Imperial College London

School of Public Health

Exploiting comorbidities from longitudinal health records to inform risk prediction of skin cancer

Authors: Carolina Richheimer, Eléonore Schneegans, Demetris Hajivassiliou, Sarvesh Dhungana, Luca Ramelli

Supervisors: Dr D. Vuckovic, Mr T. Wright, Ms A. B. Metzler

Date: March 26, 2021

Abstract:

Background: In this study, we use a wide range of EHR data, including life-course hospital episode data (i.e. comorbidities), known risk factors (i.e. lifestyle, sun exposure, family history of cancer), as well as biomarker data, in time-stamped risk prediction models for skin cancer.

Objective: We investigate whether comorbidities prior to skin cancer diagnosis can improve risk prediction compared to knwon risk factors alone, and how far back past diagnoses are informative. We also conduct a targeted analysis to see which past comorbidities are the most relevant in predicting skin cancer risk.

Methods: This study follows a retrospective nested case control design, using participants from the UK biobank. We use different combination of predictors, including known risk factors of skin cancer, comorbidities and biomarkers, in different timeframes, and compare predictive accuracy obtained from different machine learning models. Four models were selected, two of which could model linear effects (Multivariate logistic regression, Least Absolute Shrinkage and Selection Operator), and the others non-linear effects (Random Forest, Support Vector Machines).

Conclusion: Using a time-stamped approach, we showed that comorbidity information prior to skin cancer diagnosis is indeed improving risk prediction compared to using known risk factors alone, and this, up to 5 years before disease diagnosis. Models using cumulated life-course comorbidities showed an overall better predictive accuracy compared to time-windows, suggesting that any diseases that has occurred in the past may have an effect for future risk of skin cancer. Generally, history of neoplasms, skin related disorders like actinic keratosis, as well as sun exposures-related characteristics (i.e. childhood sunburn and skin color) were the strongest predictors. Overall, the targeted analysis flagged predictors that were already known to be strongly associated to the risk of developing skin cancer in the literature, but other unconfirmed risk factors were also selected by our models.

1 Introduction

Skin cancer is described as an uncontrollable proliferation of abnormal skin cells. These cancers typically develop in the epidermis, the outermost layer of the skin. Generally, skin cancers can be divided into two types: non-melanoma skin cancer (NMSC) and melanoma skin cancer. Melanoma is characterized to behave invasively, and thus is associated with a high mortality rate, while NMSCs are described to typically have aggressive features that are maintained locally [1]. It has been estimated that NMSC is the most common group of cancers, accounting for approximately 20% of all new malignancies and 90% of all skin cancers in the UK [2]. Cancer Research UK reports that there are nearly 152,000 new non-melanoma skin cancer cases yearly, and incidence rates have increased by 166% since the 1990s. On the other hand, melanoma is reported as the fifth most common cancer, with a lower incidence of 16,200 cases every year, attributing to an increase in incidence rate of 135% in the last 30 years.

The major cause of skin cancer is ultraviolet rays, from sun exposure or from tanning machines. The nature and timing of exposure to these UV rays seem to affect risk and type of skin cancer [3]. About 90% of skin cancers are associated with solar UV, and more than 419,000 cases of skin cancer in the US each year are linked to indoor tanning [4–6]. The epidemiologic evidence implicating sun exposure in the causation of melanoma is supported by biologic evidence that damage caused by UV radiation, particularly damage to DNA, plays a central part in the pathogenesis of these tumors [7]. The color of eyes, hair and skin, as well as skin type, are now well-established risk factors in melanoma development. Other known risk factors include family history of melanoma and Xeroderma pigmentosum [8].

1.1 Machine learning methods for skin cancer research

To this date, machine learning scientists have mainly explored the potential of image classification with direct applications such as diagnosis of malignant and benign lesions. These sophisticated classification algorithms achieve an accuracy comparable to board-certified dermatologists [9] [10]. Given their longitudinal and multidimensional nature, electronic health records (EHR) are now slowly gaining attention. Recent work showed that pancreatic cancer can be predicted using comorbidities occurring up to 5 years before cancer diagnosis [11]. This finding is likely generalizable to other types of cancer, which is the focus of this research. Preliminary research that leverages non-imaging health information to predict the risk skin cancer has shown promising results, but is still in its early stages. Overall:

1. These studies focus on very restricted range of potential predictors [12]
2. The timing of diagnosis of comorbidities is completely disregarded [11]
3. They leverage information from health survey and are thus subject to biases of self-reporting [13]
4. Those that leverage medical records haven't done so in parallel to lifestyle factors and laboratory data [14]

1.2 Research aims

In this study, we address this gap in the scientific literature and use a wide range of EHR data, including life-course hospital episode data (i.e. comorbidities), known risk factors (i.e. lifestyle, sun exposure, family history of cancer), as well as biomarker data, in time-stamped risk prediction models for skin cancer. We hypothesize that our combined models can be used to identify the risk than an individual will develop skin cancer at a higher precision that current models that only consider known risk factors alone. Using data from the UK Biobank (UKB), we explore:

1. Cancer risk factors: Do comorbidities prior to skin cancer diagnosis improve risk prediction?
2. Longitudinal analysis: How far back is this informative?
3. Targeted analysis: which comorbidities are the most informative in predicting cancer risk?

Leveraging health records represent a powerful source of knowledge for identifying high-risk profiles at no cost. In the UK, there is no national screening programme for skin cancer and the main reason has been claimed to be that 'the benefits don't outweigh the costs' [15]. If indeed there is some predictive value from longitudinal records, this work could pave to way to the implementation of targeted screening programmes and primary prevention strategies. More generally, this could help to reduce the economic burden of treating these skin cancers and improve the prognosis of the patient cases.

2 Methods

2.1 Study design and participants

The UKB is a national cohort that includes over 500,000 participants from 40 to 69 years old, enrolled between 2006 and 2010. It contains a wide range of pseudo-anonymized information including phenotypic, genetic, and clinical data. This retrospective nested case-control study looks at relevant skin cancer cases and matched controls extracted from UKB along (1) the Hospital Episodes Statistics (HES) base, (2) covariates of interest, (3) baseline biomarkers.

Skin cancer diagnosis was defined as any of the 20 codes from melanomas of skin (C43.0-C43.9) or unspecified malignant neoplasms of skin (C44.0-C44.9) in the International Classification of Diseases 10 (Tables 6:7, Appendix). 25,130 consenting participants of the UKB were identified as cases, predominantly in either of these three skin cancer categories: "unspecified malignant neoplasm of skin of unspecified parts of face", "unspecified malignant neoplasm of skin of trunk", and "unspecified malignant neoplasm of skin of scalp and neck". Cases were matched by age and gender to controls without any previously recorded skin cancer diagnosis, resulting in a cohort of 50,260 participants.

2.2 Datasets

2.2.1 Comorbidities

The base dataset includes participants' HES data, which is any existing disease diagnostic codes from the ICD10, referred to as 'comorbidities' in our study. The entire set of codes represented in this base was of 8,158 out of the 69,000 existing ICD10 codes. For analysis purposes, each comorbidity was transformed into a binary predictor variable. In the instance where the participant had had this diagnosis at any time point, the observation was coded as 1, and 0 otherwise. 23 cases and their respective controls were excluded given missing diagnosis dates for skin cancer or missing ICD10 code. Participant status was coded as 1 for cases and 0 for controls. To account for the timing of diagnosis, each comorbidity diagnosis date was compared to the reference date of skin cancer diagnosis in cases and their matched control. Two one-hot-encoded, time-stamped datasets were derived from the comorbidities that had occurred in the participants' life-course (1) more than 1 year prior to diagnosis ('1+ years'), (2) more than 5 years prior to diagnosis ('5+ years'), both of which included 8,138 ICD10 predictors of 50,214 observations. These time frames were chosen since the goal of the study was to identify a set of comorbidities that predict the occurrence of skin cancer 1 and 5 years in the future. Keeping in mind that cancer is a slow developing condition, this analysis can be thought of as a life-course approach to disease, where we consider that any diseases that has occurred in the past may have an effect for future risk of skin cancer.

2.2.2 Known risk factors

A literature review was conducted to find known risk factors for skin cancer. 22 of these overlapped with the available data fields in UKB and passed the quality check (i.e. less than 30% missing data). Data imputation was performed on the one-hot-encoded dataset using knn imputation package from R. Regarding data cleaning, the 'prefer not to say' group was removed from all categorical variables since it showed high degree of correlation, indicating that missingness was likely completely random. The variable family history of cancer was derived from 'illnesses of mother' and 'illnesses of father'. The resulting dataset includes variables related to sun exposure (i.e. time spent outdoor, childhood sunburn, ease of skin tanning), skin colour, overall health and diet (i.e. BMI, meat/vegetables intake), and family history of cancer. No adjustments for time periods were made, since the variables mostly described long-lasting behavioural features or stable characteristics of participants.

2.2.3 Biomarkers

Including biomarkers in the analysis was purely opportunistic and agnostic given that none of these included known skin cancer biomarkers such as Tyrosinase (i.e. synthesis of melanin), or MITF (i.e. melanocyte differentiation). Two biomarkers with over 50% of missing entries for the cohort (Oestradiol and Rheumatoid factor) and 3,380 individuals missing over 50% of measurements for biomarkers, assumed to be missing completely at random, were removed. Data imputation was performed using knn imputation package from R. Measurements for biosamples were performed at enrolment, thus, a separate subsample of cases for which the diagnosis of skin cancer occurred after their enrolment in UKB (i.e. incident cases), was used for analysis. This subsample consisted of 19,207 out of the 25,130 cases and their matched controls, yielding a subsample of 38,414 observations and 28 predictors.

2.3 Baseline analysis

2.3.1 Univariate logistic regresssion

We explore the associations between each of the 8,138 comorbidities and skin cancer status using univariate logistic regression. To correct for multiple testing, we use the Bonferroni correction, where the significance threshold is derived from:

$$\alpha' = \frac{\alpha}{p}$$

Where p is the number of independent tests performed and alpha is the significance level for a single logistic regression set as 0.05. Results are visualized in Manhattan plots where the $-log10(p\text{-values})$ are ordered by their ICD10 chapter category. This baseline model allows for a simple assessment of differences between cases and controls.

2.3.2 Networks

Networks can give insight into the functional relationships between two entities. For this research purpose, networks were used to visualize pairwise relationships between two possible comorbidities and gain further insights on which tend to co-occur together, in cases versus controls, as well as the links between informative comorbidities selected by the predictive models. The correlation between binary variables was estimated similar as to Roque FS, et al. [16]. p-values were calculated via Fisher's exact test for each comorbidity. From these candidate pairs only the pairs with a p-value equal to or below 0.05 were selected.

2.4 Predictive Models

The three different datasets comorbidities, covariates and biomarkers were used in different combinations for the two time periods. Four machine learning models were selected, two of which could model linear effects (Multivariate logistic regression, Least Absolute Shrinkage and Selection Operator (Lasso)), and the others non-linear effects (Random Forest (RF), Support Vector Machines (SVM)). Models were trained on a subsample of 80% of the data (using balanced proportion of cases and controls) and tested on the remaining 20%. The AUC was chosen as the performance metric to compare all models, since all classes are equally important when predicting for future health status.

First, to investigate whether life-course comorbidities can improve risk prediction for skin cancer, three sets of models will be compared, using known risk factors alone as a baseline predictive model (0). Models used for comparison include: (1) the set of 8,138 comorbidities, (0) + (1). The secondary analysis focuses on a predictive model for incident cases that includes baseline covariates, comorbidities and biomarkers. For the second research aim looking at the longitudinal predictive power of comorbidities, the predictive accuracy for the two different time frames is compared (i.e. 1+ versus 5+ years). Finally, we assess the most informative comorbidities in predicting cancer risk by analysing the set of selected comorbidities in variable selection models (Lasso and RF).

2.5 Sensitivity analysis

In the sensitivity analyses we use two different time frames, looking at time windows instead of the life-course approach: (1) the year preceding diagnosis ('-1 year'), (2) Five years to one year preceding diagnosis ('-5:-1 years'). This enables to assess the extent to which noise may be responsible for the results seen in the main analysis. The sensitivity analysis is carried out using the methods and models described in section 2.4.

3 Results

3.1 Cohort description

Demographics and selected characteristics for skin cancer cases, controls and the total study population are presented in table 1. All variables excluding matching criteria (sex/age) were statistically different between cases and controls (p-value < 0.001). Participants were on average 73.82 years old with 52.6% of the cohort being male. The mean BMI was over 27, which lies within the interval for overweight. Most participants rated their health as ‘good‘. The most observed skin colour was the second fairest option referred to as ‘fair‘ with over two thirds of participants, followed by ‘light olive‘ with more than 10%. Besides skin colour, another known risk factor for skin cancer is ease of tanning. Just below two fifths of participants reported that they ‘get moderately tanned‘. The other options ‘get mildly or occasionally tanned‘, ‘Never tan only burn” and “get very tanned‘ corresponded to 16 and 22% of the participants respectively. On average, control participants had had 15 recorded hospital episodes, and cases had 19 episodes on top of the cancer diagnosis. 2859 controls had no hospital episode recorded.

A χ^2-test was performed to assess differences in the prevalence of comorbidities between cases and controls. Only personal history of malignant neoplasms of other organs and systems (Z858), disorder of the skin and subcutaneous tissue unspecified (L989), actinic keratosis (L570) and Seborrheic keratosis (L82) out of the 8138 comorbidities were significantly different between the two groups. The latter two comorbidities being benign growths of skin cells, commonly caused by sun exposure and age respectively.

Table 1: Demographics and baseline characteristics of participants of the retrospective nested case-control study matched by sex and age

	Controls	Cases	Total	p-value
n	25130	25130	50260	
Age at baseline (mean (SD))	60.84 (6.72)	60.88 (6.71)	60.86 (6.72)	0.996
Bmi (mean (SD))	27.59 (4.64)	27.21 (4.47)	41.56 (6.44)	<0.001
Sex = sex_1 (%)	13220 (52.6)	13239 (52.6)	26459 (52.6)	0.949
Health rating (%)				0.001
Do not know	103 (0.4)	83 (0.3)	186 (0.4)	
Excellent	3891 (15.5)	4048 (16.1)	7939 (15.8)	
Fair	5486 (21.8)	5239 (20.9)	10725 (21.3)	
Good	14452 (57.5)	14743 (58.7)	29195 (58.1)	
Poor	1198 (4.8)	996 (4.0)	2194 (4.4)	
Height (mean (SD))	168.70 (9.21)	169.45 (9.18)		<0.001
Childhood sunburn (mean (SD))	1.48 (3.33)	2.12 (9.21)	1.8 (9.8)	<0.001
smoking status (%)				<0.001
Current smoker	2456 (9.8)	1975 (7.9)	4431 (8.8)	
Never smoker	12874 (51.2)	13353 (53.2)	26227 (52.2)	
Previous smoker	9800 (39.0)	9779 (38.9)	19579 (39.0)	
Alcohol intake frequency (%)				<0.001
Daily or almost daily	5546 (22.1)	5875 (23.4)	11421 (22.7)	
Never	2141 (8.5)	1651 (6.6)	3792 (7.5)	
Once or twice a week	6172 (24.6)	6403 (25.5)	12575 (25.0)	
One to three times a month	2596 (10.3)	2439 (9.7)	5035 (10.0)	
Special occasions only	2974 (11.8)	2558 (10.2)	5532 (11.0)	
Three or four times a week	5701 (22.7)	6181 (24.6)	11882 (23.6)	
skin-color (%)				<0.001
Black	158 (0.6)	9 (0.0)	167 (0.3)	
Brown	622 (2.5)	116 (0.5)	638 (1.3)	
Dark olive	422 (1.7)	248 (1.0)	670 (1.3)	
Do not know	339 (1.3)	252 (1.0)	591 (1.2)	
Fair	17305 (68.9)	18607 (74.1)	35912 (71,5)	
Light olive	4440 (17.7)	3210 (12.8)	7765 (15.4)	
Very fair	1844 (7.3)	2665 (10.6)	(9.0)	
Ease of skin tanning (%)				<0.001
Do not know	608 (2.4)	443 (1.8)	1051 (2.1)	
Get mildly or occasionally tanned	5044 (20.1)	5492 (21.9)	10536 (21.0)	
Get moderately tanned	9977 (39.7)	9647 (38.4)	19624 (39.0)	
Get very tanned	5436 (21.6)	4383 (17.5)	9819 (19.5)	
Never tan only burn	4065 (16.2)	5142 (20.5)	9207 (18.3)	
Time spent outdoors summer (mean (SD))	4.03 (2.36)	4.10 (2.32)	10.95 (3.31)	1
Vegetable intake (mean (SD))	2.12 (2.13)	2.10 (1.97)	2.11 (2.99	280
Water intake (mean (SD))	2.51 (2.12)	2.47 (2.12)	2.49 (3.00)	20
Family history cancer (%)	8027 (31.90)	8503 (33.90)	16530 (32.9)	<0.001
Number of hospital episodes	15.29	18.76	34.05	
At least one recorded hospital episode (%)	88	100	94	

3.2 Networks of comorbidities in cases and controls

To investigate how comorbidities were linked with each other within the two groups, we looked at the correlation between the 100 most prevalent comorbidities. Network plots were built separately for cases and controls using the methods described in section 2.3 (figures 1). The more connections a comorbidity has with others, the greater the size of its node represented in the network. Two important central nodes observed in both networks were essential primary hypertension (I50) and other forms of chronic ischemic heart disease (I258), both comorbidities with high prevalences in the UK population [17]. Other important nodes seen in cases were Type 2 diabetes mellitus (E119) and conduction disorder (I459). Controls showed a different network topology, with different central nodes: noninfective gastroenteritis and colitis, unspecified (K529), diaphragmatic hernia (K449) and pain localized to upper abdomen (R101). These highlight some obvious medical history disparities in cases versus controls.

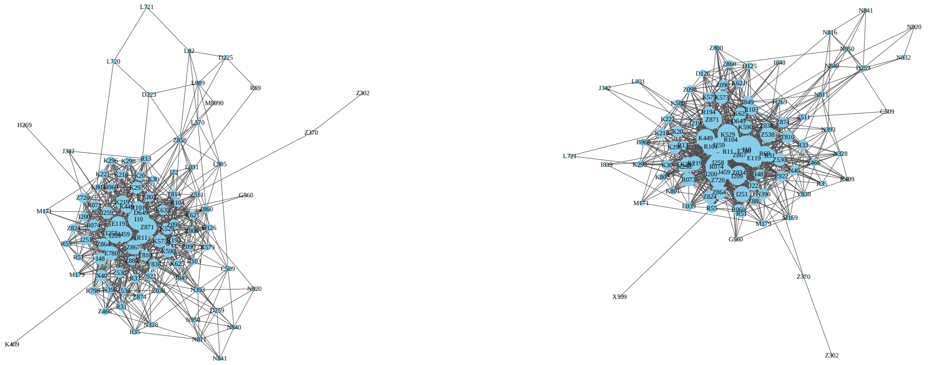

Figure 1: Network for the cases (left) and control group (right) of the 100 most prevalent comorbidities

3.3 Preliminary univariate analysis

Results for both the life-course approach and time windows are presented as Manhattan plots in figure 2. All comorbidities above the pink dashed line were above the set Bonferonni-corrected threshold and considered statistically significant findings. This allowed to filter out comorbidities which could potentially cause noise and identify a set of candidate comorbidities for the prediction of skin cancer.

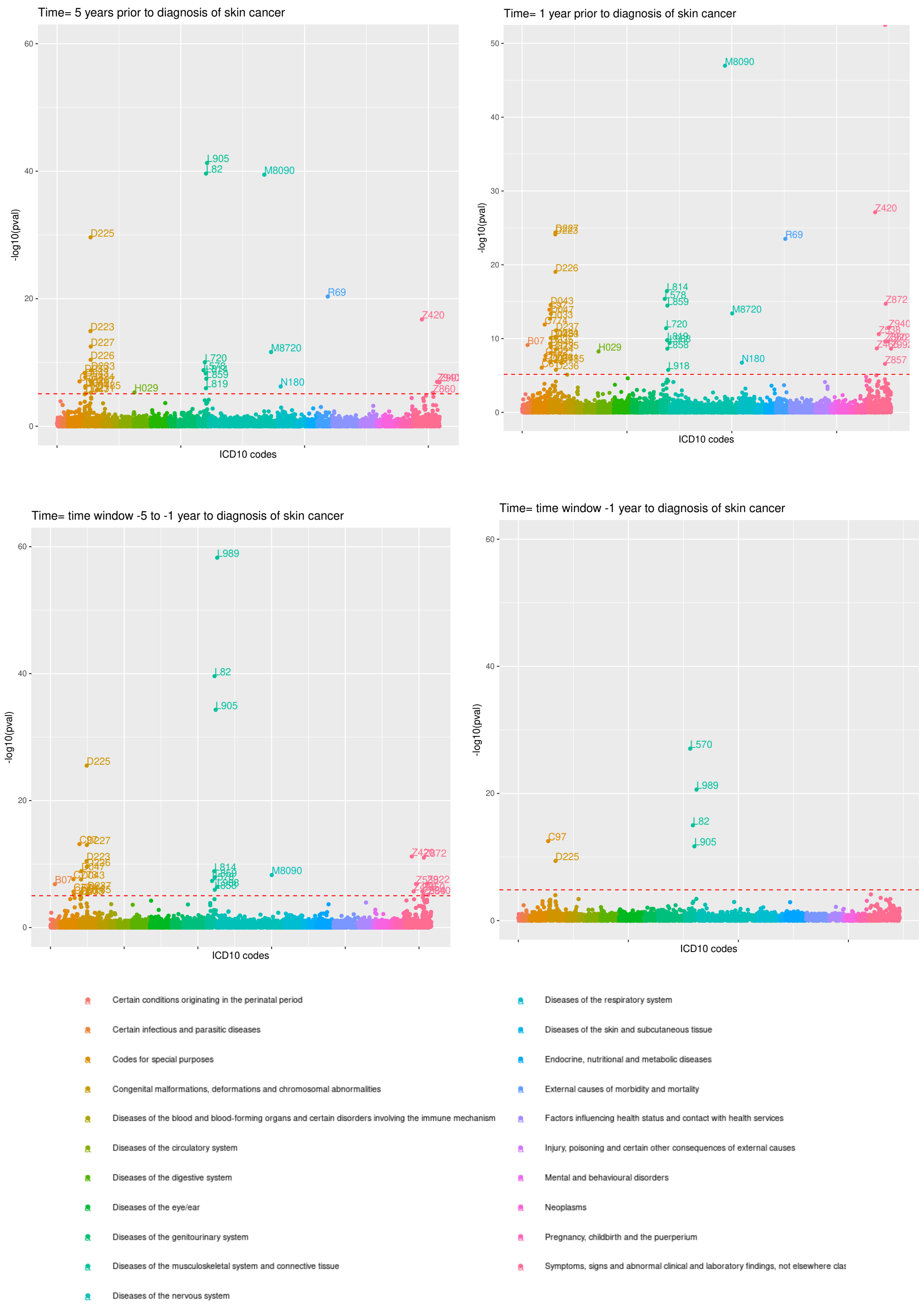

Figure 2: Manhattan plots for univariate analysis. On the top, results for the life-course approach (A, B). On the bottom, results for time windows (C,D). Plot A shows 53 significant diagnoses reaching significance threshold, B has 34, C has 35 and D has 7. *p*-value is extremely low for Z85.8 Personal history of malignant neoplasms and doesn't show on the plot.

In the life-course models, statistically significant comorbidities for 1+ years model, include all significant comorbidities found in the 5+ years model, as these are included in the former time frame. Most of the identified comorbidities for both time frames were classified as neoplasms and diseases of the skin and subcutaneous tissue. In time-windows models, we find the same 7 most significant diseases one year prior to diagnosis and in the -5 to -1 year time window, which suggests that these 7 diseases could be predictive of skin cancer years before it occurs (figure 2). This is further confirmed when comparing the short time window close to diagnosis to the 5+ years model, where again all significant findings are also detected by this latter , at the exception of C97 (*malignant neoplasm of independent primary multiple sites*). These seven diseases, listed in table 2, are strong candidates for predicting skin cancer.

Table 2: Most significant comorbidities from univariate regression on 1+ years time frame

Comorbidities	p-Value
Z85.8 Personal history of malignant neoplasms of other organs and systems	2.68×10^{-92}
L57.0 Actinic keratosis	8.87×10^{-28}
L98.9 Disorder of skin and subcutaneous tissue, unspecified	2.47×10^{-21}
L82 Seborrhoeic keratosis	1.01×10^{-15}
C97 Malignant neoplasms of independent (primary) multiple sites	3.12×10^{-13}
L90.5 Scar conditions and fibrosis of skin	2.07×10^{-12}
D22.5 Melanocytic naevi of trunk	4.06×10^{-10}

To visualize the links between these diseases and infer potential functional links, networks were modelled on the results from the time window analyses (figure 3). The red nodes denote diseases that are present in the top 100 prevalent diseases in cases. The grey nodes are diseases that are not part of the most prevalent diseases in either cases or controls, but still reach Bonferroni significance in the univariate analysis. Networks show a peripheral topology, with a few central nodes of high degree. These central nodes are found to be red coloured, which suggests that the most prevalent diseases in cases are also the diseases with the most links and may have distinct functional links with other less prevalent identified disease found predominantly on the borders of the network.

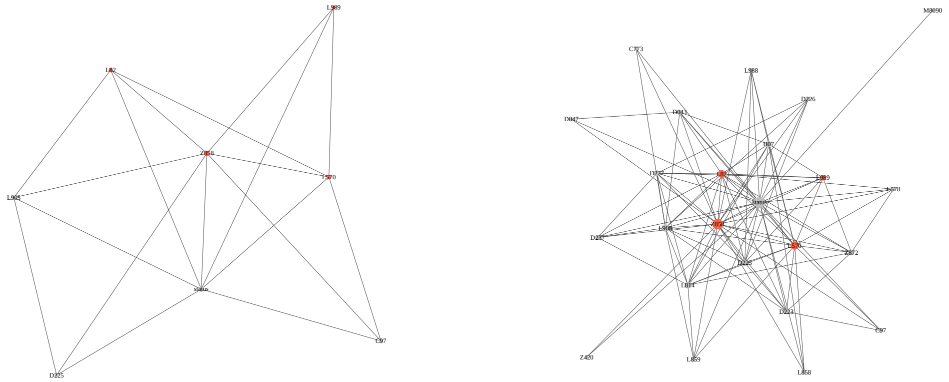

Figure 3: Network for the the significant comorbidities from the univariate logistic regression. '-5 to -1 year' time window (left) and '-1 year to skin cancer diagnosis' time window (right)

3.4 Main analysis: machine learning models

3.4.1 Predictive accuracy

AUC for all datasets and models are presented in table 3. The multivariate logistic regression was used as baseline model, as it's a simple linear effect model used for classification problems, and was fitted separately on comorbidities and covariates. When adding more predictors to the model, the prediction performance worsened. Thus, only statistically significant candidate comorbidities obtained from the univariate logistic regression were used to artificially build a sparser model (i.e. 'top comorbidities'). The 1+ years model outperforms the model using comorbidities 5+ years prior. The highest AUC of 0.659 for the multivariate logistic regression was obtained for top comorbidities in combination with covariates and the lowest AUC of 0.488, was observed when fitting the model for all 8,138 comorbidities 5+ years prior to diagnosis. Covariates and comorbidities don't have a high predictive accuracy alone but jumps to 0.659 when combined. Models with top comorbidities and covariates combined have comparable accuracy to models including biomarkers, which makes the former, the more parsimonious models, preferable for predictions.

Since the comorbidities dataset likely included noise which concealed the signal, a sparse model, the logistic Lasso, was selected next. This model was fitted for all 7 datasets. A tenfold cross validation was performed to identify the optimal λ, the standard error summed to the minimal λ, for each model. Each Lasso model was run 100 times for different samplings of cross validation. The models were considered stable since many of the predictors were included in the variable selection of the Lasso for all 100 runs. When comparing the AUCs for the different Lasso models presented in table 3, it is noticeable that predictive power increases with increasing number of predictors. Hence, the model fitted to all 3 datasets together shows the highest AUC of 0.685 and the models including only 1 dataset had the worst performance.

Support Vector Machines was performed on all datasets. Significant comorbidities were selected from the univariate logistic regression performed. The other comorbidities were filtered out, due to the computationally expensive nature of the SVM algorithm. Linear and radial kernels were selected for all C-classification SVM models, where the cost value was 1, and the gamma value for the radial kernel was the value of the inverse of the dimensions of the different datasets. The AUC values for all the datasets for SVM showed that radial kernels performed better, with the highest AUC being 0.626 for the dataset that combines comorbidities, covariates and biomarkers.

The non-linear RF model showed equivalent performance to both Lasso and multivariate logistic regression. The optimisation of the RF was conducted through a grid search of the best parameters for maximum depth, number of estimators, minimum number of samples per leaf and criterion. Changing the parameters in the grid for a potentially highly influential parameter, led to a modest improvement in the AUC and was not worth the additional computational time required.

Table 3: AUCs for all models run using the different combinations of datasets at 2 time frames

		5+ years prior to diagnosis AUC	1+ years prior to diagnosis AUC
Multivariate logistic regression	Covariates	0.598	
	Comorbidities	0.488	0.550
	Top comorbidities	0.586	0.552
	Top comorbidities + covariates	0.634	***0.659***
	Comorbidities + covariates + biomarkers*	0.529	0.551
	Top comorbidities + covariates + biomarkers*	0.632	***0.659***
Lasso	Covariates	0.597	
	Comorbidities	0.561	0.595
	Comorbidities + covariates	0.628	0.652
	Comorbidities + covariates + biomarkers*	0.656	***0.685***
SVM	Covariates	0.549	
	Comorbidities	0.548	0.582
	Comorbidities + covariates	0.583	0.599
	Comorbidities + covariates + biomarkers*	0.605	***0.626***
RF	Covariates	0.595	
	Comorbidities	0.561	0.587
	Comorbidities + covariates	0.625	0.651
	Comorbidities + covariates + biomarkers*	0.639	***0.665***

*Modelled on incident cases only

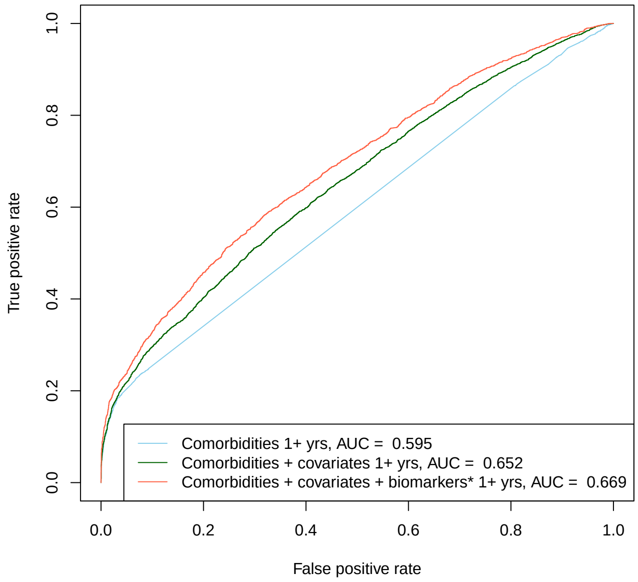

Figure 4: ROC-curve for the Lasso fitted to the 3 different datasets for the time frame up to 1 year before diagnosis of skin cancer.

*Modelled on incident cases only

3.4.2 Most informative features

The most informative features were assessed for the best predictive models that combined covariates, comorbidities and biomarkers, for both Lasso and RF. For Lasso models, the best selected predictive features were almost identical in both time frames: 1+ years and 5+ years which suggests that important features for risk prediction are independent of the time frame, as can be seen in figure 5. All already identified comorbidities in the univariate logistic regression analysis and χ^2-test (Z858, L989, L570 and L82) were selected by Lass.This shows that the two linear methods used for variable selection agree on the most predictive comorbidities.

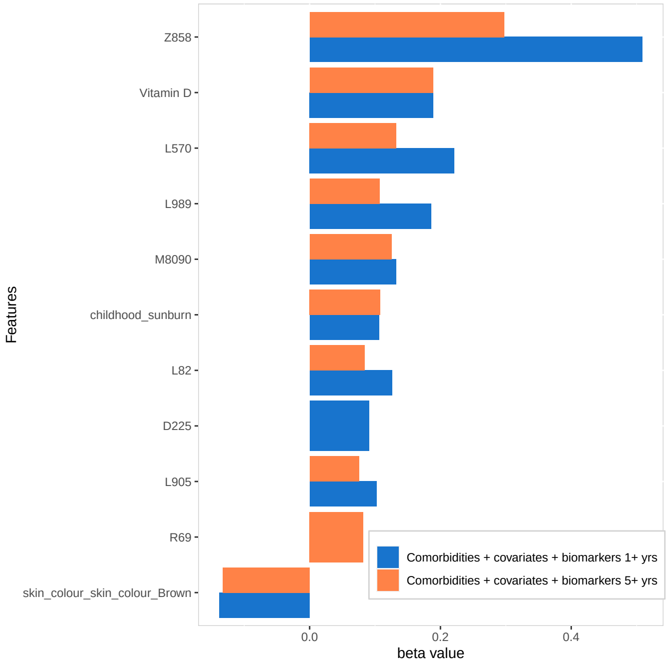

Figure 5: Selection of 10 most important features in Lasso for 1+ years and 5+ years prior to skin cancer diagnosis

RF models showed a completely different set of important features. None of the covariates were included in the best features for skin cancer prediction, contrary to Lasso, where childhood sunburn and skin colour played an important role. The featured biomarkers differed as well, with Vitamin D, total protein and Glycated haemoglobin selected by Lasso, versus HDL cholesterol, Cystatin C and Triglycerides in RF. RF models showed a very unstable selection of features across the considered time frames and combinations of predictors, with no obvious inferential conclusions from best predictive features. The underlying instability observed with RF results most likely come from the greedy nature of the algorithm itself, whereby it gives a series of local optima to give the final feature importance. On the other hand, penalized regression allows for the assessment of a global optima which could explain the high level of stability in feature selection.

3.4.3 Sensitivity Analysis

Overall, predictive accuracy results are close to what is observed in the main analysis, with a slight down performance. This suggests that there is some predictive value in including hospital episodes that occurred more than 5 years before skin cancer diagnosis.

Table 4: Sensitivity analysis using time windows between 1 and 5 years prior to diagnosis and between 1 year prior until the date of diagnosis of skin cancer, presenting AUCs for all models

		-5 until -1 years AUC	-1 year AUC
Multivariate logistic regression	Covariates	0.598	
	Top comorbidities	0.552	0.586
	Top comorbidities + covariates	0.630	0.610
	Top comorbidities + covariates + biomarkers*	0.634	0.603
Lasso	Covariates	0.597	
	Comorbidities	0.552	0.518
	Comorbidities + covariates	0.626	0.605
	Comorbidities + covariates + biomarkers*	0.648	0.621
SVM	Covariates	0.549	
	Comorbidities	0.544	0.518
	Comorbidities + covariates	0.570	0.555
	Comorbidities + covariates + biomarkers*	0.558	0.558
RF	Covariates	0.594	
	Comorbidities	0.561	0.541
	Comorbidities + covariates	0.627	0.599
	Comorbidities + covariates + biomarkers*	0.643	0.616

*Modelled on incident cases only

4 Discussion

4.1 Comparison of the predictive accuracy

To answer whether comorbidities prior to skin cancer diagnosis improve prediction, we compared the predictive accuracy of the models based on the different combinations of groups of predictors. Unsurprisingly, models that included comorbidities only don't perform well. Models with covariates alone didn't perform well either while these known risk factors usually show good performance in the literature. In previous research, Vuong et al. have built a logistic regression model based on hair colour, nevus density, first-degree family history of melanoma, previous melanoma skin cancer and lifetime sunbed use, and reached an overall prediction accuracy of 0.70 [12]. Nevertheless, we systematically see an improvement in prediction scores for models that include both comorbidities and covariates, with the highest performance at 0.66 for the Lasso model ($t = 1 + years$). Since comorbidities and covariates are independent, these results highlight the predicative value of these two sets of predictors, and how comorbidities can be used as an additional source of information to improve risk prediction. Finally, adding biomarkers to the prediction comorbidities/covariates models showed very limited value, with a marginal increase in the Lasso models. This isn't surprising considering that most candidate biomarkers had very restricted functional link with the considered outcome based on the current scientific literature.

To comment on the longitudinal value of health records, one must see how the predictive accuracy of our models vary in the different time frames considered. Out of all considered time frames, the 1+ years model outperformed all other models, which suggests that the more information on patient's clinical history is included in the prediction model, the better the capacity of prediction gets. Moreover, the life-course approach looking at all the comorbidities before 1 year to diagnosis slightly outperformed the -5 to –1 time window, which suggests that there is some value in looking at diseases that occurred 5 years prior to diagnosis. Nonetheless, 5+ health records alone are very partially informative. Finally, adding comorbidity information in the year prior to diagnosis seem to yield minor improvement compared to the covariates alone models. Thus most informative time window seems to occur between 5 years to 1 year prior to skin disease diagnosis.

To comment on the performance of the different predictive models, we showed that both linear models seem to best capture the underlying trend in our data, where RF gives very close performance and shows a very unstable selection of feature across the different time frames considered. For this reason, the Lasso model was defined as the best performing model, which also allowed to draw inferential results given the stability of its prediction. We took a closer look at different performance metrics to evaluate strength and weaknesses of the Lasso model(table 5). The accuracy of 0.629 indicates that overall the number of truly identified skin cancer cases and truly identified non-cases by our model before the cancer would be diagnosed in a hospital setting is satisfactory. However, precision is considerably lower than recall, indicating that the model favours false positives to false negatives. Thus, this model could be useful as a 2^{nd} screening test, since it could help filter out further cases from the 1^{st} screening test, without adding any cost.

Table 5: Accuracy, Precision, Recall an F1-score for the Lasso model fitted to the dataset consisting of comorbidities, covariates and biomarkes 1+ years recorded prior to skin cancer diagnosis

Accuracy	Precision	Recall	F1 score
0.629	0.476	0.687	0.562

4.2 Assessing variable importance

To assess which comorbidities are the most informative in predicting skin cancer risk we assessed which were stably selected in Lasso, our best predictive model, and compared those to the results from univariate regression. For the 1+ years model, 35 out of the 53 ICD10 diagnoses that reached significance threshold in the univariate analysis were stably selected (100% of time in cross validation) by the Lasso model. Other stably selected variables include diagnoses from ICD10 chapters not seen in the univariate analysis, including: Endocrine, nutritional and metabolic diseases, mental and behavioural disorders, and diseases of the circulatory/respiratory/digest system. More than half of the candidate covariates were included in the model 100% of times, which validates known risk factors from the scientific literature such as skin colour or ease of skin tanning. Interestingly, other covariates for which there is no scientific consensus on their predictive risk for skin cancer were stably selected by our model. These included: Alcohol intake, smoking, BMI and diet related variables. The strongest predictors in the Lasso model include personal history of malignant neoplasms, actinic keratosis, unknown disorder of the skin, which also had the lowest p-values in the univariate models in both life-course and time window approaches. In the literature, actnitic keratosis has been proposed as a potential precancerous form of skin cancer which arises from long-term exposure to UV. Similarly, 'squamous cell carcinoma', a benign form of skin cancer, is stably selected in the Lasso models and appears in all univariate models except for the year prior to diagnosis. Known risk factors such as skin colours and childhood sunburn were the strongest covariate predictors. Finally, Vitamin D, which is strongly correlated with sun exposure status and skin type also showed to be strong predictors in our model [18–20].

4.3 Value, limitations and avenues for research

In this work, we showed that information from EHR, especially on life-course episode diagnoses, has some value for predicting the risk of skin cancer. Using a time-stamped approach, we showed that information up to 5 years before skin cancer diagnosis is improving risk prediction compared to using known risk factors alone. Moreover, the findings from the targeted analysis further strengthened our approach since it flagged strong predictors that were already known to be strongly associated to the risk of developing skin cancer in the literature. As a result, this work builds a strong case for the combined use of health records with other lifestyle variables when assessing high-risk patient profiles. Health records and their temporal component are an invaluable source of clinical information that could easily be accessed at no cost. The use of previous diagnoses from EHR over known risk factors warrants further investigation before clinical implementation, but has the potential to automate the detection of high-risk in routine care at no cost. Bypassing expensive awareness campaigns would allow for a more efficient allocation of screening program resources to cost-dependent measures (i.e. sponsoring for sun cream).

One of the main limitations of our work was linked to the nature of the variables available in the UKB. Some important risk factors would have been included upon availability, especially strong known predictors such as UV exposure, use of UV protection, or frequency of solarium use. Given the consistent predictive power seen in the literature, we could have hoped for an overall better predictive accuracy in all of our models. Moreover, the temporal component of our prediction was imperfect. We did not look at the length of disease diagnoses, which could have informed on the severity and chronic aspect of some comorbidities. Future work could try to consider previous diagnoses severity and duration in their models. Furthermore, time collinearity is a challenge in any risk prediction models with a temporal component, and further work would need to disentangle the role of time varying covariates like sun exposure. Similarly, this work did not explore the heterogeneous nature of the considered cancer while more than half of the cases came from the single subcategory 'unspecified malignant neoplasm of skin of unspecified parts of face'. Further work should address histological subtypes in a rigorous manner, for instance by clustering cases as in a preliminary step and then perform classification tasks in a second step. Since benign forms of skin cancer were excluded from our cases, controls in our sample could have well ended with undiagnosed forms of skin cancers. Given the availability of other national biobanks, there is opportunity to replicate our approach using health records from different ethnic groups and subsequently pave the way to a systematic integration of health records in routine care screening for skin cancer.

Authors contributions

Carolina Richheimer: Analysis plan, Methodology, Dataset pre-processing, Formal analysis, Writing original draft, Writing - review & editing, Visualization. Eléonore Schneegans: Analysis plan, Methodology, Dataset pre-processing, Formal analysis, Writing original draft, Writing - review & editing, Visualization. Demetris Hajivassiliou: Methodology, Dataset pre-processing, Formal analysis. Sarvesh Dhungana: Dataset pre-processing, Formal analysis, Github. Luca Ramelli: Formal analysis

References

[1] Nawal Alsadi. *MicroRNA-200b Signature in the Prevention of Skin Cancer Stem Cells by Polyphenol-Enriched Blueberry Preparation (PEBP).* PhD thesis, Université d'Ottawa/University of Ottawa, 2016.

[2] National Cancer Intelligence Network (NCIN). Non-melanoma skin cancer in england, scotland, northern ireland, and ireland, 2013.

[3] Richard P Gallagher, Tim K Lee, Chris D Bajdik, and Marilyn Borugian. Ultraviolet radiation. *Chronic Diseases and Injuries in Canada*, 29, 2010.

[4] Howard K Koh, Alan C Geller, Donald R Miller, Ted A Grossbart, and Robert A Lew. Prevention and early detection strategies for melanoma and skin cancer: current status. *Archives of dermatology*, 132(4):436–443, 1996.

[5] DM Parkin, D Mesher, and P Sasieni. 13. cancers attributable to solar (ultraviolet) radiation exposure in the uk in 2010. *British journal of cancer*, 105(2):S66–S69, 2011.

[6] Mackenzie R Wehner, Mary-Margaret Chren, Danielle Nameth, Aditi Choudhry, Matthew Gaskins, Kevin T Nead, W John Boscardin, and Eleni Linos. International prevalence of indoor tanning: a systematic review and meta-analysis. *JAMA dermatology*, 150(4):390–400, 2014.

[7] Barbara A Gilchrest, Mark S Eller, Alan C Geller, and Mina Yaar. The pathogenesis of melanoma induced by ultraviolet radiation. *New England Journal of Medicine*, 340(17):1341–1348, 1999.

[8] JÒRGEN Lock-Andersen, Krzysztof T Drzewiecki, and Hans Christian Wulf. Eye and hair colour, skin type, and constitutive skin pigmentation as risk factors for basal cell carcinoma and cutaneous malignant melanoma. *ACTA DERMATOVENEREOLOGICA-STOCKHOLM-*, 79:74–80, 1999.

[9] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Fröhling, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019.

[10] Adria Romero Lopez, Xavier Giro-i Nieto, Jack Burdick, and Oge Marques. Skin lesion classification from dermoscopic images using deep learning techniques. In *2017 13th IASTED international conference on biomedical engineering (BioMed)*, pages 49–54. IEEE, 2017.

[11] Maxwell Salvatore, Lauren J Beesley, Lars G Fritsche, David Hanauer, Xu Shi, Alison M Mondul, Celeste Leigh Pearce, and Bhramar Mukherjee. Phenotype risk scores (phers) for pancreatic cancer using time-stamped electronic health record data: Discovery and validation in two large biobanks. *Journal of Biomedical Informatics*, 113:103652, 2021.

[12] Kylie Vuong, Bruce K Armstrong, Elisabete Weiderpass, Eiliv Lund, Hans-Olov Adami, Marit B Veierod, Jennifer H Barrett, John R Davies, D Timothy Bishop, David C Whiteman, et al. Development and external validation of a melanoma risk prediction model based on self-assessed risk factors. *JAMA dermatology*, 152(8):889–896, 2016.

[13] David Roffman, Gregory Hart, Michael Girardi, Christine J Ko, and Jun Deng. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Scientific reports*, 8(1):1–7, 2018.

[14] Hsiao-Han Wang, Yu-Hsiang Wang, Chia-Wei Liang, and Yu-Chuan Li. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer. *JAMA dermatology*, 155(11):1277–1283, 2019.

[15] Screening for melanoma skin cancer | Cancer Research UK. `https://www.cancerresearchuk.org/about-cancer/melanoma/getting-diagnosed/screening`, 2021. [Online; accessed 21 March 2021].

[16] Francisco S Roque, Peter B Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Søeby, Søren Bredkjær, Anders Juul, Thomas Werge, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol*, 7(8):e1002141, 2011.

[17] NHS. Health Survey for England : Cardiovascular diseases. `http://healthsurvey.hscic.gov.uk/media/78646/HSE17-CVD-rep.pdf`, 2017. [Online; accessed 21 March 2021].

[18] Ann R Webb. Who, what, where and when—influences on cutaneous vitamin d synthesis. *Progress in biophysics and molecular biology*, 92(1):17–25, 2006.

[19] Mark D Farrar, Richard Kift, Sarah J Felton, Jacqueline L Berry, Marie T Durkin, Donald Allan, Andy Vail, Ann R Webb, and Lesley E Rhodes. Recommended summer sunlight exposure amounts fail to produce sufficient vitamin d status in uk adults of south asian origin. *The American journal of clinical nutrition*, 94(5):1219–1224, 2011.

[20] Actinic Keratosis. `https://www.skincancer.org/skin-cancer-information/actinic-keratosis//`, 2021. [Online; accessed 21 March 2021].

5 Appendix

Table 6: Skin cancer subtypes

ICD CODE	DESCRIPTION	Title
C430	Malignant melanoma of lip	Certain infectious and parasitic diseases
C431	Malignant melanoma of eyelid, including canthus	Neoplasms
C432	Malignant melanoma of ear and external auricular canal	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
C433	Malignant melanoma of other and unspecified parts of face	Endocrine, nutritional and metabolic diseases
C434	Malignant melanoma of scalp and neck	Mental and behavioural disorders
C435	Malignant melanoma of trunk	Diseases of the nervous system
C436	Malignant melanoma of upper limb, including shoulder	Diseases of the eye and adnexa
C437	Malignant melanoma of lower limb, including hip	Diseases of the ear and mastoid process
C438	Malignant melanoma of overlapping sites of skin	Diseases of the circulatory system
C439	Malignant melanoma of skin, unspecified	Diseases of the respiratory system
C440	Other and unspecified malignant neoplasm of skin of lip	Diseases of the digestive system
C441	Other and unspecified malignant neoplasm of skin of eyelid, including canthus	Diseases of the skin and subcutaneous tissue
C442	Other and unspecified malignant neoplasm of skin of ear and external auricular canal	Diseases of the musculoskeletal system and connective tissue
C443	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	Diseases of the genitourinary system
C444	Other and unspecified malignant neoplasm of skin of scalp and neck	Pregnancy, childbirth and the puerperium
C445	Other and unspecified malignant neoplasm of skin of trunk	Certain conditions originating in the perinatal period
C446	Other and unspecified malignant neoplasm of skin of upper limb, including shoulder	Congenital malformations, deformations and chromosomal abnormalities
C447	Other and unspecified malignant neoplasm of skin of lower limb, including hip	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
C448	Other and unspecified malignant neoplasm of overlapping sites of skin	Injury, poisoning and certain other consequences of external causes
C449	Other and unspecified malignant neoplasm of skin, unspecified	External causes of morbidity and mortality

Table 7: ICD10 chapters

Chapter	Block	Title
I	A00–B99	Certain infectious and parasitic diseases
II	C00–D48	Neoplasms
III	D50–D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00–E90	Endocrine, nutritional and metabolic diseases
V	F00–F99	Mental and behavioural disorders
VI	G00–G99	Diseases of the nervous system
VII	H00–H59	Diseases of the eye and adnexa
VIII	H60–H95	Diseases of the ear and mastoid process
IX	I00–I99	Diseases of the circulatory system
X	J00–J99	Diseases of the respiratory system
XI	K00–K93	Diseases of the digestive system
XII	L00–L99	Diseases of the skin and subcutaneous tissue
XIII	M00–M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00–N99	Diseases of the genitourinary system
XV	O00–O99	Pregnancy, childbirth and the puerperium
XVI	P00–P96	Certain conditions originating in the perinatal period
XVII	Q00–Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00–R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00–T98	Injury, poisoning and certain other consequences of external causes
XX	V01–Y98	External causes of morbidity and mortality
XXI	Z00–Z99	Factors influencing health status and contact with health services
XXII	U00–U99	Codes for special purposes