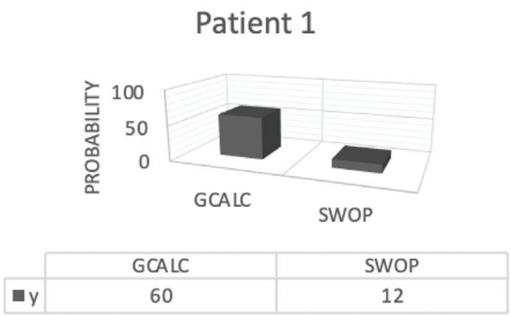


FIGURE 8 Evaluation example 1

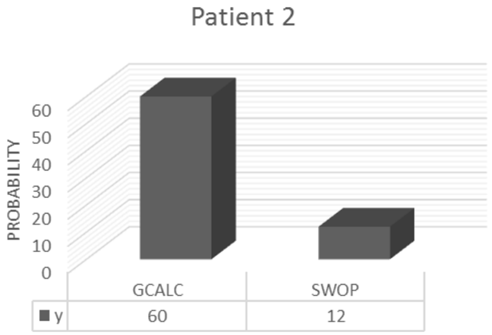


FIGURE 9 Evaluation example 2

from the baseline questionnaires for each patient was used for SWOP prediction, and as GCALC uses lifestyle risk factors, lifestyle risk factors specific to each patient and age were used for prediction, as each patient is at an age where the probability of disease occurrence is high. Only the permissive results for GCALC were used for comparison with the SWOP results as SWOP only predicts the chances of disease occurrence and not both occurrence and prevention as GCALC, and only the risk calculator 1 for the online SWOP prediction tool was used for evaluation as risk calculator 2 requires the prostate specific antigen value (PSA), which GCALC does not currently use.

This evaluation presents 2 algorithms designed to solve the same problem but using different methods, information from the same individual data source is used by both algorithms for prediction and the results are further analysed to predict possible reasons for result similarities or differences.

Evaluation results are as follows with more details after the diagrams:

Patient 1 is a 70 year old obese male with frequent use of Aspirin and some urinary issues, predicting the chances of disease occurrence based on only patient age and lifestyle risk factors with no account for race or family history. As shown in Figure 8, GCALC presented a 60% chance of disease occurrence, while SWOP predicted a 12% chance of disease occurrence based on the patient urinary issues.

Required data for the SWOP online risk calculator were inputted with results recorded for comparison with GCALC results from the same patient data for all patients.

Patient 2 is a 72 year old diabetic male with exposure to Arsenic. As shown in Figure 9, GCALC predicted a 60% chance of disease occurrence based on patient age and lifestyle risk factors and SWOP predicted 12% chance of disease occurrence based on the specific patient urinary issues.

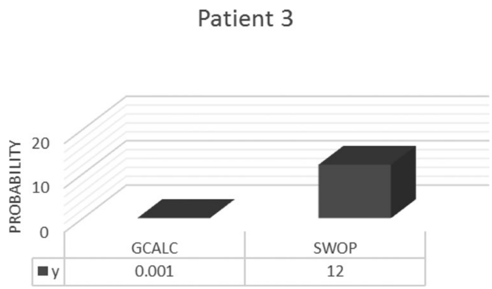


FIGURE 10 Evaluation example 3

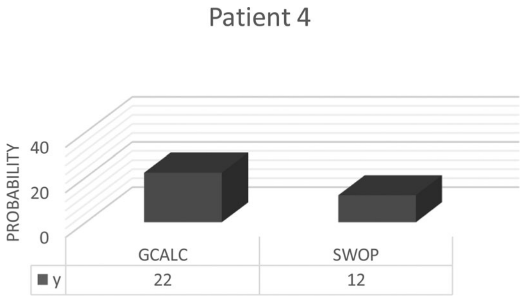


FIGURE 11 Evaluation example 4

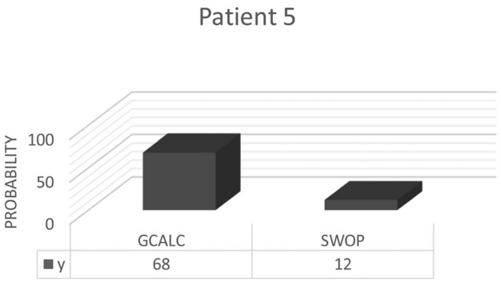


FIGURE 12 Evaluation example 5

Patient 3 is a 70 year old male with constant exposure to Vitamin D, Arsenic and Calcium. Based on age and those combined lifestyle risk factors, as shown in Figure 10, GCALC predicted a 0.001% chance of disease occurrence while SWOP predicted a 12% chance of disease occurrence based on the patient specific urinary issues.

Patient 4 is a 74 year old diabetic male with constant use of Aspirin, as shown in Figure 11, GCALC predicted a 22% chance of disease occurrence based on those lifestyle risk factors while SWOP predicted a 12% chance of disease occurrence.

Patient 5 is an obese 71 year old male with constant use of Aspirin and exposure to Arsenic, as shown in Figure 12, GCALC predicted a 68% chance of disease occurrence based on those lifestyle risk factors while SWOP predicted a 12% chance of disease occurrence.

Based on the comparisons, GCALC tends to show higher degrees of probability for disease occurrence when compared to the SWOP risk calculator except in Patients 3 where GCALC and SWOP had similar results as both algorithms presented low chances of disease occurrence.

7 | CONCLUSION AND FUTURE WORK

In this paper, we propose a risk predictive framework for aggregating lifestyle risk factors and demonstrated it via case studies. The aggregated prediction model generates possible leads of risk factors research for domain experts. The objective of the paper was to suggest a base model for predicting both positive and negative outcomes for complex diseases. This base model was not intended to replace any current testing method but rather to be used as an ad-hoc test to support currently existing testing methods. Once fully validated by clinical trials, it can be used as a knowledge base for many IoT based heath applications, alongside with monitored data for individual patient, a personalised risk prediction can be developed to provide advice during patients' self-management, coaching services, and patient education.

FUNDING

None.

CONFLICT OF INTEREST

The authors do not have a conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Enjie Liu https://orcid.org/0000-0003-4617-595X

References

1. Norrish, A., Jackson, R., Sharpe, S., Skeaff, C.: Prostate cancer and dietary carotenoids. Am. J. Epidemiol. 151(2), 119–123 (2000). https://doi.org/10.1093/oxfordjournals.aje.a010176
2. Food Quality Protection Act of 1996 (1996). https://www.congress.gov/104/plaws/publ170/PLAW-104publ170.pdf. Accessed Nov 2019
3. Delgadillo, J., Moreea, O., Lutz, W.: Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. Behav. Res. Ther. 79, 15–22 (2016). https://doi.org/10.1016/j.brat.2016.02.003
4. O'Flynn, N., Staniszewska, S.: Improving the experience of care for people using NHS services: summary of NICE guidance. BMJ 344(mar16 1), d6422 (2012). https://doi.org/10.1136/bmj.d6422
5. Mahmood, S., Levy, D., Vasan, R., Wang, T.: The Framingham heart study and the epidemiology of cardiovascular disease: A historical perspective. Lancet 383(9921), 999–1008 (2014). https://doi.org/10.1016/s0140-6736(13)61752-3
6. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. 1st Ed. Pearson Addison Wesley, Boston (2005)
7. Cooney, M.T., Dudina, A.L., Ian, M., Graham, I.M.: Value and limitations of existing scores for the assessment of cardiovascular risk, a review for clinicians. J. Am. Coll. Cardiol. 54(14) (2009). https://doi.org/10.1016/j.jacc.2009.07.020
8. Aktas, M.K., Ozduran, V., Pothier, C.E., Lang, R., Lauer, M. S.: Global risk scores and exercise testing for predicting all-cause mortality in a preventive medicine program. JAMA 292, 1462–1468 (2004)
9. Framingham Heart Study (2019). https://www.framinghamheartstudy.org/fhs-about/history/epidemiological-background/. Accessed Nov 2019
10. Dawber, T.: The Framingham study. Ann. Intern. Med. 94(2), 286 (1981). https://doi.org/10.7326/0003-4819-94-2-286_1
11. Feinlieb, M.: The Framingham study: Sample selection, follow-up and methods of analyses. Natl. Cancer Inst. Monogr. (67), 20–35 (1983)
12. The Personalized Medicine Technology Landscape. pp. 1–188. PHG Foundation, Cambridge, UK (2018)
13. IBM Micromedex Care Delivery Evidence-Based Clinical Decision Support for Healthcare Decision-Makers. pp. 1–7. IBM Watson Health, New York 2018
14. QResearch Survey: https://www.qresearch.org/
15. Hippisley-Cox, J., Coupland, C.: Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ Open 5(3), e007825–e007825 (2015). https://doi.org/10.1136/bmjopen-2015-007825
16. Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R.: Using three machine learning techniques for predicting breast cancer recurrence. J. Health Med. Inf. 04(02), 124 (2013). https://doi.org/10.4172/2157-7420.1000124
17. Effiok, E., Liu, E., Hitchcock, J.: Life style related risk association mining. In: International Conference on Internet of Things, Embedded Systems and Communications (IINTEC). Hamammet, Tunisia (2018)
18. Sexton, K.: Cumulative health risk assessment: Finding new ideas and escaping from the old ones. Human Ecol. Risk Assess. 21(4), 934–951 (2014). https://doi.org/10.1080/10807039.2014.946346
19. Effiok, E., Liu, E., Hitchcock, J.: Lifestyle risk association aggregation. In: Proceeding of the Fifth IEEE International Workshop on Internet of Things: Networking Applications and Technologies. Limerick (2019)
20. Dietrich, F., List, C.: Probabilistic opinion pooling generalised. Part two: the premise-based approach. Social Choice Welfare 48(4), 787–814 (2017)
21. SWOP. The Prostate Cancer Research Foundation. (2019). http://www.prostatecancer-riskcalculator.com/. Accessed Nov 2019
22. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, JSTOR 65–70 (1988)
23. WHO, GLOBAL HEALTH RISKS: (2009). https://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf. Accessed Nov 2019

How to cite this article: Effiok, E., Liu, E., Hitchcock, J.: Predicting cumulative effect of lifestyle risk factors for complex disease. Healthc. Technol. Lett. 9, 34–42 (2022). https://doi.org/10.1049/htl2.12021