PLOS ONE

RESEARCH ARTICLE

Improving five-year survival prediction via multitask learning across HPV-related cancers

Andre Goncalves[1]*, Braden Soper[1], Mari Nygård[2], Jan F. Nygård[2], Priyadip Ray[1], David Widemann[1], Ana Paula Sales[1]

1 Lawrence Livermore National Laboratory, Livermore, CA, United States of America, **2** Cancer Registry of Norway, Oslo, Norway

* goncalves1@llnl.gov

Citation: Goncalves A, Soper B, Nygård M, Nygård JF, Ray P, Widemann D, et al. (2020) Improving five-year survival prediction via multitask learning across HPV-related cancers. PLoS ONE 15(11): e0241225. https://doi.org/10.1371/journal.pone.0241225

Editor: Randall J. Kimple, University of Wisconsin, UNITED STATES

Received: February 26, 2020

Accepted: October 11, 2020

Published: November 16, 2020

Data Availability Statement: The data is publicly available and can be requested from the SEER website: https://seer.cancer.gov/data/.

Funding: AG, BS, PR, DW, MA, AS were supported by the U.S. Department of Energy through the Lawrence Livermore National Laboratory. MN and JFN work was supported by the Cancer Registry of Norway. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Oncology is a highly siloed field of research in which sub-disciplinary specialization has limited the amount of information shared between researchers of distinct cancer types. This can be attributed to legitimate differences in the physiology and carcinogenesis of cancers affecting distinct anatomical sites. However, underlying processes that are shared across seemingly disparate cancers probably affect prognosis. The objective of the current study is to investigate whether multitask learning improves 5-year survival cancer patient survival prediction by leveraging information across anatomically distinct HPV related cancers. Data were obtained from the Surveillance, Epidemiology, and End Results (SEER) program database. The study cohort consisted of 29,768 primary cancer cases diagnosed in the United States between 2004 and 2015. Ten different cancer diagnoses were selected, all with a known association with HPV risk. In the analysis, the cancer diagnoses were categorized into three distinct topography groups of varying specificity. The most specific topography grouping consisted of 10 original cancer diagnoses differentiated by the first two digits of the ICD-O-3 topography code. The second topography grouping consisted of cancer diagnoses categorized into six distinct organ groups. Finally, the third topography grouping consisted of just two groups, head-neck cancers and ano-genital cancers. The tasks were to predict 5-year survival for patients within the different topography groups using 14 predictive features which were selected among descriptive variables available in the SEER database. The information from the predictive features was shared between tasks in three different ways, resulting in three distinct predictive models: 1) Information was not shared between patients assigned to different tasks (single task learning); 2) Information was shared between all patients, regardless of task (pooled model); 3) Only relevant information was shared between patients grouped to different tasks (multitask learning). Prediction performance was evaluated with Brier scores. All three models were evaluated against one another on each of the three distinct topography-defined tasks. The results showed that multitask classifiers achieved relative improvement for the majority of the scenarios studied compared to single task learning and pooled baseline methods. In this study, we have demonstrated that sharing information among anatomically distinct cancer types can lead to improved predictive survival models.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Humans have the ability to transfer relevant knowledge from previous experiences to new ones, mastering them more easily and faster. Analogously in machine learning, multitask learning (MTL) methods attempt to improve model generalization and performance by using shared representations to exploit commonalities and differences across related tasks, while avoiding using information from unrelated tasks [1, 2]. MTL has shown to be an effective approach for overcoming the challenges of low quality datasets, such as scarce or highly skewed training data, that can degrade predictive performance [3–6].

While MTL has a history of over 20 years in machine learning, only recently has it started making its way into the field of predictive oncology. In the vast majority of such applications, MTL methods (and the related methods of transfer learning) are used to allow sharing of information across related datasets of the same type of cancer (e.g., breast cancers [7], skin cancers [8], and lung or prostate cancers [9]). In other words, tasks are defined in terms of datasets limited to a specific type of cancer.

However, as stark as the differences between distinct cancers may be, there probably exist underlying processes that are shared, even across seemingly very disparate cancer types. For example, it is known that human papillomavirus (HPV) plays a role in roughly 5% of all cancers worldwide [10]. While HPV is considered a necessary cause of cervical cancer, it is also linked to cancers in other anatomical sites, with rates varying according to susceptibility to oncogenic types of HPV. For instance, 90% of anal and 74% of vaginal cancers appear to be induced by HPV [11, 12]. Nearly 30% of penile and vulvar cancers may be caused by this virus [13, 14]. Likewise, about 30% of all head and neck cancers are HPV positive [15]. This proportion has been increasing over time, approaching 80% of incident tonsil cancers in some countries [16]. In addition to the similarities in the etiology, better survival has been observed among patients with HPV-related cancers who were tested positive for HPV, such as patients with penile cancer [17, 18], nasopharyngeal cancer [19], anal cancer [20], and vulvar cancer [17, 21]. In contrast to traditional tobacco- and alcohol-associated oropharyngel cancers, patients with positive HPV findings have demonstrated improved survival and significantly higher cure rates [22–24]. In spite of the commonalities described for HPV-related cancers, these patients have been managed through different oncology disciplines such as gynecologic oncologist, otolaryngologist or head and neck surgeons, onco-gastroenterologists, and onco-urologists. With low annual incidence rate, some less than 4 per 100,00 individuals, the clinical experience is not only modest, but also rarely shared across the segregated fields of oncology.

In the majority of developed countries, oncology care units regularly report selected information on each cancer patient to the nation-wide cancer registry surveillance program. While the information sent to the surveillance programs is less nuanced than information available in the clinics, it is standardized following international conventions and the quality of the selected variables are formally assured for accuracy over time. Data-driven statistics are computed on incidence, mortality and survival, informing aspects on evidence-based care and changes in disease burden on a regular basis. Consequently, cancer registry data has been a valuable source of research for hypothesis generation and testing [25–27].

Encouraged by recent developments in the field of multitask learning (MTL) and increased availability of accurate registry data, our aim is to investigate whether MTL can be used to leverage registry data for 5-year survival prediction. Potentially, the prediction performance for rare cancer types with limited patient records may be improved via leveraging patient records of similar cancer types but with larger number of patient records. Hence, as opposed to sharing information across different datasets of the same cancer type, we approach MTL for cancer from a different perspective: leverage MTL to share information across anatomically

distinct cancer types. In this context, HPV related cancers serve an excellent model to test the accuracy of the different survival prediction models.

Many existing MTL methods rely on explicit assumptions about the relationships between tasks. These assumptions are incorporated into machine learning algorithms through specifically designed priors [6, 28] or regularization functions [29–31]. More recently proposed methods are capable of learning the relationships between tasks from the data and incorporating this information into the learning process [4, 5]. The MTL method proposed in this study is built upon the MSSL approach proposed in Goncalves et al. [5], which, aside from learning tasks coefficients, also estimates the relationship among the tasks represented as an undirected graph. This is useful in inferring how information is shared across the different tasks during model training.

When predicting whether or not patient will survive for at least five years (i.e., binary classification problem) based on patient features, it is important that the learning algorithms are capable of handling censored data, to avoid potential bias [32]. In Vock et al. [33], a general-purpose technique for adapting machine learning algorithms to right-censored, time-to-event data is presented. The method is based on computing inverse probability of censoring weights (IPCW) which are then used to construct a weighted loss function and weighted performance metrics used in the training and testing of the given learning algorithm. In this paper, we extend the MSSL formulation of Goncalves et al. [5] to appropriately handle right-censored data using the inverse probabilities of censoring weights.

In summary, the objective of the current study is to investigate whether multitask learning improves 5-year survival cancer patient survival prediction by leveraging information across anatomically distinct HPV related cancers.

Material and methods

Cohort selection

Data was obtained from the Surveillance, Epidemiology, and End Results (SEER) program database [34], which provides de-identified information on cancer statistics of the United States' population. Specifically, data from nine SEER registries were used: Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah. Although data is available for cases diagnosed from 1973 through 2015, we only used data from 2004 onwards due to the fact that in that year there was a major change in the criteria used for both cancer stage and grade definitions. In this study, we focus on cancers in anatomical sites for which evidence of an association between HPV and cancer has been established. Based on evidence suggesting etiological link between infection with human papillomavirus (HPV) infection and cancer, we selected the following cancer sites to the study using the International Classification of Diseases for Oncology, 3rd edition (ICD-O-3) topography codes [35]: cervix (C53), anus and anal canal (C21), vulva (C51), vagina (C52), and penile cancer (C60). Regarding head and neck cancers, we included all cancers coded as C01, C09, and C10 where all sub-sites are HPV related [36]. C02, C05 and C11 are types of head and neck sited with mixed etiology in respect to HPV infection. Regarding cancers coded as C02 (other and unspecified parts of tongue), sub-sites C02.0-3 and C02.9 have been referred to as not HPV-related and therefore not included, while sub-sites C02.4 and C02.8 are linked to HPV infection and are included in this study. Regarding cancers coded as C05 (palate) sub-sites C05.1 (soft palate) and C05.2 (uvula) are HPV related (included) while sub-sites C05.8 and C05.9 are typically not HPV related (excluded). We did not include cancers in lip, gum, floor and other unspecified parts of mouth, cancers in glands, sinuses and hypopharynx due to the lack of strong evidence of being associated with HPV-infection. We also excluded C11

Table 1. Population data description per anatomical site, classified by ICD-O-3 topography codes: Number of cancer cases (N), age information, and 5-year survival rate.

ICD-O-3 code	Anatomical site	N	Age			5 year Survival (%)
			mean	median	min–max	
C01	Base of tongue	4421	61.03	60	19–102	63.3
C02	Other/Unsp tonge *	244	59.79	59	26–96	59.7
C05	Palate**	488	61.06	61	20–101	56.2
C09	Tonsil	5511	57.91	57	19–102	72.0
C10	Oropharynx	943	60.79	59	25–94	42.2
C21	Anus & Anal canal	4287	60.29	59	19–105	63.5
C51	Vulva	2733	66.11	66	19–102	62.6
C52	Vagina	645	65.43	64	23–100	41.9
C53	Cervix Uteri	9729	48.90	47	19–103	68.7
C60	Penis	767	65.85	66	26–98	58.0
	Total	29,768	57.05	57	19–105	62.5

*Subsite C02.0-3/9 excluded.

**Subsite C05.8/9 excluded.

https://doi.org/10.1371/journal.pone.0241225.t001

(nasopharynx), which is linked to infection with Ebstein-Barr virus [37] and where the HPV-etiology is not firmly established [38] partly because this site is difficult to study as the deep structures of the skull base related to the nasopharynx are inaccessible to routine clinical examinations.

In addition we removed all cases for which: 1) survival time and/or event information were missing; 2) age at diagnosis was under 18 years; 3) the case is a pre-cancer (cases in which the cancer stage is 0); 4) the number of survival months after being diagnosed is zero; 5) the cancer was not the first diagnosed cancer case of the patient.

After exclusions, the cohort consisted of 29,768 primary cancer cases diagnosed in the United States between 2004 and 2015 with anatomical sites associated with HPV risk. The cohort contains a total of 11,887 men and 17,818 women, with mean age of 57 years. A descriptive table of the population data used in this paper is shown in Table 1.

Tasks definition

In MTL an important step is to define the "tasks". In many applications of MTL, tasks are easily identified. In our application, defining the tasks is not straightforward, as there are many possible ways to do so. We next provide details of our criteria for tasks definition.

Because we aim to implement MTL methodologies across cancer types, the task definition will be determined by how cancer "type" is defined. For the purposes of this study we will define a cancer type based on the anatomical location of the cancer diagnosis. We consider three distinct strategies for grouping individual cancer sites with a decreasing degree of specificity. The most specific grouping, *Topography group 1* (TP1), consists of 10 cancer sites which are defined by the first two digits of the ICD Topography code. In the next most specific grouping, *Topography group 2* (TP2), cancers from *Topography group 1* were grouped by organ, resulting in a total of six cancer sites. Finally, in the least specific grouping, *Topography group 3* (TP3), related organs are grouped into broad anatomical regions, resulting in just two cancer sites. Table 2 presents the groupings utilized in our experiments.

Given a particular topography group, a task is defined as the binary classification problem of predicting whether a patient with a cancer diagnosis at a particular cancer cite (as defined

Table 2. ICD-O-3 codes included in the cohort and topography group division used in our experiments.

	PRIMSITE	Topography Group 1		Topography Group 2		Topography Group 3	
		Task ID		Task ID		Task ID	
Tasks	C01	T-01:	Base Tongue	T-01:	Tongue	T-01:	Head & Neck
	C02	T-02:	Other/Unsp. Parts of Tongue				
	C05	T-03:	Palate	T-02:	Palate		
	C09	T-04:	Tonsil	T-03:	Oropharynx & Tonsils		
	C10	T-05:	Oropharynx				
	C21	T-06:	Anus & Anal Canal	T-04:	Anus & Anal Canal	T-02:	Ano-Genital
	C51	T-07:	Vulva	T-05:	Genital Female		
	C52	T-08:	Vagina				
	C53	T-09:	Cervix Uteri				
	C60	T-10:	Penis	T-06:	Genital Male		

Topography group 1 through 3 represent groupings of ICD-O-3 codes with decreasing degree of specificity.

https://doi.org/10.1371/journal.pone.0241225.t002

by the topography group) will survive less than five years or more from the time of diagnosis. Demographic and cancer-related information are used as predictive features. By looking into distinct topography groups of anatomical specificity, we aim to investigate the performance of the multitask learning methods under the different strategies of splitting the data into tasks.

Variable selection and re-coding

Inspired by the work of Lynch et al. [39], a total of 14 predictive variables, listed in Table 3, are chosen. Variables with prefix "X_" are derived from features in the SEER database. The process is described in the next paragraph. Variables indicated with an asterisk are re-coded versions of the original variables in the SEER database. The re-coding process for each of these variables are presented in the S1 File. The main reason for re-coding is to group similar categories into larger groups, reducing the number of categories for the modelling to be more effective, without loosing predictive power.

X_PRIMSITE_1 is derived from SEER variable PRIMSITE, such that the first two digits of PRIMSITE are assigned to X_PRIMSITE_1. X_TUMSIZ_COMB_NUM is derived from SEER variable CSTUMSIZ, which contains both numeric values (actual sizes of tumors in millimeters) as well as ordinal values (codes that indicate that the tumor size lies within a range of 10 millimeters). In X_TUMSIZ_COMB_NUM all the numeric values are retained while the categorical values were mapped to the median of the range of the bin to which the given tumor was assigned.

Variables other than the ones with prefix "X_" or "*" are maintained as originally coded.

Feature encoding

To be used by the machine learning algorithms, all predictive variables must be represented numerically. This process is referred to as *feature encoding*. Different encoding strategies are applied to different types of predictive variables, the selection of which depends on the characteristics of the variable under consideration. In what follows we use the terms variable and feature interchangeably.

Numerical predictive variables such as AGE_DX and X_TUMSIZ_COM_NUM are already real numbers, so they can directly be used as features by the methods without any additional encoding.

Table 3. Variables obtained from SEER database and used as 5-year survival prediction features.

SEER variable	Description	Type	# Levels
REG	Registry ID	Categorical	9 levels
AGE_DX	Age at diagnosis	Numerical	-
SEX	Sex	Categorical	2 levels
RAC_RECA	Race recode (White, Black, Other)	Categorical	5 levels
SURGSCOF	Scope of regional lymph node surgery	Categorical	3 levels
HISTREC	Histology recode, broad groupings	Categorical	3 levels
*GRADE	(Recoded) Grade	Numerical	-
*DAJCCT	(Recoded) AJCC 'T' component (6th Ed.)	Numerical	-
*DAJCCN	(Recoded) AJCC 'N' component (6th Ed.)	Numerical	-
*DAJCCM	(Recoded) AJCC 'M' component (6th Ed.)	Numerical	-
*DAJCCSTG	(Recoded) AJCC 'stage group' component (6th Ed.)	Numerical	-
*SURGPRIF	(Recoded) Surgery of primary site, generic	Numerical	-
X_PRIMSITE_1	First two digits of ICD-O-3 code for anatomical site	Categorical	10 levels
X_TUMSIZ_COMB_NUM	Tumor size	Numerical	-
SURV_TIME_MON	Survival time in months	Numerical	-
X_SURV_TIME_5Y	Five year survival	Binary	-

Variables indicated with "*" are re-coded versions of the original SEER variables (see S1 File). Variables with prefix "X_" refer to modified variables derived from original SEER variables (see text). All other variables are kept as originally coded. SURV_TIME_MON and X_SURV_TIME_5Y are predictands of the model, thus they are not considered in the input feature set. AJCC stands for American Joint Committee on Cancer.

https://doi.org/10.1371/journal.pone.0241225.t003

Variables that are purely categorical (qualitative variables with no clear ordering or associated numerical values) are usually represented via one-hot-encoding. In this encoding strategy the variable is represented by a binary vector of the same size as the number of categories. The vector contains zeros everywhere except for the position corresponding to the given category, in which case a 1 appears. This is equivalent to introducing dummy variables for categorical variables in standard regression analysis. The following variables are encoded using this strategy: REG, SEX, RAC_RECA, HISTREC, X_PRIMSITE_1, SURGSCOF, and X_SURGPRIF_GEN. Cases with "Unknown" / "Not Applicable" categories are treated as an additional category for that feature.

Stage-related variables have an intrinsic ordering of the categories. For example, the categories in the feature DAJCCSTG have an increasing order related to the severity of the cancer diagnosis: stage I, stage II, stage III, and stage IV. To preserve this ordinal relationship we used a label encoding strategy in which each stage category is assigned an integer value corresponding to its relative severity. For example, in the feature DAJCCSTG stage I is represented by '1', stage II by '2' and so on. However, in the SEER dataset these variables also have an "Unknown" / "Not applicable" category that breaks the natural ordering of the categories. To deal with these cases, we propose to represent this particular category as the *empirical mean* of all assigned integer valued labels in the observed data. Note that this process is a type of imputation which treats the "Unknown" / "Not applicable" category as data that is missing at random. We used this encoding approach for the following features: GRADE, DAJCCT, DAJCCN, DAJCCM, and DAJCCSTG.

Outcome definition

A binary outcome, X_SURV_TIME_5Y, derived from the SEER feature SRV_TIME_MON, was used as the outcome variable. A value of 1 indicates that a patient has survived at least five

years from the time of diagnosis, and a value of 0 indicates that the patient survived less than five years from the time of diagnosis. Five-year survival has been the de facto method for reporting cancer survival in major epidemiological studies, the most recent being the Concorde Programme [40]. The use of five-year survival originates from the fact that until recently, cancer was a fatal disease and patients surviving for that long could be considered cured [41]. Although the proportion of patients who survive for 5 years has been increasing over the years, it remains a widely used benchmark, even though it cannot be directly interpreted as the proportion of patients who are cured [42].

Censored cases

The censor variable used in this study was built from SEER's variable `STAT_REC`, which describes whether the patient is dead or alive at the end of follow-up. All follow-up is censored at the cut-off date (Dec 31st, 2015). Any patient that dies after this date is considered alive as of the cut-off date. Since we are focusing on 5-year survival prediction, any patient that is alive but its survival time (`SRV_TIME_MON`) is less than 60 months, due to the cut-off date, is considered censored. Therefore, all alive cases that were diagnosed after 2011 are censored. If `STAT_REC` indicates death, then it is uncensored. In case the patient is alive and has already survived for at least 60 months, then it is uncensored.

Methods

To evaluate our hypothesis that combining data from apparently disparate cancer types could lead to model performance improvements, a multitask classifier was compared against two single task baselines. The three methods are described below in the section *Classifiers*.

Our learning tasks are classification problems using distinct datasets. We denote by T the number of tasks, d the number of features in each dataset, assumed to be identical for all learning tasks, and n_t the number of samples for the t-th task. $\mathbf{X}_t \in \mathbb{R}^{n_t \times d}$ and $\mathbf{y}_t \in \{0,1\}^{n_t}$ are the feature (covariate) matrix and the binary outcome vector for the t-th task. $\mathbf{W} \in \mathbb{R}^{d \times T}$ is the MTL parameter matrix, where columns are vector parameters $\mathbf{w}_t \in \mathbb{R}^d$, $t = 1, \ldots, T$, for each task. For any matrix $\mathbf{A}$, $\mathrm{tr}(\mathbf{A})$ is the trace operator and $\|\mathbf{A}\|_1$ is the ℓ_1-norm of matrix $\mathbf{A}$, defined as the sum of the absolute values of its entries.

Treatment of censored data

We applied a general-purpose technique for adapting machine learning algorithms to right-censored time-to-event data [33]. Inverse probability weighting is a method of constructing estimators and likelihood functions that account for sampling biases and missing data [43]. The idea is to use the probability of being sampled, or estimates thereof, to calculate weights that adjust the number of unlikely samples by inflating their representation in the observed data to better reflect the true sampling population. For example, suppose the probability of being sampled is known, and a particular sample x_i has a probability of being sampled $p_i > 0$. The value $\frac{1}{p_i}$ can be used to weight the observed sample x_i when constructing estimators or likelihood functions to essentially create a set of pseudo-samples that increases the effective sample size of the observed data. Intuitively if p_i is close to one, then the sample was very likely to have been selected, thus no adjustment is needed. Consequently, $\frac{1}{p_i}$ is also close to one resulting in minimal adjustments. On the other hand if p_i is very small, then there are many more samples similar to x_i in the true population that were not selected. To represent these samples, the inverse probability weight $\frac{1}{p_i}$, which is now much larger than one, is used to inflate the number of samples similar to x_i in the observed data. More concretely, if $p_i = \frac{1}{10}$ then on average for

every x_i in the observed data, there are 10 such samples in the true population that were not chosen. Thus we inflate the number of samples of the type x_i by a factor of $\frac{1}{p_i} = 10$.

Censoring is a type of sampling bias, and the method of inverse probability of censoring weights (IPCW) constructs weights that use the empirical distribution of censoring times to compute inverse probability weights to adjust estimators and likelihood functions in survival models. To see how censoring can lead to biased predictions in the case of binary classification, consider the following standard survival analysis set-up. Indexing patients by i, let t_i be the survival time, c_i the censoring time and $\mathbf{x}_i$ the features of patient i. We assume the existence of a joint probability distribution $P(c, t, \mathbf{x})$ such that the *complete data* is $(c_i, t_i, \mathbf{x}_i) \sim P(c, t, \mathbf{x})$. Defining $v_i = \min\{c_i, t_i\}$ and $\delta_i = I(t_i < c_i)$, where $I(\cdot)$ is the indicator function, the *observed data* is $(v_i, \delta_i, \mathbf{x}_i)$. Traditional survival analysis seeks to make inference about the distribution $P(t|\mathbf{x})$ given the observed data $\{v_i, \delta_i, \mathbf{x}_i\}$. In the binary classification problem, we want to predict whether patient i survives for $\tau > 0$ years after the date of diagnosis, given the data $\mathbf{x}_i$. Define the binary random variable $y_i = I(t_i \geq \tau)$ with expected value $\pi(\mathbf{x}_i) = E(y_i|\mathbf{x}_i)$. In the presence of right-censoring, the value of y_i will be unknown for some patients, namely those patients with $v_i < \tau$ and $\delta_i = 0$. Restricting analysis to patients with known y_i will lead to bias in the predictions since patients with small event times, i.e. small t_i, are less likely to be censored. Thus we will over sample patients with $y_i = 0$, leading to potentially biased predictions.

To correct for this oversampling, IPCW can be used to essentially artificially inflate the dataset by weighting the influence of uncensored individuals with large event times. The intuition behind this weighting is that the longer a patient survives, the more unobserved patients with similar survival times there are that dropped out due to censoring. These longer-surviving patients are weighted to represent these unobserved censored patients. Weights are computed via inverse probabilities [43] and require estimates of the censoring distribution's survival function (the complement of the cumulative distribution function). If we assume that the censoring distribution is independent of event times and patient features, i.e. $P(c, t, \mathbf{x}) = P(c)P(t, \mathbf{x})$, then we can estimate the censoring distribution's survival function via the Kaplan-Meier estimator, which is defined as

$$\hat{G}(t) = \prod_{i:v_i<t} \left(1 - \frac{k_i}{m_i}\right), \tag{1}$$

where k_i are the number of events observed at time v_i and m_i is the number of individuals who have not yet experienced an event and are not yet censored just before time v_i. The IPCW for patient i is then defined as

$$\omega_i = \begin{cases} \dfrac{\delta_i}{\hat{G}(\min\{v_i, \tau\})} & \text{if } \min\{t_i, \tau\} < c_i, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that patients who are censored before the threshold τ have a weight of zero and do not contribute directly to the data, but instead are incorporated indirectly through the weights. As outlined in [33], these weights can be used to adjust loss functions and performance metrics to account for right-censored data.

Classifiers

In this study, in addition to the MTL classifier, two baseline methods were considered: Single task learning (STL) models and pooled models. These two baselines represent the two extremes in the spectrum of information sharing across cancer types. At one end of the

spectrum, STL utilizes one model per task, so there is no sharing of any information across tasks. At the other end, the pooled model utilizes a single model for all tasks, so all information is shared across all tasks. The MTL method lies somewhere in between these two baselines, providing a principled way of controlling the level and nature of information sharing across subsets of tasks.

Baseline 1: Single task classifiers. This baseline consists of building individual classifiers separately for each task, i.e., cancer type (listed in Table 2). The data presented to each classifier consists of samples from the same cancer group, and hence is much more homogeneous than the data presented to the other two types of classifiers described below. However, as some cancer groups have low incidence, the corresponding STL classifiers are trained with relatively small training sets. The classification model is a ℓ_1-penalized (lasso) logistic regression. The choice for ℓ_1-penalization is due to its variable selection property and success in practical applications [44, 45].

To deal with censored data inputs, we adapt the lasso logistic regression formulation using the method proposed in Vock et al. [33] discussed in section *Treatment of censored data*. The result is a weighted version of the traditional ℓ_1-penalized logistic regression. The adapted formulation is defined as

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \omega_i \mathcal{L}(y_i, \mathbf{x}_i, \mathbf{w}) + \lambda \|\mathbf{w}\|_1, \tag{3}$$

where $\mathcal{L}(y_i, \mathbf{x}_i, \mathbf{w})$ is the cross-entropy loss function, $\boldsymbol{\omega}$ the IPCW weights, $y_i \in \{0, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^d$ are the label and features of the i-th data instance, and $\lambda \geq 0$ is a hyper-parameter that controls the amount of regularization. The cross-entropy loss function is defined as:

$$\mathcal{L}(y_i, \mathbf{x}_i, \mathbf{w}) = -(y_i \log(\sigma(\mathbf{w}^\intercal \mathbf{x}_i)) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^\intercal \mathbf{x}_i)))$$

where $\sigma(\cdot)$ is a sigmoid function. Note that in Eq 3, λ must be specified by the user or selected via cross-validation. The regularization aims to attenuate the overfitting that is likely to occur, particularly for tasks with small sample sizes. Our implementation of the IPCW ℓ_1-penalized logistic regression is based on the Scikit-learn package [46].

Baseline 2: Pooled classifier. The pooled baseline consists of a single classifier for the entire cohort. In this case, data from all tasks are pooled into one monolithic task and a single classifier is trained over all cancer types. The feature used to define the tasks in the MTL model (described in section *Classifiers*) and to define individual classifiers in Baseline 1 (described in section *Classifiers*), is passed to this classifier as an additional predictor feature. Similar to the model in Baseline 1, our implementation of the IPCW ℓ_1-penalized logistic regression uses Scikit-learn [46]. The advantages of this pooled classifier baseline are that the training set is much larger (albeit more heterogeneous compared to Baseline 1), and smaller model complexity than MTL, which implies a lower risk of model overfitting with smaller datasets. This pooled classifier makes the strong assumption that all tasks have a high level of similarity, ignoring particularities of individual cancer groups.

Multitask learning classifier. For this study, we extended the MTL formulation proposed in Goncalves et al. [5] called Multi-task Sparse Structure Learning (MSSL) to deal with right-censored data using the IPCW adaptation. The MSSL formulation has shown promising results on classification problems from a variety of domains. Aside from learning task coefficients, MSSL also estimates the relationship among the tasks represented by an undirected

graph, which can be further analyzed. The adapted IPCW-MSSL formulation is defined as

$$\mathbf{W} = \underset{\mathbf{W},\boldsymbol{\Omega}\succeq 0}{\arg\min} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \omega_i^t \mathcal{L}(y_i^t, \mathbf{x}_i^t, \mathbf{w}_t) + \lambda_1 \mathrm{tr}(\mathbf{W}\boldsymbol{\Omega}\mathbf{W}^T) - d \log|\boldsymbol{\Omega}| + \lambda_2 \|\boldsymbol{\Omega}\|_1, \quad (4)$$

where $\mathcal{L}(\mathbf{y}_t, \mathbf{X}_t, \mathbf{w}_t)$ is the cross-entropy loss function on the t-th task, and d is the problem dimension. The matrix $\boldsymbol{\Omega}$ is an inverse covariance (precision) matrix that captures the dependence among tasks and is learned together with the task-specific parameters $\mathbf{W}$. Finally, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are hyper-parameters that control the trade-off between the corresponding terms and need to be specified by the user. A detailed discussion on the role of each term is provided in Goncalves et al. [5]. For our experiments, we adapted the python code made publicly available by the authors.

Experimental setup

For each experiment, we randomly selected 70% of the available data for training and the remaining 30% for testing. Each experiment was repeated 30 times with different random train/test partitions to account for the variability of training and test splits. In every repetition, the three methods received exactly the same training and test sets. The hyper-parameters of the methods were selected by cross-validation. The hyper-parameters values resulting in the smallest average performance metric (Brier score) over all tasks were selected.

To assess the performance of the methods, the Brier score [47] was used. In the MTL setting, there are two complementary approaches for evaluating predictive performance: 1) 'per-sample basis', where for each experiment repetition, test sets from all tasks are pooled together and a score is computed; and 2) 'per-task basis', where the performance metric is computed for each task individually, such that each task contributes equally regardless of sample size.

To determine whether the improvement in performance obtained by MTL was practically significant, two measures of effect size were used: *Cohen's d* [48] and *common language effect size (CLES)* [49]. The notion of effect size is typically associated with randomized experiments with both control and treatment group. In our context, the population of Brier scores from randomly sampled test/train splits from the various algorithms constitute our control and treatment groups. Results from STL and Pooled models will be considered control groups while the MTL results will be the treatment group. Thus larger effect sizes are indicative of more significant differences in model performance. Each Brier score in the control group is paired with one of the treatment group by nature of the fact that they were trained on the same test/train split. Let b_i^c and b_i^t be the Brier scores from the i-th test/train split in the control group and the treatment group respectively. Let $\Delta_i = b_i^c - b_i^t$ be the difference of the ith Brier scores. Then Cohen's d for paired samples is the sample mean of Δ_i divided by the sample standard deviation of the Δ_i:

$$d = \frac{\bar{\Delta}_i}{SD(\Delta_i)}.$$

As an empirical measure of the signal-to-noise ratio, larger values are indicative of a larger effect of the treatment (MTL). In contrast to Cohen's d, CLES is a probabilistic measure of effect size. Specifically, CLES is defined as the probability that two individuals randomly chosen from each population will differ in a particular way. In our case, we would like to know if the mean Brier score is significantly less in the treatment group than the control group. Thus in our case CLES is the probability that a randomly chosen Brier score from the treatment group is less than a randomly chosen Brier score from the control group. Because our samples

are paired we simply compute the fraction of paired samples in which the treatment group's Brier score is lower than the control group's:

$$CLES = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{b_i^c > b_i^t\}}.$$

All measures of effect size can be found in Appendix B of S1 Appendix.

Results

The sample sizes for the different tasks vary widely, particularly for the most fine-grained task definition (*Topography group 1*). In this case, the total number of samples per task is as low as 244 for *other/unsp. parts of tongue* cancer (ICD code C02), to higher than 9,729 for *cervical* cancer (ICD code 53) (Table 1). Likewise, 5-year survival also varies significantly across tasks, with rates as low as 42.2% for *oropharyngeal* cancer (ICD code 10) and as high as 72% for *tonsil* cancer (ICD code C09).

We start by comparing MSSL, STL, and pooled classifiers on a per-sample basis for the three tasks split definitions. Brier score is used as the performance metric: lower Brier score indicates that the method has a better prediction performance. For each method we combined the true and predicted values from all tasks, and computed a single Brier score by combining all test observations, regardless of task assignment. The per-sample basis comparison was performed for each task definition. As shown in Fig 1, the MSSL classifier consistently outperformed both STL and pooled classifiers, across all task definitions. Specifically, we grouped the ICD-O-3 codes into two broader anatomical classes, listed in Table 2, and performed the same type of comparisons above. The groups were designed in increasing level of specificity within a hierarchy: *Topography group 1* (first two digits of ICD-O-3 anatomical codes) with 10 tasks, *Topography group 2* with 6 tasks, and finally *Topography group 3* with only 2 tasks. All models show a decreasing trend in the Brier score as we go from *Topography group 1* to *Topography*

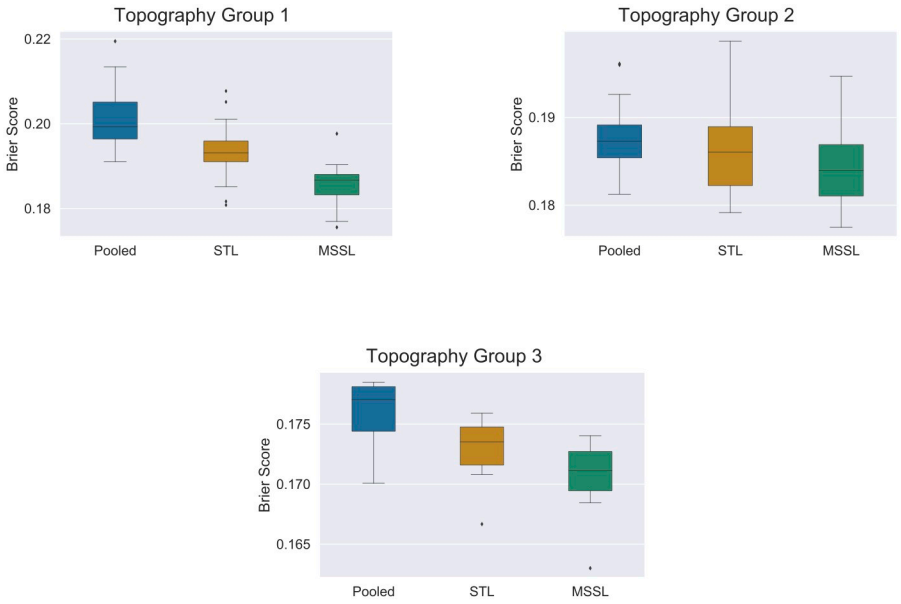

Fig 1. Brier score (y-axis) performance by classifiers (pooled, STL, and MSSL). Results show the aggregated performance from all tasks. Boxplots are composed of the mean Brier score over all 30 independent runs. MSSL shows superior performance in all three tasks splitting approaches.

https://doi.org/10.1371/journal.pone.0241225.g001

group 3. The best performance for all methods was obtained in the coarsest task split (*Topography group 3*). This is intuitively pleasing because in the coarsest task split we have more data per task, and all models benefit from this. Nevertheless, across all comparisons, the MSSL classifier outperformed both STL and pooled classifiers for the large majority of cases.

Figs 2, 3 and 4 present per-task performance for the three topography groups. A prevalent pattern across all three topography groups can be observed (particularly for TP2 and TP3): MSSL obtained the best performance (lowest Brier score), followed by STL and then the pooled classifier. For *other/unspecified parts of tongue*, the pooled method presented better performance than MSSL and STL. It indicates that information from other related cancers is helpful for predicting survival for patients diagnosed with *other/unspecified parts of tongue* cancer. Further investigation is required to properly determine the reason for MSSL's poor performance in *other/unspecified parts of tongue*, in which a significant difference in Brier score was observed. On the other hand, for *oropharynx* and *vagina* (Fig 2), pooled performed much worse than its counterparts, indicating that these two groups do not share much with the other cancer types.

When comparing the three different strategies for task definition based on topography groups (Table 2) in Fig 1, we observe that the definition based on *Topography group 3* shows the best results (lowest Brier score) for all methods. However, even in the coarsest scenario, MSSL produced better results than STL and pooled methods, indicating that intelligent information sharing helps improve 5-year survival predictability at any level of cancer grouping.

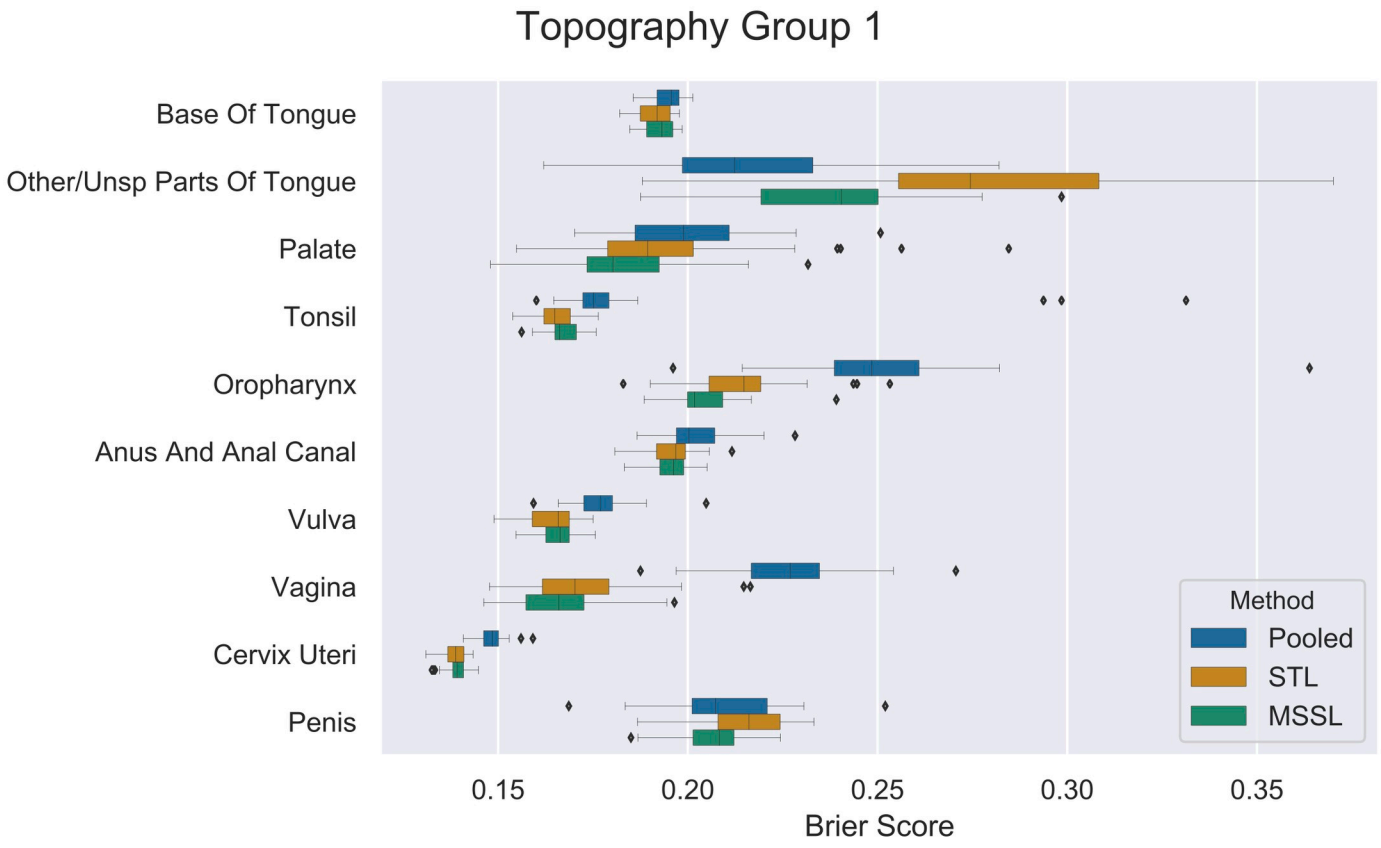

Fig 2. Brier scores (x-axis) for pooled, STL, and MSSL classifiers for *Topography group 1* data split. Boxplots show results from 30 independent runs.

https://doi.org/10.1371/journal.pone.0241225.g002

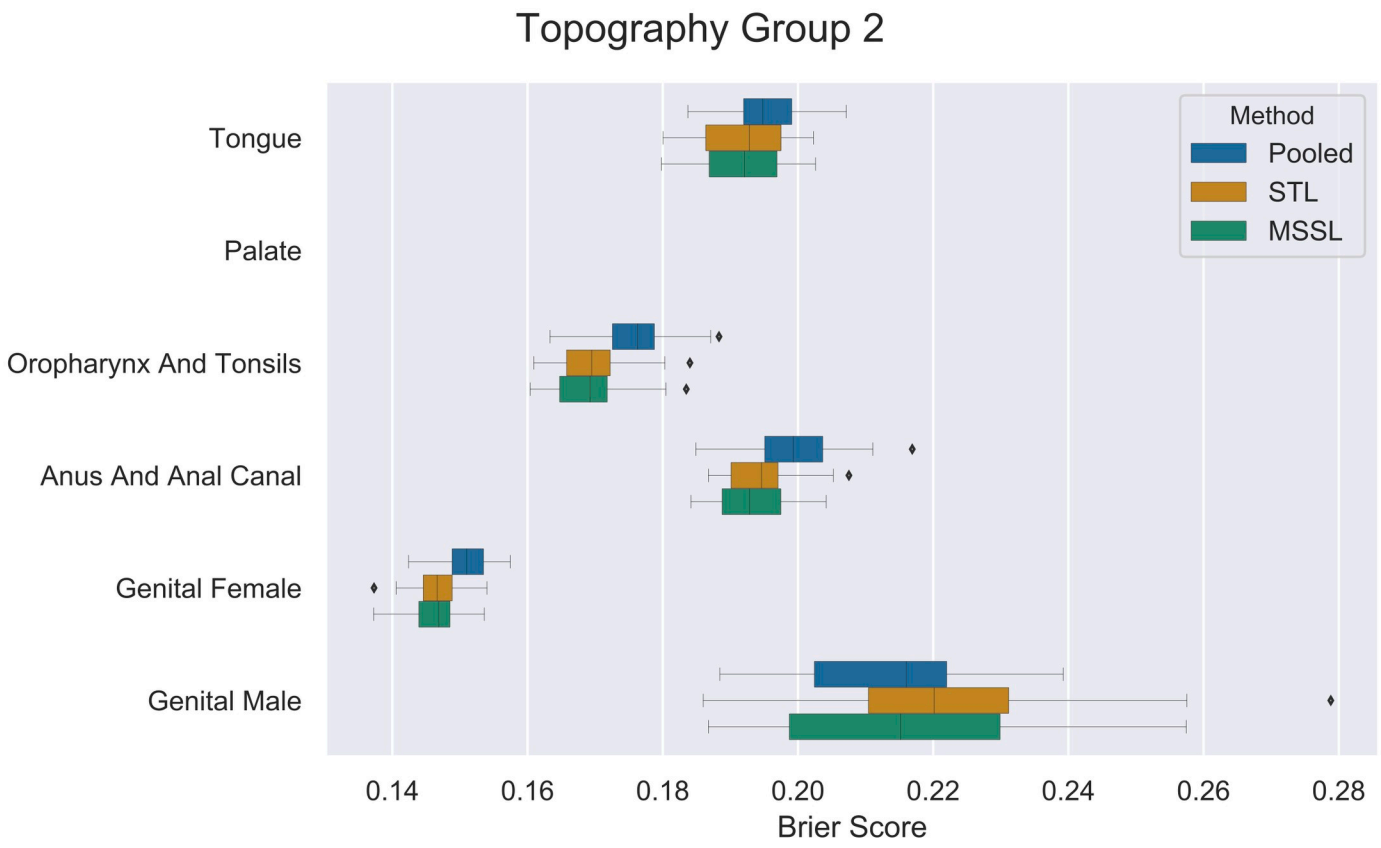

Fig 3. Brier scores (x-axis) for pooled, STL, and MSSL, and pooled classifiers for *Topography group 2* data split. Boxplots show results from 30 independent runs.

https://doi.org/10.1371/journal.pone.0241225.g003

Relative variable importance

Fig 5 shows the relative importance of the variables for STL, MSSL, and pooled models considering the task division by *Topography group 1*. The values showed are the average variable relevance computed over 30 independent model runs. To conserve space, the figures for *Topography group 2* and *Topography group 3* are provided in the Appendix A of S1 Appendix. The relative importance of the i-th variable (r_i) is computed as:

$$r_i = \frac{|w_i| - \min(|\mathbf{w}|)}{\max(|\mathbf{w}|) - \min(|\mathbf{w}|)} \tag{5}$$

where $\mathbf{w}$ is the array of coefficients estimated by the model for all variables, and $|w_i|$ denotes the absolute value of the model coefficient associated with the i-th variable in the model. For the variables encoded with the one-hot-encoding strategy, $|w_i|$ is calculated as the Euclidean norm of the coefficient vector associated with their binary variables. Note that relative importance is 0 for variables not relevant for predicting 5-year survival, and is close to 1 for variables which are highly predictive of the outcome. For MSSL and STL, this metric is computed independently for each task, implying that $r_i = 1$ in two different tasks does not imply identical importance in each task. The value is relative to the task. Note that we have ignored the sign of the variable while computing importance, as the large majority of the variables used in the model are categorical and not ordinal.

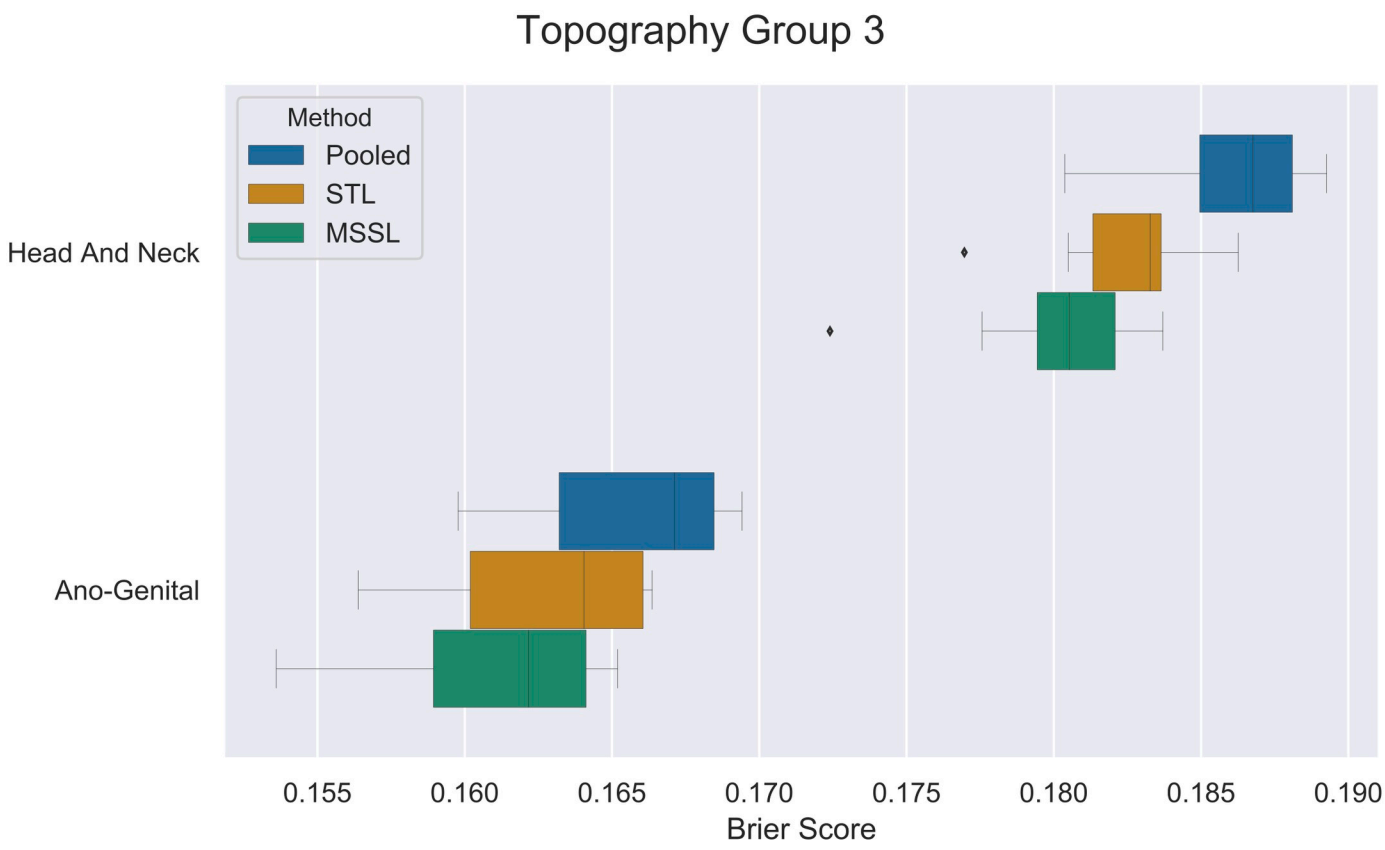

Fig 4. Brier scores (x-axis) for pooled, STL, and MSSL classifiers for *Topography group 3* data split. Boxplots show results from 30 independent runs.

https://doi.org/10.1371/journal.pone.0241225.g004

Fig 5 shows the heat map of the relative variable importance for *Topography group 1*. We observe that the age at the time of diagnosis (AGE_DX) is the most relevant variable for Pooled and MSSL models, followed by DAJCCT, DAJCCN, and DAJCCM. These are stage-related variables that indicate the level of severity of the cancer, which clearly reflects in the 5-year survival prediction. We also noticed that for MSSL, registry ID (REG) is relevant for some tasks, indicating that there might be differences across the registries for those anatomical sites. For the STL model, the variables' importance is more dependent on the anatomical site, which can be explained by the fact that STL approach fits one logistic regression model for each task separately. Therefore, STL model is more susceptible to spurious correlations in the task's dataset.

To conserve space, relative variable importance heat maps for *Topography group 2* (S1 Fig) and *Topography group 3* (S2 Fig) are presented in Appendix A of S1 Appendix.

(a) STL relative coefficients (variable) importance.

(b) MSSL relative coefficients (variable) importance.

(c) Pooled relative coefficients (variable) importance.

Learned tasks relationship

Aside from estimating the regression coefficients ($\mathbf{w}$), MSSL also learns the precision matrix (Ω) that reveals tasks relationship. Precision matrix is the inverse of the covariance matrix, which is estimated based on the regression coefficients as we can see from Eq 4. Therefore, looking at either the precision or the covariance matrix inferred by MSSL can provide relevant

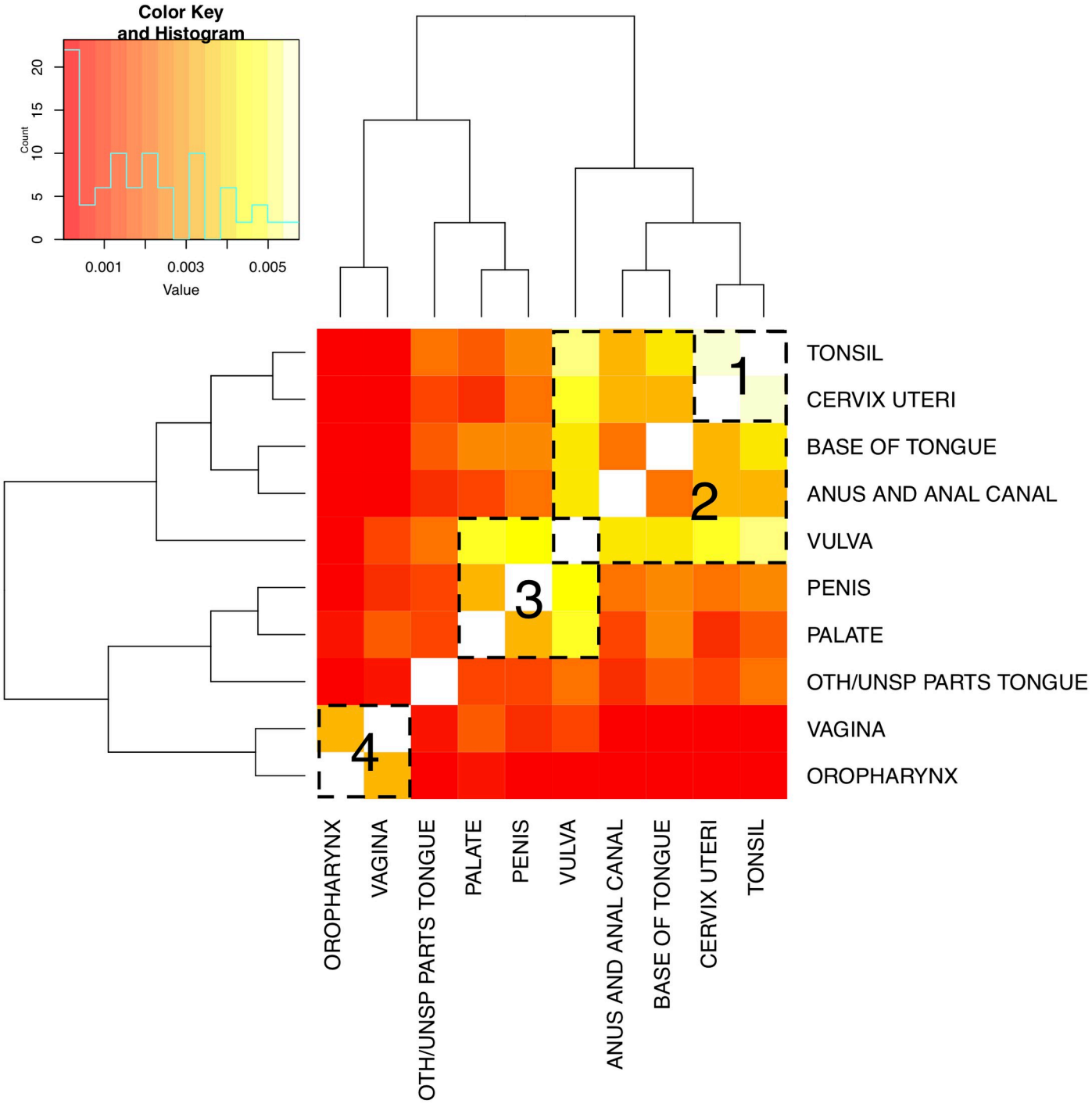

Fig 5. Relative variable importance for STL, MSSL, and Pooled models for the *Topography group 1* experiment. AGE_DX, DAJCCT, DAJCCN, and DAJCCM are the most relevant variables in the Pooled and MSSL models. For the STL model, the importance of the variables is more dependent on the anatomical site, as each model is trained separately.

https://doi.org/10.1371/journal.pone.0241225.g005

information about cancer commonalities, conditioned on the set of variables used in the model. Figs 6 and 7 present the covariance matrices for *Topography group 1* and *Topography group 2* task split. Lighter colors (yellow) are associated with higher tasks commonalities, while darker (red) means weaker relationship between the pair of tasks.

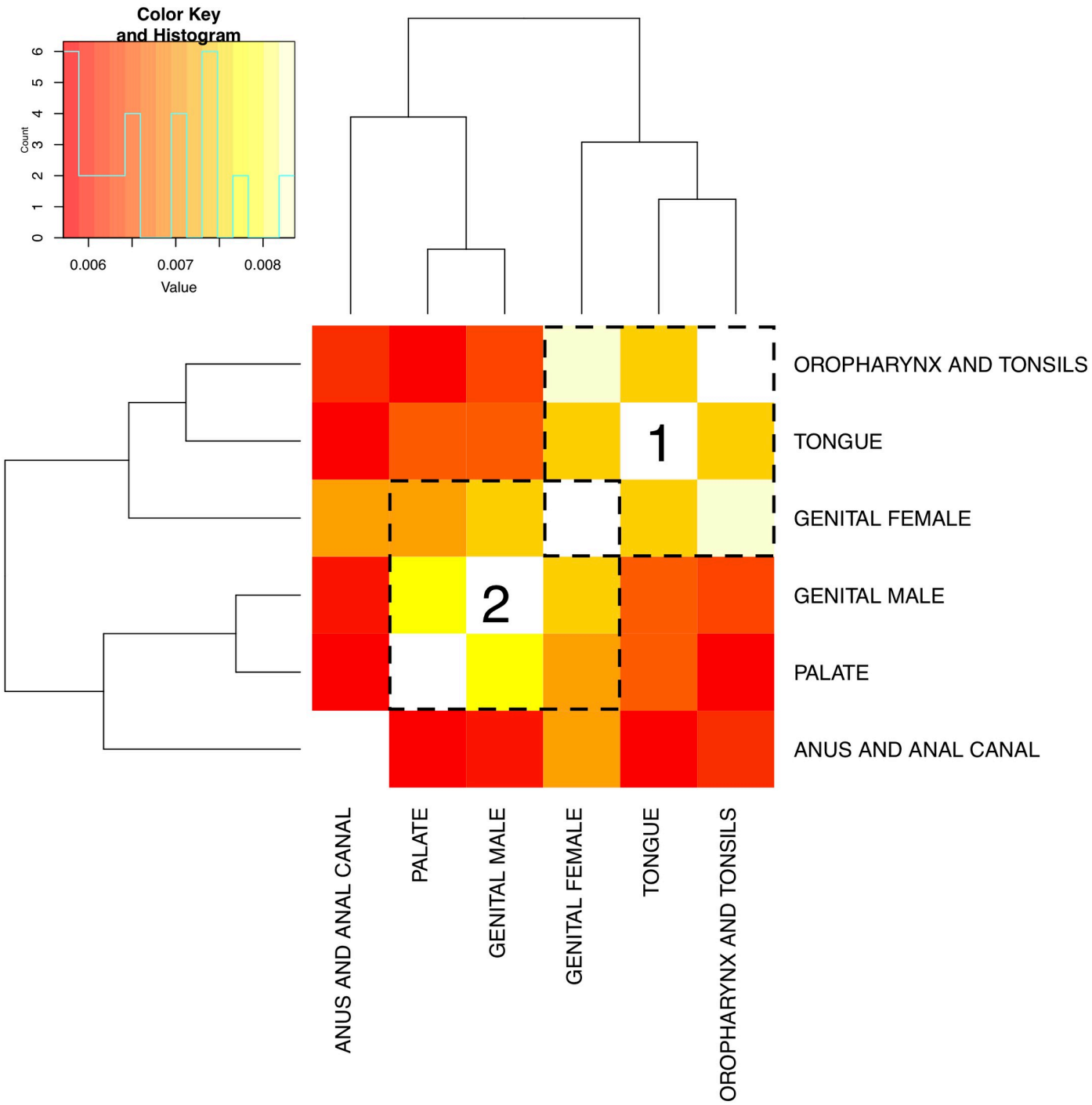

Fig 6. Task relationship learned by MSSL for task split *Topography group 1*. Lighter values indicate stronger relationships. Four groups of cancer sites were found: 1) *tonsil* and *cervix uteri*; 2) *tonsil*, *cervix uteri*, *base of tongue*, *anus and anal canal*, and *vulva*; 3) *vulva*, *penis*, and *palate*; and 4) *vagina* and *oropharynx*. Groups are highlighted in the plot.

https://doi.org/10.1371/journal.pone.0241225.g006

For the *Topography group 1*, MSSL found four groups of related tasks with distinct sizes and magnitudes: 1) *tonsil* and *cervix uteri*; 2) *tonsil*, *cervix uteri*, *base of tongue*, *anus and anal canal*, and *vulva*; 3) *vulva*, *penis*, and *palate*; and 4) *vagina* and *oropharynx*. Firstly, we notice that *tonsil* and *cervix uteri* forms a group (group 1) within a larger group of tasks (group 2).

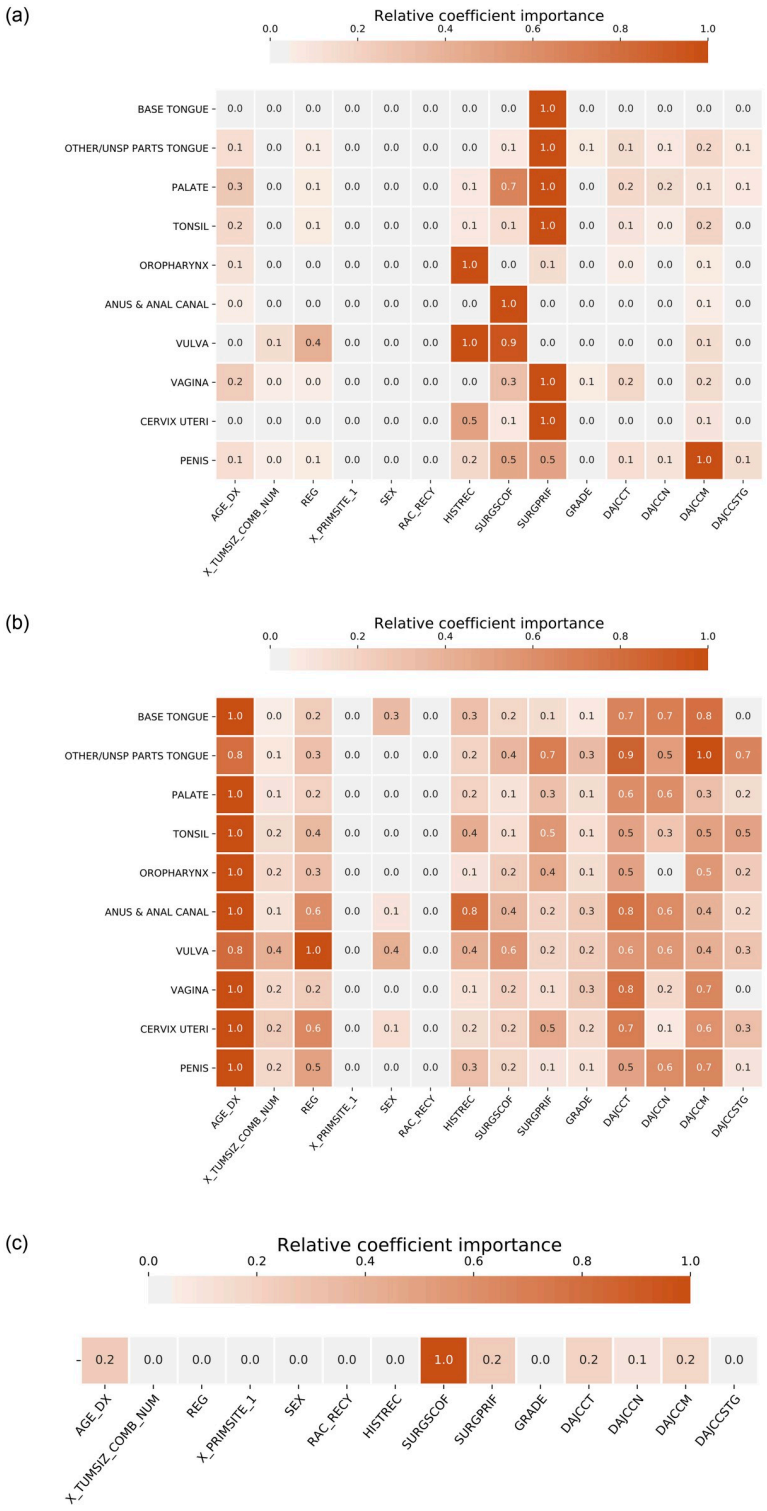

Fig 7. Task relationship learned by MSSL for task *Topography group 2*. Lighter values indicate stronger relationships. *Genital male* and *genital female* are strongly related. *Nasopharynx and palate*, *tongue*, *oropharynx and tonsils* forms a mutually related group. *Oropharynx and tonsils* is related with the majority of tasks, except for *anus and anal canal*. Groups are highlighted in the plot.

https://doi.org/10.1371/journal.pone.0241225.g007

This structure encourages MSSL to share more information between the two tasks in group 1 and with a lesser extent with tasks in the larger group 2. We observe that *vulva* is associated with two different groups of tasks (groups 2 and 3). Therefore, the part of the model associated with *vulva* cancers will share information and also be influenced by data from the two groups of cancers simultaneously. That is possible to be captured in MSSL due to the pairwise nature of the precision matrix learned by the model.

The four groups in Fig 6 appear to be closely related to the 5-year survival rate. For group 1, the survival rate is the highest varying from 68.7 to 72.0%. Group 4 has the lowest survival and varies between 41.9% and 42.4%, while groups 2 and 3 have 5-year survival varying from 62.6%-72.0% and 56.2%-63.5%, respectively. Since we are modeling the 5-year survival prediction, it makes sense that the coefficients of the tasks with similar survival time (tasks in the same group) are more similar.

In the case of *Topography group 2*, Fig 7, *oropharynx and tonsils*, *tongue*, and *genital female* forms a group of mutually related tasks (group 1). A second group of tasks is formed with *genital female*, *genital male*, and *palate* (group 2). Similar to what is observed in Fig 6 with *vulva*, *genital female* appears in two different groups simultaneously. *Anus and anal canal* that appeared in the larger group 2 for *Topography group 1* in Fig 6, now is less related to other cancer groups. This indicates that task definition has a significant effect on how data is shared within the MTL model.

As observed for *Topography group 1*, the two groups are closely related to the 5-year survival rates. The rate for group 1 is the highest, varying from 62.5% to 66.8%, while group 2 has lower survival varying from 56.0% to 58.2%.

Sample size vs MTL performance improvement

In this study, we have five cancer sites with relative few cases: *other/unsp tongue* (C02), *palate* (C05), *oropharynx* (C10), *vagina* (C52) and *penis* (C60) which all have less then 1000 cases. Five others have relative more cases: *base of tongue* (C01), *tonsil* (C09), *anus & anal canal* (C21), *vulva* (C51), and *cervix uteri* (C53) which have between 2,733 and 9,729 cases, see Table 1. For all cancer sites with relative many cases, the Brier scores are rather similar. While for cancer diagnosis with few cases, MSSL performs clearly better, except for C02 where the pooled model is best, and the STL is doing the worst (see Fig 2).

Fig 8 shows the relationship between the number of training samples in the tasks split by *Topography group 1* and the relative performance improvement (RPI) of MSSL over STL. The performance improvement for task k ($\mathrm{RPI_k}$) is computed as the relative gain in performance of MSSL over STL in terms of the Brier Score:

$$RPI_{\mathrm{k}} = \frac{BS(STL_k) - BS(MTL_k)}{BS(STL_k)} \tag{6}$$

where BS(·) is the Brier score performance on the test set obtained by the method. We clearly observe that small-sample-size tasks benefit the most from the MSSL model. And, as the sample size increase, the difference in performance between MSSL and STL reduces.

Discussion

In this paper we demonstrated that multitask classifiers achieved significant improvements in predicting survival for the majority of scenarios investigated, when compared to baseline approaches such as single task and pooled classifiers. To the best of our knowledge this is the first demonstration that sharing information across anatomically distinct cancer types can lead to improved predictive survival models.

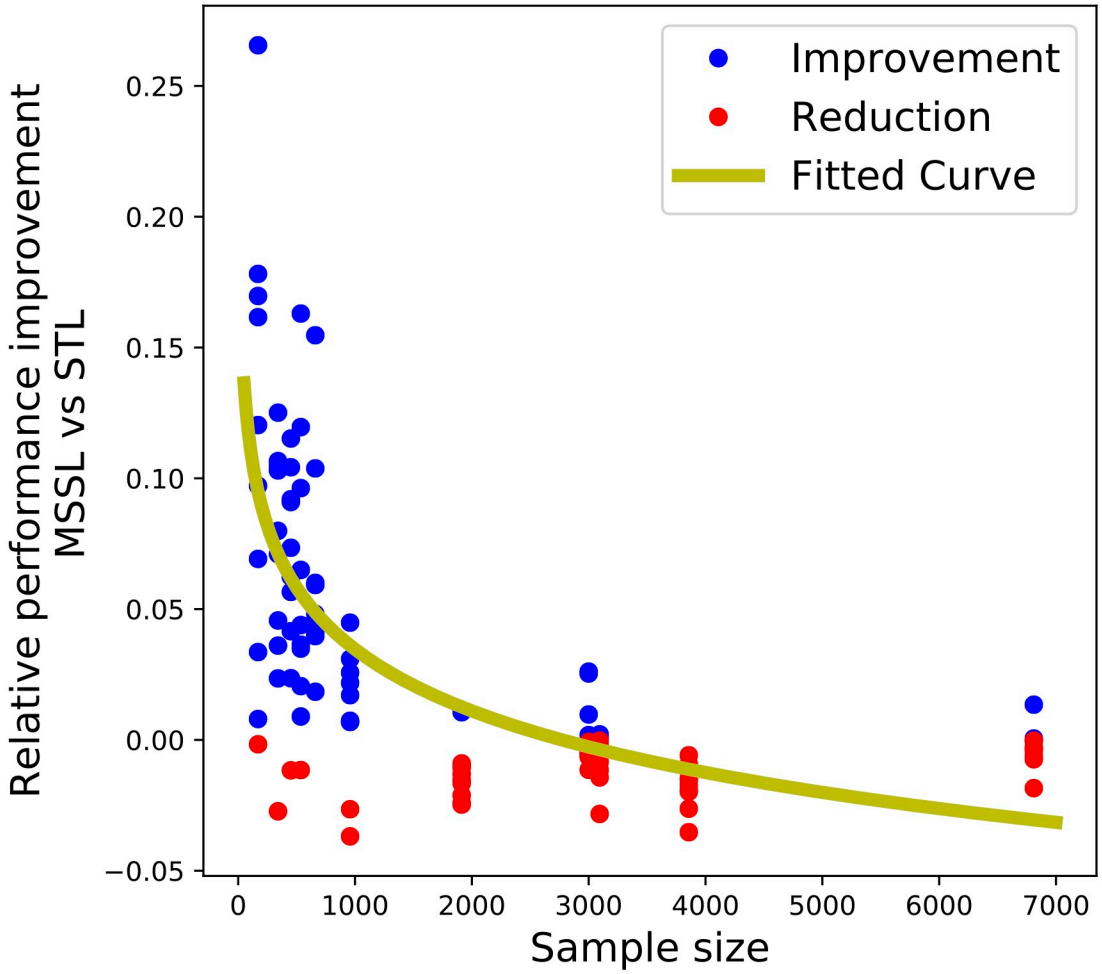

Fig 8. Correlation between task training sample size and MSSL relative performance improvement over STL. Tasks split by *Topography group 1*. MSSL shows higher improvement over STL particularly for tasks with smaller sample sizes.

https://doi.org/10.1371/journal.pone.0241225.g008

The present study demonstrates the benefits of leveraging a multitask learning approach to combine clinical data from disparate cancer types in order to improve prediction of cancer patient survival. Previous work has applied MTL approaches to cancer data; however, for the most part, they focused on very specific and homogeneous cancer types, with MTL being deployed for related datasets of the same cancer type. The one exception is the study by [50] in which transfer learning (a machine learning technique related to MTL) was applied across breast and ovarian cancer DNA copy number datasets. While molecular data such as DNA copy number and genetic markers carry relatively high predictive power (in comparison to the type of clinical data used here), they are not ubiquitously deployed in clinical practice for a number of reasons, such as cost or lack of studies demonstrating their translation to courses of action. In contrast, clinical surveillance data, such as the one used in this study and made available by SEER, is ubiquitous in clinical practice. This means that from the machine learning point of view, larger datasets are available; but more importantly, methods trained on this type of data have greater likelihood of actually being deployed to clinical practice to provide decision support to physicians.

Survival and time-to-event data are susceptible to incomplete observations or censoring. While predicting patient survival times based on patient features is of interest, it is important

to be able to adapt the learning algorithms to handle censored cases to avoid potential bias [32]. We thus extended IPCW to be used in MTL.

For almost every cancer site selected in this study a common etiological factor linked to infection with human papillomavirus has been reported [51]. Worldwide these virus related cancer sites show an increasing trend, which has been linked to increased exposure to sexually transmitted HPV in the population due to changes in sexual behaviour [52]. Furthermore, some studies indicate a favorable survival pattern in HPV-positive tumors of anal [53], oropharyngeal [54], vaginal [55], and penile cancers [56]. This suggests that these seemingly disparate cancer sites share commonalities which can be leveraged to improve the accuracy of predictive algorithms. To evaluate our hypothesis, we compared the accuracy of a MTL classifier against two baselines. The first baseline consisted of several classifiers (STL classifiers), one for each cancer group. The second baseline consisted of a single classifier applied to the entire cohort (pooled classifier), where the cancer group information used to define the MTL tasks and the different STL classifiers was incorporated as an extra predictive feature. These two baselines represent the approaches most commonly used in research and clinical practice, and both make incorrect assumptions about the data.

In the first baseline (STL), different cancers types were treated individually as if they were independent from each other. The accuracy of the predictions, however, depends heavily on the size of the population, and the independence assumption becomes increasingly more consequential with a decreasing number of patients. Generally, the small sample sizes can lead to severe over-fitting, even for relatively simple models. The one-size-fits-all pooled baseline makes the strong assumption that data from all cancers groups are identically distributed. In other words, this approach completely ignores any differences between individual cancer groups. These pooled models tend to approximate the mean distribution, which will be heavily impacted by the categories with the most samples. MTL approaches make more nuanced assumptions about the relationships between cancer types, allowing them to be treated as non-identically distributed, but also not as entirely independent of one another. In this way certain parameters' sub-spaces can be shared across all or a subset of cancer groups, whereas others are specific to individual cancer types.

Overall, MTL methods tend to outperform STL methods when there is latent information shared across tasks. Task sample size also plays an important role in model performance. For tasks with large sample sizes, the improvements of MTL over STL can be limited, whereas substantial improvements can be seen for tasks with a smaller number of observations. This, of course, is all dependent on the assumption that there is in fact similarities across the different tasks. If tasks are truly independent, ideally the MTL approach will perform similarly to its STL counterpart.

When tasks are defined at *Topography Group 2 and 3* (see Table 2), MTL improved predictive performance over STL and pooled for the large majority of tasks, meaning that MTL was able to exploit the commonality existing in the tasks (which we speculate is probably due to HPV). While the MTL improvements were seen for all three anatomical cancer groups, significant improvements were not seen in all individual tasks. In our view, these results provide an optimistic perspective on our proposed approach of combining data from disparate cancer types, but there remains many possible directions for improvement.

Analogously to the standard STL approach, the proposed MTL model estimates a set of coefficients (one for each data attribute) for each task, that is, cancer site. The only and crucial distinction is that, during training, the MTL model encourages coefficients of different but related tasks to be similar. However, the model still permits attributes to have very different coefficients for tasks that are not related. Recall that MSSL learns a matrix Ω to capture tasks relationship from the data. For example, the importance of "Tumor size" or "Stage" for *base of tongue* could be different to *palate*, if the data say so. The key point of MSSL is that even the

model encouraging coefficients of tasks to be close to each other in the parameter space, it is still flexible enough to accommodate any possibly unrelated tasks or particular attributes. Thus, even if "Stage" does not have exactly the same meaning for one particular cancer type compared to the others, including this feature into the MSSL model can still contribute to the improvement of survival prediction.

Multitask learning methods tend to outperform single task learning approaches in the low sample size regime. In such regimes, the implicit information transfer procedure in MSSL diminish the impact of data scarcity, while in large sample sizes, STL models have already enough data to construct an accurate model. In our experiment, we see that the MSSL and STL models have similar performance with regard to Brier score when dealing with larger sample sizes. For cancer sites with less than 1,000 cases, we see marked improvement when using the MSSL methodology.

We attribute the better performance of both MTL and STL methods over pooled methods to the heterogeneity of cancer cases when pooling all data into a single classifier. Even though HPV is a common trigger of the cancer types considered in our cohort, the dissimilarities among the cancer types also play an important role when determining the classifier to use. Thus, the MTL model appears as a suitable candidate to find the correct balance between the one-size-fits-all and completely independent models.

While our working example focuses on HPV-related cancers, there exist numerous other examples in the literature of shared commonalities across distinct cancer types. In this work we explicitly formulated the problem to deal with cancers that share a latent potential causal factor. Many other such examples could similarly be formulated. For example, mutations in the oncogenic signaling protein Ras are found in upwards of 30% of all human cancers [57]. Alternatively we may apply these methods to larger sets of data where we do not explicitly filter the data on known latent causes, but instead extend the MTL framework to help discover the latent connections between cancers automatically. Another possible extension is to use observed survival times instead of binary survival outcomes.

One possible direction is to augment the patients' information used as predictive features by the machine learning model. Prescribed medications, type and length of treatments, medical notes (unstructured text), medical test results, including images and physiological measurements, are additional sources of relevant information for predicting patient survival. Dealing with unstructured data, e.g., images and text, poses additional challenges, as computational representations of such data need to be extracted. Fortunately, image and text data processing have seen significant advancements in the last decade, particularly due to the development of deep learning models [58]. Going forward, we will investigate how much these additional sources of data can improve the performance of the machine learning models for survival prediction.

In conclusion, we have proposed a new approach for predicting 5-year cancer survival in which data from anatomically distinct cancers can be combined via multitask learning to improve overall prediction accuracy. While this work represents a proof-of-concept demonstration, and extrapolation to larger and broader cohorts remains to be demonstrated, there are a number of potential research directions that could amplify the improvements in performance obtained by MTL. If indeed this type of improvement can be shown to extrapolate across different cohorts, and the improvements can be increased, it has the potential for real-world impact in clinical research and practice.

Supporting information

S1 File.
(DOCX)

S1 Appendix.
(PDF)

Author Contributions

Conceptualization: Braden Soper, Mari Nygård, Jan F. Nygård, Priyadip Ray, David Widemann, Ana Paula Sales.

Data curation: Andre Goncalves, Braden Soper, Jan F. Nygård, Ana Paula Sales.

Funding acquisition: Ana Paula Sales.

Investigation: Mari Nygård, David Widemann.

Methodology: Andre Goncalves, Jan F. Nygård, David Widemann, Ana Paula Sales.

Project administration: Ana Paula Sales.

Resources: Andre Goncalves.

Software: Andre Goncalves.

Supervision: Mari Nygård, Jan F. Nygård, Priyadip Ray, Ana Paula Sales.

Validation: Andre Goncalves, Braden Soper, David Widemann.

Visualization: Andre Goncalves.

Writing – original draft: Andre Goncalves, Braden Soper, Mari Nygård, Jan F. Nygård, Priyadip Ray, Ana Paula Sales.

Writing – review & editing: Andre Goncalves, Braden Soper, Mari Nygård, Jan F. Nygård, Priyadip Ray, Ana Paula Sales.

References

1. Caruana R. Multitask Learning. Machine Learning. 1997; 28(1):41–75. https://doi.org/10.1023/A:1007379606734
2. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
3. Kim S, Xing EP. Tree-guided group lasso for multi-task regression with structured sparsity. In: International Conference on International Conference on Machine Learning; 2010. p. 543–550.
4. Yang M, Li Y, Zhang Z. Multi-task learning with Gaussian matrix generalized inverse Gaussian model. In: International Conference on Machine Learning; 2013. p. 423–431.
5. Goncalves AR, J VF, Banerjee A. Multi-task Sparse Structure Learning with Gaussian Copula Models. Journal of Machine Learning Research. 2016; 17(33):1–30.
6. Goncalves A, Ray P, Soper B, Widemann D, Nygård M, Nygård JF, et al. Bayesian multitask learning regression for heterogeneous patient cohorts. Journal of Biomedical Informatics: X. 2019; 4.
7. Kandaswamy C, Silva LM, Alexandre LA, Santos JM. High-Content Analysis of Breast Cancer Using Single-Cell Deep Transfer Learning. Journal of Biomolecular Screening. 2016; 21(3):252–259. https://doi.org/10.1177/1087057115623451
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639):115–118. https://doi.org/10.1038/nature21056 PMID: 28117445
9. Chen AH, Huang ZW. A New Multi-Task Learning Technique to Predict Classification of Leukemia and Prostate Cancer. In: Zhang D, Sonka M, editors. Medical Biometrics; 2010. p. 11–20.
10. Parkin DM. The global health burden of infection-associated cancers in the year 2002. International Journal of Cancer. 2006; 118(12):3030–3044. https://doi.org/10.1002/ijc.21731 PMID: 16404738

11. Alemany L, Saunier M, Alvarado-Cabrero I, et al. Human papillomavirus DNA prevalence and type distribution in anal carcinomas worldwide. Int J Cancer. 2015; 136(1):98–107. https://doi.org/10.1002/ijc.28963 PMID: 24817381

12. Alemany L, et al. Large contribution of human papillomavirus in vaginal neoplastic lesions: A worldwide study in 597 samples. European Journal of Cancer. 2014; 50(16):2846–2854. https://doi.org/10.1016/j.ejca.2014.07.018 PMID: 25155250

13. Alemany L, et al. Role of Human Papillomavirus in Penile Carcinomas Worldwide. European Urology. 2016; 69(5):953–961. https://doi.org/10.1016/j.eururo.2015.12.007 PMID: 26762611

14. de Sanjosé S, et al. Worldwide human papillomavirus genotype attribution in over 2000 cases of intraepithelial and invasive lesions of the vulva. European Journal of Cancer. 2013; 49(16):3450–3461. https://doi.org/10.1016/j.ejca.2013.06.033 PMID: 23886586

15. Marur S, D'Souza G, Westra WH, Forastiere AA. HPV-associated head and neck cancer: a virus-related cancer epidemic. The Lancet Oncology. 2010; 11(8):781–789. https://doi.org/10.1016/S1470-2045(10)70017-6 PMID: 20451455

16. Nasman A, Attner P, Hammarstedt L, et al. Incidence of human papillomavirus (HPV) positive tonsillar carcinoma in Stockholm, Sweden: an epidemic of viral-induced carcinoma? Int J Cancer. 2009; 125:362–368. https://doi.org/10.1002/ijc.24339 PMID: 19330833

17. Allo G, Yap ML, Cuartero J, Milosevic M, Ferguson S, Mackay H, et al. HPV-independent Vulvar Squamous Cell Carcinoma is Associated With Significantly Worse Prognosis Compared With HPV-associated Tumors. Int J Gynecol Pathol. 2020; 39(4):391–399. https://doi.org/10.1097/PGP.0000000000000620 PMID: 31274700

18. Ferrandiz-Pulido C, Masferrer E, de Torres I, Lloveras B, Hernandez-Losa J, Mojal S, et al. Identification and genotyping of human papillomavirus in a Spanish cohort of penile squamous cell carcinomas: correlation with pathologic subtypes, p16(INK4a) expression, and prognosis. J Am Acad Dermatol. 2013; 68(1):73–82. https://doi.org/10.1016/j.jaad.2012.05.029 PMID: 22863066

19. Wookey VB, Appiah AK, Kallam A, Ernani V, Smith LM, Ganti AK. HPV Status and Survival in Non-Oropharyngeal Squamous Cell Carcinoma of the Head and Neck. Anticancer Res. 2019; 39(4):1907–1914. https://doi.org/10.21873/anticanres.13299 PMID: 30952732

20. Urbute A, Rasmussen CL, Belmonte F, Obermueller T, Prigge ES, Arbyn M, et al. Prognostic Significance of HPV DNA and p16(INK4a) in Anal Cancer: A Systematic Review and Meta-Analysis. Cancer Epidemiol Biomarkers Prev. 2020; 29(4):703–710. https://doi.org/10.1158/1055-9965.EPI-19-1259 PMID: 32051192

21. Zhang J, Zhang Y, Zhang Z. Prevalence of human papillomavirus and its prognostic value in vulvar cancer: A systematic review and meta-analysis. PLoS One. 2018; 13(9):e0204162. https://doi.org/10.1371/journal.pone.0204162 PMID: 30256833

22. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010; 363(1):24–35. https://doi.org/10.1056/NEJMoa0912217 PMID: 20530316

23. Lassen P, Eriksen JG, Krogdahl A, Therkildsen MH, Ulhoi BP, Overgaard M, et al. The influence of HPV-associated p16-expression on accelerated fractionated radiotherapy in head and neck cancer: evaluation of the randomised DAHANCA 6&7 trial. Radiother Oncol. 2011; 100(1):49–55. https://doi.org/10.1016/j.radonc.2011.02.010 PMID: 21429609

24. Posner MR, Lorch JH, Goloubeva O, Tan M, Schumaker LM, Sarlis NJ, et al. Survival and human papillomavirus in oropharynx cancer in TAX 324: a subset analysis from an international phase III trial. Ann Oncol. 2011; 22(5):1071–7. https://doi.org/10.1093/annonc/mdr006 PMID: 21317223

25. Wright JL, Dalkin BL, True LD, Ellis WJ, Stanford JL, Lange PH, et al. Positive Surgical Margins at Radical Prostatectomy Predict Prostate Cancer Specific Mortality. The Journal of Urology. 2010; 183(6):2213–2218. https://doi.org/10.1016/j.juro.2010.02.017 PMID: 20399459

26. Weiser MR, Gönen M, Chou JF, Kattan MW, Schrag D. Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. Journal of Clinical Oncology. 2011; 29(36):4796. https://doi.org/10.1200/JCO.2011.36.5080 PMID: 22084366

27. Luxembourg A, Kjaer SK, Nygard M, Ellison MC, Group T, Marshall JB, et al. Design of a long-term follow-up effectiveness, immunogenicity and safety study of women who received the 9-valent human papillomavirus vaccine. Contemp Clin Trials. 2017; 52:54–61. https://doi.org/10.1016/j.cct.2016.10.006 PMID: 27777126

28. Skolidis G, Sanguinetti G. Bayesian multitask classification with Gaussian process priors. IEEE Transactions on Neural Networks. 2011; 22(12). https://doi.org/10.1109/TNN.2011.2168568 PMID: 21990334

29. Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. In: Advances in Neural Information Processing Systems (NIPS); 2007. p. 41–48.

30. Ji S, Ye J. An accelerated gradient method for trace norm minimization. In: International Conference on Machine Learning. ACM; 2009. p. 457–464.

31. Jalali A, Sanghavi S, Ruan C, Ravikumar PK. A Dirty Model for Multi-task Learning. In: Advances in Neural Information Processing Systems (NIPS); 2010. p. 964–972.

32. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. ACM Computing Surveys. 2017; 1(1):38.

33. Vock DM, Wolfson J, Bandyopadhyay S, Adomavicius G, Johnson PE, Vazquez-Benitez G, et al. Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. Journal of Biomedical Informatics. 2016; 61:119–131. https://doi.org/10.1016/j.jbi.2016.03.009 PMID: 26992568

34. SEER. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2015) National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission.; 2018.

35. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, et al. International Classification of Diseases for Oncology: Third Edition. World Health Organization: Geneva; 2000.

36. Razzaghi H, Saraiya M, Thompson TD, Henley SJ, Viens L, Wilson R. Five-year relative survival for human papillomavirus-associated cancer sites. Cancer. 2018; 124(1):203–211. https://doi.org/10.1002/cncr.30947 PMID: 29105738

37. Simon J, Schroeder L, Ingarfield K, Diehl S, Werner J, Brenner N, et al. Epstein-Barr virus and human papillomavirus serum antibodies define the viral status of nasopharyngeal carcinoma in a low endemic country. Int J Cancer. 2020; 147(2):461–471. https://doi.org/10.1002/ijc.33006 PMID: 32279316

38. Wotman M, Oh EJ, Ahn S, Kraus D, Costantino P, Tham T. HPV status in patients with nasopharyngeal carcinoma in the United States: A SEER database study. Am J Otolaryngol. 2019; 40(5):705–710. https://doi.org/10.1016/j.amjoto.2019.06.007 PMID: 31277887

39. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. International Journal of Medical Informatics. 2017; 108:1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013 PMID: 29132615

40. Allemani C, et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37.513.025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. The Lancet. 2018; 391:1012–1075. https://doi.org/10.1016/S0140-6736(17)33326-3

41. Berkson J, Gage RP. Survival curve for cancer patients following treatment. Journal of the American Statistical Association. 1952; 47:501–515. https://doi.org/10.1080/01621459.1952.10501187

42. Dickman PW, Adami HO. Interpreting trends in cancer patient survival. Journal of Internal Medicine. 2006; 260:103–117. https://doi.org/10.1111/j.1365-2796.2006.01677.x PMID: 16882274

43. Tsiatis AA. Semiparametric theory and missing data. Springer Series in Statistics. Springer; 2010.

44. Lin Y, Wang S, Chappell RJ. Lasso tree for cancer staging with survival data. Biostatistics. 2012; 14(2):327–339. https://doi.org/10.1093/biostatistics/kxs044 PMID: 23221681

45. Zhang Y, Biswas S. An improved version of logistic Bayesian LASSO for detecting rare haplotype-environment interactions with application to lung cancer. Cancer informatics. 2015; 14:CIN–S17290. https://doi.org/10.4137/CIN.S17290 PMID: 25733797

46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

47. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine. 1999; 18(17-18):2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5 PMID: 10474158

48. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, N.J: L. Erlbaum Associates; 1988.

49. McGraw KO, Wong SP. A common language effect size statistic. Psychological Bulletin. 1992; 111(2):361–365. https://doi.org/10.1037/0033-2909.111.2.361

50. Zhang H, Tian Z, Kuang R. Transfer Learning across Cancers on DNA Copy Number Variation Analysis. In: IEEE International Conference on Data Mining; 2013. p. 1283–1288.

51. Hansen BT, Campbell S, Nygard M. Long-term incidence trends of HPV-related cancers, and cases preventable by HPV vaccination: a registry-based study in Norway. BMJ Open. 2018; 8(2):e019005. https://doi.org/10.1136/bmjopen-2017-019005 PMID: 29476028

52. Liu G, Hariri S, Bradley H, Gottlieb SL, Leichliter JS, Markowitz LE. Trends and patterns of sexual behaviors among adolescents and adults aged 14 to 59 years, United States. Sexually transmitted diseases. 2015; 42(1):20. https://doi.org/10.1097/OLQ.0000000000000231 PMID: 25504296

53. Serup-Hansen E, Linnemann D, Skovrider-Ruminski W, Hogdall E, Geertsen PF, Havsteen H. Human papillomavirus genotyping and p16 expression as prognostic factors for patients with American Joint Committee on Cancer stages I to III carcinoma of the anal canal. J Clin Oncol. 2014; 32(17):1812–7. https://doi.org/10.1200/JCO.2013.52.3464 PMID: 24821878

54. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010; 363(1):24–35. https://doi.org/10.1056/NEJMoa0912217 PMID: 20530316

55. Gadducci A, Fabrini MG, Lanfredini N, Sergiampietri C. Squamous cell carcinoma of the vagina: natural history, treatment modalities and prognostic factors. Crit Rev Oncol Hematol. 2015; 93(3):211–24. https://doi.org/10.1016/j.critrevonc.2014.09.002 PMID: 25476235

56. Djajadiningrat RS, Jordanova ES, Kroon BK, van Werkhoven E, de Jong J, Pronk DT, et al. Human papillomavirus prevalence in invasive penile cancer and association with clinical outcome. The Journal of urology. 2015; 193(2):526–531. https://doi.org/10.1016/j.juro.2014.08.087 PMID: 25150641

57. Fernández-Medarde A, Santos E. Ras in Cancer and Developmental Diseases. Genes & Cancer. 2011; 2(3):344–358. https://doi.org/10.1177/1947601911411084 PMID: 21779504

58. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature medicine. 2019; 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z PMID: 30617335