Building and Evaluating Cancer Prescreening Models with Electronic Health Records

by

Pasapol Saowakon

S.B., Computer Science and Engineering, Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by:	Pasapol Saowakon Department of Electrical Engineering and Computer Science May 19, 2023
Certified by:	Martin C. Rinard Professor of Computer Science and Engineering Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Building and Evaluating Cancer Prescreening Models with Electronic Health Records

by

Pasapol Saowakon

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Cancer is a leading cause of death that kills over ten million people every year, and many times delayed treatment is the culprit. Building on a recent framework, we used electronic health records from TriNetX to develop prescreening models for ten different cancer types: biliary tract, brain, breast (female), colon, esophageal, gastric, kidney, liver, lung, and ovarian. The models showed great performance, with neural network models consistently but marginally outperforming their logistic regression counterparts. As expected, we found that models trained to detect specific cancer types performed noticeably better than ones trained more generally to detect any cancer. All models proved to be reasonably robust in geographical, racial, and temporal external validations, although a prospective study is still needed to verify the performance and the potential impact of our models.

Thesis Supervisor: Martin C. Rinard
Title: Professor of Computer Science and Engineering

Acknowledgments

First, I would like to extend my sincere gratitude to Professor Martin Rinard, my research supervisor. Martin, thank you so much for trusting me to pursue this fascinating, potentially life-saving research. You have provided me with a positive, stress-free environment to work and learn in, which I appreciate immensely. It was truly an honor to work alongside you.

Next, I would like to thank my impressive colleagues Kai Jia and Limor Appelbaum for their endless and proactive support throughout my time working with them. Kai, thanks for answering my questions even early in the morning or late at night. This work would have been extremely difficult without your assistance. Limor, you have been no less dedicated and prompt than Kai when it comes to giving help. Your extensive medical insights and cheerful energy towards our work are what has allowed us to keep progressing at a fast pace. It was my pleasure to work with both of you.

I also would like to thank TriNetX and the awesome people there for providing us with arguably the most critical piece for our work: the data. Kathryn Haapala and Jeff Warnick especially, thank you for being reliable points of contact and for being there whenever I needed assistance.

Lastly, I must acknowledge the ultimate component of my success today: my family. Thank you, Mom and Dad, for helping me plan and prepare for my future proactively throughout my childhood. Thanks for supporting me in whatever endeavors I enjoy and for working hard to earn me the privilege of not having to worry about anything except myself. I would never have gotten to where I am today, where I had always wanted to be, had it not been for your persistent work and efforts. I really cannot thank you enough.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a leading cause of death that kills over ten million people every year. Many times, delayed treatment is the culprit: Hanna et al. estimated that every delayed month of cancer treatment can increase the risk of death by ten percent [5]. The development of tools that enable early detection of cancer is therefore critically important, for their potential to save countless lives.

The recent digitization of health records has brought about a new, powerful possibility for developing such tools. This advent of electronic health records (EHRs) has allowed researchers to easily conduct retrospective studies to identify and evaluate causes of cancer and develop cancer risk models accordingly. Unfortunately, these records are considered sensitive and are typically maintained as private within the institutions that own them, limiting the potential for breakthroughs and advancements in the field.

Through a collaboration agreement between MIT[1], BIDMC[2], and TriNetX—a global federated network of health records—we were able to access a vast EHR database for the development of cancer risk models. In this thesis, we iterate upon the framework outlined in a precursor work on pancreatic cancer [7] by the author's own research group to develop risk models for other types of cancer: biliary tract, brain, breast (female), colon, esophageal, gastric, kidney, liver, lung, and ovarian cancers.

[1] MIT stands for Massachusetts Institute of Technology

[2] BIDMC stands for Beth Israel Deaconess Medical Center

Chapter 2

Related work

Many EHR-based cancer risk models have emerged in recent years, with the vast majority of efforts focused on individual cancer types. For hepatocellular carcinoma, in 2021, Liang et al. developed a prediction model with EHR data from a Taiwanese national database [9]. For colorectal cancer, in 2020, Cooper et al. created a risk model with data from a UK network of providers [1]. For breast cancer, in 2017, Wu et al. trained and evaluated their models with data from a Wisconsin network of hospitals and clinics [11]. The list goes on, with some other notable works including esophageal cancer models for the Chinese population by Han et al. in 2023 [4], gastric cancer models trained on Bay Area data by Huang et al. in 2022 [6], and lung cancer models derived from a Taiwanese national database by Yeh et al. in 2021 [12].

Earlier this year, our collaborative research group between MIT and BIDMC also developed pancreatic cancer risk models [7]. The models were trained and evaluated on US data from a large federated network of health care organizations with over 60k patients in the cancer group and over 3.6m patients in the control group. Remarkably, the group externally validated their models both geographically and racially, ensuring the generalizability of the models. This thesis was an effort to extend that work to other cancer types: biliary tract, brain, colon, esophageal, gastric, kidney, liver (HCC), lung, and ovarian.

While similar previous modeling work exists for cancer types explored in this research, most such studies were subject to limited data. For instance, Liang et al.'s

2021 HCC model was developed with a dataset of fewer than 50,000 patients. On the other hand, our datasets obtained through TriNetX include hundreds of thousands, up to millions, of total patients for each cancer type. This allowed us to develop more sophisticated models and produce reliable results.

Chapter 3

Methodology

Our methodology was largely identical to that employed in the group's precursor work on pancreatic cancer [7]. We built on that to explore the possibility of prescreening for other types of cancer: biliary tract, brain, colon, esophageal, gastric, kidney, liver (HCC), lung, and ovarian.

3.1 Data source

We used anonymized medical records from the TriNetX federated EHR database, which we accessed through a collaboration agreement between BIDMC, MIT, and TriNetX. Available data include patient demographics and timestamped medical history for encounters, diagnoses, lab results, medications, and procedures.

For this study, we used data from over 50 US healthcare organizations (HCOs) from diverse geographical settings. The TriNetX network handles the varied record structures from different HCOs and harmonizes the data into a uniform format.

Diagnosis, lab, and medication records are marked with codes under different code systems. Diagnosis codes follow either the ICD-9-CM code set (deprecated, less extensive) or the ICD-10-CM code set (currently in use, more extensive), both of which are United States' adaptations of the International Classification of Disease (ICD) system. The vast majority ($> 99.9\%$) of lab codes conform to the Logical Observation Identifier Names and Codes (LOINC) standard, while the rest follow

TriNetX's internal lab encoding system. All medication records follow the RxNorm nomenclature system.

Some records in the TriNetX database were manually and structurally entered into the system, while many others were parsed from free text using natural language processing (NLP) techniques. TriNetX attaches to each entry a field specifying the corresponding derivation source.

For each cancer prediction task, we requested cancer and control datasets with hundreds of thousands or millions of total patients (see the "Raw" values in Table A.1). Dataset composition is provided in Table A.2. Relevant sample data of two *synthetic* patients is given in Table A.3 for better clarity.

ICD-9 and ICD-10 codes used to identify patients with cancer are shown in Table A.4 for biliary tract cancer, Table A.5 for brain cancer, Table A.7 for colon cancer, Table A.8 for esophageal cancer, Table A.9 for gastric cancer, Table A.10 for kidney cancer, Table A.11 for liver cancer, Table A.12 for lung cancer, and A.13 for ovarian cancer.

3.2 Framework

3.2.1 Overview

Our models solve a binary classification task: whether a patient will develop cancer in the next 6 to 18 months[1].

We considered two types of models: *singular* and *unified*. A singular model evaluates whether a patient is likely to develop a specific type of cancer. On the other hand, a unified model evaluates whether a patient is likely to develop some (any) type of cancer among all cancer types of interest. Please take note that in this work, we limited our attention to 10 cancer types, not all cancers that exist.

We employed the same, consistent framework in developing singular and unified

[1] This prediction window was taken directly from the precursor work on pancreatic cancer [7]. We believe that this standard timeframe would be a basis for a reasonable prediction timeline for most cancers and should be appropriate for our broad-scale exploratory work.

models, which we describe below.

3.2.2 Dataset construction

We constructed datasets for the purpose of model development from the raw datasets acquired from TriNetX in the following way.

We assigned a potentially different *cutoff date* to each patient. Medical records dated no later than the patient's cutoff date were aggregated into summary statistics to be used as inputs to the models. In a sense, we set it up so that the models would make a prediction for a particular patient on their specific cutoff date, using the medical records up to that day.

Patients who have been diagnosed with a cancer of interest were labeled *positive*. Others were labeled *negative*. Below, we outline how we determined cutoff dates and list the summary statistics we generated.

Cutoff dates

A cutoff date is specific to each patient. Records past the cutoff date would be excluded. The processes of generating cutoff dates differed for positive and negative patients, with that for positive patients taking place first and that for negative patients following. For positive patients, their cutoff dates were uniformly sampled to be between 6 months and 18 months prior to their first diagnosis of cancer of interest. For negative patients, we sampled their cutoff dates from the distribution of the positive patients' cutoff dates in order to prevent time-induced bias. To handle the possibility of cancer going undiagnosed, we additionally restricted the cutoff dates for negative patients to be no later than 18 months prior to the dataset retrieval date or the patients' death record (if applicable).

Summary statistics

For each patient, the longitudinal medical entries no later than their cutoff date were aggregated into summary statistics of five categories to be used as model inputs:

basic, *weight*, *diagnosis*, *lab*, and *medication*.

Basic features comprised age, whether age was known, sex, whether sex was known, and the numbers of recent (within the past 18 months) and earlier (older than 18 months) diagnosis, lab, or medication records.

Weight features consisted of normalized median weight measurements from three different timeframes: within the past 4 months, between 4 months and 1 year old, and more than 1 year old.

Only diagnosis, lab, and medication codes that appeared in at least 1 percent of the positive patients in the training set were considered. We say that such codes are *common*. We presumed that most uncommon codes were likely less relevant and removed them in an attempt to limit the number of features in the model.

Diagnosis features included, for each common diagnosis code, whether it existed in the patient, the first date on which it appeared, and the last date on which it appeared.

Lab features included, for each common lab code, whether it existed in the patient, the frequency of the lab test, the first date on which it appeared, the last date on which it appeared, the latest lab value, whether such a value existed, the rate of change of the value (for numerical lab results) or the average (for boolean lab results), and whether such a value could be computed.

Medication features included, for each common medication code, whether it existed in the patient, the frequency of the medication, the span between the first and the last prescription, and the date on which the medication was last ordered.

Additional considerations

Similar codes were grouped and treated as one according to the terminology defined by TriNetX.

Entries derived from textual records through NLP were discarded due to occasional inaccuracies.

When constructing female breast cancer and ovarian cancer datasets, we also discarded male patients and patients whose gender was unknown. While men should intrinsically never develop an ovarian cancer or be associated with a diagnosis code for female breast cancer, there existed a small number of erroneous records indicating otherwise.

Patients were sampled from the TriNetX database without bias when constructing a unified dataset. Therefore, the distribution of the 10 cancer types of interest in the constructed unified dataset closely approximated the proportions found in the TriNetX database.

Finally, patients with insufficient medical records were discarded. The sufficiency criteria were as empirically defined in the preceding work on pancreatic cancer [7]: **(1)** there were at least 16 diagnosis, lab, or medication entries in the 2-year window preceding the cutoff date **and (2)** the first and the last diagnosis, lab, or medication entries prior to the cutoff date were at least 3 months apart. The final numbers of remaining patients are shown in the "Qualified" columns of Table A.1.

3.2.3 Dataset partitioning

The data were partitioned into training (75%), test (10%), and validation sets (15%).

3.2.4 Models

We developed two classes of models: logistic regression (LR) models and neural network (NN) models. To train our logistic regression models, we used the SAGA solver [2] with balanced class weights. Our neural network models featured three fully connected layers (with 48, 16, and 1 output neurons) with tanh nonlinearity in hidden layers. We also used the BinMask sparsification technique to mitigate overfitting [8].

3.3 Evaluation

3.3.1 Singular models

We independently evaluated our algorithm with respect to each pair (t, c) (where t is a cancer type and $c \in \{\text{LR}, \text{NN}\}$ is a model class) on two areas.

General performance

To measure general performance, we repeated datasplitting and model training for a total of nine times with different random seeds and weight initialization. This resulted in nine distinct models $m_0, \ldots, m_8$ of the same model class c and specialized for the same cancer type t. We then tested each model by using it to predict relative cancer risk in test-set patients. The average area under the receiver operating curve (AUC) across all nine models was computed and reported. The actual receiver operating curve (ROC) for model m_0 was also recorded and analyzed.

Model generalizability

We evaluated model generalizability by performing external validations on HCO geographical locations (Midwest, Northeast, South, or West) and patient races (AIAN[2], Asian, Black, NHPI[3], or White).

Let A be the set of HCO locations and B be the set of patient races, as listed above. Then, our external validation process was as follows.

For each partition label $l \in A \cup B$, we trained a model on the train set *excluding* all entries associated with label l. Then, we evaluated this model on **(1)** the validation set *excluding* all entries associated with label l **and (2)** the test set *with only* entries associated with label l. We call the difference in AUCs observed across these two sets the val/test AUC gap. The more generalizable a model is, the smaller (and closer to zero) we would expect to see this number be.

Additionally, we evaluated time-wise generalizability through temporal validation. Let $D = \{d_5, d_6, \ldots, d_9\}$ be the set of 50th, 60th, …, 90th percentiles from the distribution of cancer diagnosis dates. For each dataset split date $d \in D$, we trained models with medical records up to that date d and tested them on a test set with cancer diagnosis dates after d_9. The more temporally generalizable a model is, the more consistent we would expect performance to be across all $d \in D$.

[2] AIAN stands for American Indian and Alaska Native

[3] NHPI stands for Native Hawaiian and Pacific Islander

3.3.2 Unified models

Unified models were evaluated in all ways that singular models were. More precisely, we treated the unified models as if they were specialized for an abstract, imaginary cancer type that corresponds to the union of the 10 original cancer types of interest.

Additionally, in evaluating the general performance, we also used model m_0 to predict by-type cancer risks by running m_0 on individual-cancer test sets. The AUC performance for each cancer type was observed and recorded. We hoped that comparing this number to the performance of singular models would reveal insights on the commonalities between the factors that cause or correlate with different types of cancer.

Chapter 4

Results

4.1 ROC curves

Areas under ROC curves were as described in Table 4.1. Singular models achieved an average AUC of between 0.761 and 0.920 for LR models and between 0.771 and 0.928 for NN models. Unified models showed worse performance on individual cancer prediction tasks than the specialized singular models themselves, with AUCs ranging from 0.723 to 0.815 for the unified LR model and from 0.733 to 0.825 for the unified NN model.

The actual ROC curves from one of the nine runs are shown in Figure B-1 for the singular models and in Figure B-2 for the unified models. All curves appear to exhibit a smooth concave profile without remarkable characteristics.

4.2 External validation

4.2.1 By race

Results of external validation by race are shown in Figure B-3 for singular models and in Figure B-4 for unified models.

Most singular models saw AUC gaps of approximately no greater than 0.05. In other models, the gaps may occasionally be as large as roughly 0.1 for external vali-

Table 4.1: Areas under ROC curves (AUCs)

Cancer	Singular model		Unified model	
	LR	NN	LR	NN
Biliary Tract	0.811*	0.813*	0.778	0.781
Brain	0.793*	0.808*	0.723	0.733
Breast	0.761*	0.771*	0.734	0.742
Colon	0.787*	0.789*	0.742	0.750
Esophageal	0.825*	0.834*	0.733	0.738
Gastric	0.799*	0.805*	0.747	0.756
Kidney	0.813*	0.820*	0.733	0.741
Liver	0.920*	0.928*	0.815	0.825
Lung	0.846*	0.852*	0.797	0.802
Ovarian	0.768*	0.780*	0.736	0.742
All of the above (Unified)	-	-	0.772*	0.779*

Note: Figures marked with an asterisk were the average AUCs from nine random runs. Other figures were the AUCs from a single run.

dation on the NHPI race group, for which we had little data available (take note of the abnormally wide confidence intervals). The val/test AUC gaps observed by the unified models were also no more than 0.05.

We conclude that our models generalize well in terms of race.

4.2.2 By HCO geographical location

Results of external validation by HCO location are shown in Figure B-5 for singular models and in Figure B-6 for unified models.

In all singular models, the val/test AUC gaps were no more than approximately 0.06. In the unified models, the gaps were smaller than 0.04.

We conclude that our models generalize well in terms of HCO location.

4.2.3 Temporal validation

Results of temporal validation are shown in Figure B-7 for singular models and in Figure B-8 for unified models.

In all singular models, the fluctuations in test AUC across different data splitting dates (spanning approximately three to five years) were less than 0.05, with the majority registering even below 0.03. In the unified models, the variations were noticeably smaller, at less than 0.01.

Hence, we conclude that our models demonstrate robust generalization across time periods.

4.3 Feature analysis

In addition to the evaluations above, we ranked the predictive power of features in the NN models. For each feature f, we computed the AUCs achieved by running the NN models on the test set with no information other than feature f (i.e., all other features would be blanked out). We call such quantities *univariate AUCs.* Results are shown in Figure B-9 for singular NN models and in Figure B-10 for the unified NN model.

Chapter 5

Discussion

5.1 Findings

Singular models we obtained achieved average AUCs ranging from 0.76 to 0.93 for different cancer types. Unified models performed noticeably worse on individual cancer prediction tasks when compared to the correspondingly specialized individual models, which was an expected finding. Furthermore, neural network models consistently albeit marginally outperformed logistic regression models on all cancer prediction tasks, individual or unified. Geographical, racial, and temporal external validations all substantiated the robustness of our models.

Many empirically important predictors in our models align with the current understanding in oncology. For instance, in our singular LR models, cirrhosis came out to be the absolute top feature for biliary tract and liver cancers, and dysphagia was the absolute top feature for esophageal cancer. Lab results such as the measurements of alkaline phosphatase, bilirubin, or lymphocytes in blood and a complete blood count also proved to be powerful predictors for some cancers.

Model performance varied quite widely across different cancer types, mirroring the reality that certain cancers are more predictable than others. For example, cirrhosis can be found in up to 80 to 90 percent of liver cancer patients [3], so a cirrhosis diagnosis can be indicative of an elevated liver cancer risk, relative to the non-cirrhosis population. On the other hand, it was a surprising finding that we were able to achieve an AUC of up to 0.81 in predicting the brain cancer, for which little is known in the medical field today. Analyzing the list of top important model features revealed that our models might be taking advantage of an existing suspicion or diagnosis of a brain tumor (not yet classified as cancerous), through information such as the use of dexamethasone and records of encounters for chemotherapy. This suggests that an evaluation by oncologists beyond the numbers may be necessary for us to gain an accurate understanding of the capabilities and the potential impact of our models.

5.2 Potential use scenarios

Compatible with the TriNetX network, our models can directly be deployed to detect high risk individuals across the country. Medical professionals and researchers may also review the interaction between the features and the predictions in our models to develop a more in-depth understanding of cancer risk factors and potentially discover currently unknown relationships through further investigation.

5.3 Limitations

Our study has a number of limitations.

First, the study used data only from HCOs within the United States. While location-based and race-based external validation verified that our models are geographically and racially generalizable within the US, our results may not extend to outside the country, for instance possibly due to varying care practices and patient demographics.

Second, the study was retrospective and used data only from the past. While temporal validation demonstrated great model generalizability across time periods, we will still need to evaluate the efficacy of our models clinically in a prospective study in order to understand their true impact.

Third, our models used an uncurated list of features to make predictions. Some of them may be, for example, lab tests that providers would order when they are already suspecting an ongoing cancer case, as alluded to in Section 5.1. Such features would allow our models to make potentially many more accurate diagnoses, but then the value added of our models may not be interpreted solely from looking at the prediction performance within our dataset (such as the computed AUCs). Again, this calls for a prospective study so that real world performance of our models can be assessed.

Finally, the model architectures investigated in this work expect flat-array, fixed-size inputs. Some trends that exist in the initial, temporal data may not be captured through our feature extraction pipeline. On the other hand, sequence models that intrinsically take in sequential inputs would not be susceptible to such information loss, which in turn might allow for better prediction capabilities, for instance as employed in another recent work on pancreatic cancer by Placido et al. [10]. In other words, our prediction performance can potentially be improved further, as our work can benefit from a more extensive model architecture exploration.

Chapter 6

Conclusion and future work

In this study, we took the framework our research group had used to develop pancreatic cancer risk models and extended it to ten other cancer types: biliary tract, brain, breast (female), colon, esophageal, gastric, kidney, liver, lung, and ovarian. The resulting models showed great overall performance. Neural network models performed better than logistic regression models in all cancers, achieving average AUC scores ranging from 0.771 to 0.928 for different cancer types. We found that, on individual cancer prediction tasks, unified models performed noticeably worse than the specialized singular models. The models were robust and experienced only minor AUC drops when externally validated on HCO locations and patient races as well as across different time periods.

Moving forward, it will be a good idea to consult with specialized professionals to review the important model features for their practical relevance. As suggested in the previous section, a prospective study should also be conducted through deployment in the TriNetX network to fully evaluate the models and potentially begin making a real-world impact. We hope that our models will allow the TriNetX network to promptly notify providers when patients under their care are determined to be at risk. Through this, our models will help save lives by promoting early cancer detection and thereby enabling effective treatment.

Appendix A

Tables

Table A.1: Patient counts in the datasets

Cancer	Cancer set		Control set	
	Raw	Qualified	Raw	Qualified
Biliary Tract	45,499	21,949	100,007	55,559
Brain	96,601	40,044	500,000	271,984
Breast	768,335	362,788	3,000,001	883,176
Colon	309,748	145,318	296,082	164,532
Esophageal	62,638	27,387	299,871	164,427
Gastric	61,044	28,129	99,997	55,368
Kidney	152,606	81,141	299,871	164,004
Liver	490,440	49,380	1,999,999	1,093,961
Lung	425,900	205,630	999,563	546,845
Ovarian	130,602	40,411	2,000,002	613,092
All of the above (Unified)	500,002	248,395	1,499,998	801,032

Table A.2: Dataset contents

Filename	Used?	Description
`patient.csv`	✓	Patient demographics
`diagnosis.csv`	✓	Diagnosis records
`lab_result.csv`	✓	Lab results
`standardized_term.csv`	✓	Standardized terminology
`vitals_signs.csv`	✓	Vital signs records
`dataset_details.csv`	✓	Dataset details
`med_ingredients.csv`	✓	Medication records
`medication_drug.csv`		Medication records (subset of the above file)
`procedure.csv`		Records of procedures undergone
`genomic.csv`		Genomic records
`tumor.csv`		Tumor records
`tumor_properties.csv`		Tumor properties
`oncology_treatment.csv`		Oncology treatment records
`cohort_details.csv`		Dataset cohort details
`patient_cohort.csv`		A mapping from each patient to their corresponding cohort
`chemo_lines.csv`		Records of chemotherapy lines of treatment
`encounter.csv`		Encounter records

Note: Some filenames shown are abbreviated for conciseness.

Table A.3: Sample data with two synthetic patients

(a) patient.csv

patient_id	sex	race	yob	...	region	source_id
42abc	M	White	1942	...	Midwest	EHR
mit77	F	Asian	1989	...	South	NLP

(b) diagnosis.csv

patient_id	system	code	date	...	source_id
42abc	ICD-9-CM	401.9	20180720	...	EHR
42abc	ICD-10-CM	C22.0	20200101	...	EHR
mit77	ICD-9-CM	155.0	20200103	...	EHR

(c) lab_result.csv

patient_id	system	code	val	unit	date	...	source_id
42abc	LOINC	11050-2	100	mg/DL	20180720	...	EHR

(d) standardized_term.csv

system	code	code_desc	path	unit
ICD-10-CM	C22.0	Liver cell carcinoma	.../C15-C26/C22/C22.0	N/A
ICD-9-CM	155.0	Malignant neoplasm of liver, primary	.../C15-C26/C22/C22.0/155.0	N/A
ICD-9-CM	401.9	Unspecified essential hypertension	.../I10-I15/I10/401.9	N/A

(e) vitals_signs.csv

patient_id	system	code	val	unit	date	...	source_id
42abc	LOINC	3141-9	150.2	lb	20180720	...	EHR

(f) dataset_details.csv

num_unique_patients	num_HCOs	date_created
2	2	20230501

(g) med_ingredients.csv

patient_id	system	code	brand	strength	start_date	...	source_id
42abc	RxNorm	25480	Neurontin	300 MG	20180720	...	EHR

Note: Some column names shown are abbreviated for conciseness. Filenames are consistent with Table A.2.

Table A.4: ICD codes for biliary tract cancer

System	Code	Description
ICD-10-CM	C22.1	Intrahepatic bile duct carcinoma
	C23	Malignant neoplasm of gallbladder
	C24.0	Malignant neoplasm of extrahepatic bile duct
	C24.8	Malignant neoplasm of overlapping sites of biliary tract
	C24.9	Malignant neoplasm of biliary tract, unspecified
ICD-9-CM	155.1	Malignant neoplasm of intrahepatic bile ducts
	156.0	Malignant neoplasm of gallbladder
	156.1	Malignant neoplasm of extrahepatic bile ducts
	156.8	Malignant neoplasm of other specified sites of gallbladder and extrahepatic bile ducts
	156.9	Malignant neoplasm of biliary tract, part unspecified site

Table A.5: ICD codes for brain cancer

System	Code	Description
ICD-10-CM	C71.0	Malignant neoplasm of cerebrum, except lobes and ventricles
	C71.1	Malignant neoplasm of frontal lobe
	C71.2	Malignant neoplasm of temporal lobe
	C71.3	Malignant neoplasm of parietal lobe
	C71.4	Malignant neoplasm of occipital lobe
	C71.5	Malignant neoplasm of cerebral ventricle
	C71.6	Malignant neoplasm of cerebellum
	C71.7	Malignant neoplasm of brain stem
	C71.8	Malignant neoplasm of overlapping sites of brain
	C71.9	Malignant neoplasm of brain, unspecified
ICD-9-CM	191.0	Malignant neoplasm of cerebrum, except lobes and ventricles
	191.1	Malignant neoplasm of frontal lobe
	191.2	Malignant neoplasm of temporal lobe
	191.3	Malignant neoplasm of parietal lobe
	191.4	Malignant neoplasm of occipital lobe
	191.5	Malignant neoplasm of ventricles
	191.6	Malignant neoplasm of cerebellum nos
	191.7	Malignant neoplasm of brain stem
	191.8	Malignant neoplasm of other parts of brain
	191.9	Malignant neoplasm of brain, unspecified

Table A.6: ICD codes for female breast cancer

System	Code	Description
ICD-10-CM	C50	Malignant neoplasm of breast
	C50.0	Malignant neoplasm of nipple and areola
	C50.01	Malignant neoplasm of nipple and areola, female
	C50.011	Malignant neoplasm of nipple and areola, right female breast
	C50.012	Malignant neoplasm of nipple and areola, left female breast
	C50.019	Malignant neoplasm of nipple and areola, unspecified female
	C50.1	Malignant neoplasm of central portion of breast
	C50.11	Malignant neoplasm of central portion of breast, female
	C50.111	Malignant neoplasm of central portion of right female breast
	C50.112	Malignant neoplasm of central portion of left female breast
	C50.119	Malignant neoplasm of central portion of unspecified female breast
	C50.2	Malignant neoplasm of upper-inner quadrant of breast
	C50.21	Malignant neoplasm of upper-inner quadrant of breast, female
	C50.211	Malignant neoplasm of upper-inner quadrant of right female breast
	C50.212	Malignant neoplasm of upper-inner quadrant of left female breast
	C50.219	Malignant neoplasm of upper-inner quadrant of unspecified female breast
	C50.3	Malignant neoplasm of lower-inner quadrant of breast
	C50.31	Malignant neoplasm of lower-inner quadrant of breast, female
	C50.311	Malignant neoplasm of lower-inner quadrant of right female breast
	C50.312	Malignant neoplasm of lower-inner quadrant of left female breast
	C50.319	Malignant neoplasm of lower-inner quadrant of unspecified female breast

Table A.6: ICD codes for female breast cancer (continued)

System	Code	Description
	C50.4	Malignant neoplasm of upper-outer quadrant of breast
	C50.41	Malignant neoplasm of upper-outer quadrant of breast, female
	C50.411	Malignant neoplasm of upper-outer quadrant of right female breast
	C50.412	Malignant neoplasm of upper-outer quadrant of left female breast
	C50.419	Malignant neoplasm of upper-outer quadrant of unspecified female breast
	C50.5	Malignant neoplasm of lower-outer quadrant of breast
	C50.51	Malignant neoplasm of lower-outer quadrant of breast, female
	C50.511	Malignant neoplasm of lower-outer quadrant of right female breast
	C50.512	Malignant neoplasm of lower-outer quadrant of left female breast
	C50.519	Malignant neoplasm of lower-outer quadrant of unspecified female breast
	C50.6	Malignant neoplasm of axillary tail of breast
	C50.61	Malignant neoplasm of axillary tail of breast, female
	C50.611	Malignant neoplasm of axillary tail of right female breast
	C50.612	Malignant neoplasm of axillary tail of left female breast
	C50.619	Malignant neoplasm of axillary tail of unspecified female breast
	C50.8	Malignant neoplasm of overlapping sites of breast
	C50.81	Malignant neoplasm of overlapping sites of breast, female
	C50.811	Malignant neoplasm of overlapping sites of right female breast
	C50.812	Malignant neoplasm of overlapping sites of left female breast
	C50.819	Malignant neoplasm of overlapping sites of unspecified female breast
	C50.9	Malignant neoplasm of breast of unspecified site
	C50.91	Malignant neoplasm of breast of unspecified site, female
	C50.911	Malignant neoplasm of unspecified site of right female breast
	C50.912	Malignant neoplasm of unspecified site of left female breast
	C50.919	Malignant neoplasm of unspecified site of unspecified female breast

Table A.6: ICD codes for female breast cancer (continued)

System	Code	Description
ICD-9-CM	174.0	Malignant neoplasm of nipple and areola of female breast
	174.1	Malignant neoplasm of central portion of female breast
	174.3	Malignant neoplasm of lower-inner quadrant of female breast
	174.4	Malignant neoplasm of upper-outer quadrant of female breast
	174.5	Malignant neoplasm of lower-outer quadrant of female breast
	174.6	Malignant neoplasm of axillary tail of female breast
	174.8	Malignant neoplasm of other specified sites of female breast
	174.9	Malignant neoplasm of breast (female), unspecified

Table A.7: ICD codes for colon cancer

System	Code	Description
ICD-10-CM	C18	Malignant neoplasm of colon
	C18.0	Malignant neoplasm of cecum
	C18.1	Malignant neoplasm of appendix
	C18.2	Malignant neoplasm of ascending colon
	C18.3	Malignant neoplasm of hepatic flexure
	C18.4	Malignant neoplasm of transverse colon
	C18.5	Malignant neoplasm of splenic flexure
	C18.6	Malignant neoplasm of descending colon
	C18.7	Malignant neoplasm of sigmoid colon
	C18.8	Malignant neoplasm of overlapping sites of colon
	C18.9	Malignant neoplasm of colon, unspecified
ICD-9-CM	153	Malignant neoplasm of colon
	153.0	Malignant neoplasm of hepatic flexure
	153.1	Malignant neoplasm of transverse colon
	153.2	Malignant neoplasm of descending colon
	153.3	Malignant neoplasm of sigmoid colon
	153.4	Malignant neoplasm of cecum
	153.5	Malignant neoplasm of appendix vermiformis
	153.6	Malignant neoplasm of ascending colon
	153.7	Malignant neoplasm of splenic flexure
	153.8	Malignant neoplasm of other specified sites of large intestine
	153.9	Malignant neoplasm of colon, unspecified site

Table A.8: ICD codes for esophageal cancer

System	Code	Description
ICD-10-CM	C15.3	Malignant neoplasm of upper third of esophagus
	C15.4	Malignant neoplasm of middle third of esophagus
	C15.5	Malignant neoplasm of lower third of esophagus
	C15.8	Malignant neoplasm of overlapping sites of esophagus
	C15.9	Malignant neoplasm of esophagus, unspecified
ICD-9-CM	150.0	Malignant neoplasm of cervical esophagus
	150.1	Malignant neoplasm of thoracic esophagus
	150.2	Malignant neoplasm of abdominal esophagus
	150.3	Malignant neoplasm of upper third of esophagus
	150.4	Malignant neoplasm of middle third of esophagus
	150.5	Malignant neoplasm of lower third of esophagus
	150.8	Malignant neoplasm of other specified part of esophagus
	150.9	Malignant neoplasm of esophagus, unspecified site

Table A.9: ICD codes for gastric cancer

System	Code	Description
ICD-10-CM	C16	Malignant neoplasm of stomach
	C16.0	Malignant neoplasm of cardia
	C16.1	Malignant neoplasm of fundus of stomach
	C16.2	Malignant neoplasm of body of stomach
	C16.3	Malignant neoplasm of pyloric antrum
	C16.4	Malignant neoplasm of pylorus
	C16.5	Malignant neoplasm of lesser curvature of stomach, unspecified
	C16.6	Malignant neoplasm of greater curvature of stomach, unspecified
	C16.8	Malignant neoplasm of overlapping sites of stomach
	C16.9	Malignant neoplasm of stomach, unspecified
ICD-9-CM	151	Malignant neoplasm of stomach
	151.0	Malignant neoplasm of cardia
	151.1	Malignant neoplasm of pylorus
	151.2	Malignant neoplasm of pyloric antrum
	151.3	Malignant neoplasm of fundus of stomach
	151.4	Malignant neoplasm of body of stomach
	151.5	Malignant neoplasm of lesser curvature of stomach, unspecified
	151.6	Malignant neoplasm of greater curvature of stomach, unspecified
	151.8	Malignant neoplasm of other specified sites of stomach
	151.9	Malignant neoplasm of stomach, unspecified site

Table A.10: ICD codes for kidney cancer

System	Code	Description
ICD-10-CM	C64.1	Malignant neoplasm of right kidney, except renal pelvis
	C64.2	Malignant neoplasm of left kidney, except renal pelvis
	C64.9	Malignant neoplasm of unspecified kidney, except renal pelvis
ICD-9-CM	189.0	Malignant neoplasm of kidney, except pelvis

Table A.11: ICD codes for liver cancer

System	Code	Description
ICD-10-CM	C22.0	Liver cell carcinoma
ICD-9-CM	155.0	Malignant neoplasm of liver, primary

Table A.12: ICD codes for lung cancer

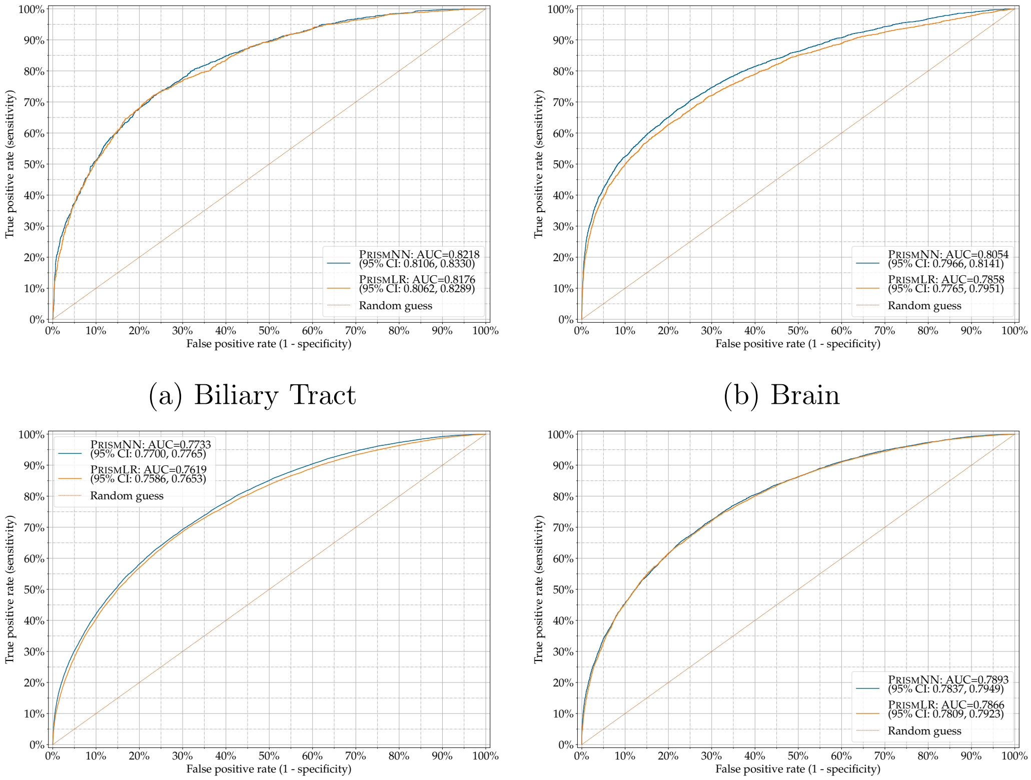
System	Code	Description
ICD-10-CM	C34.0	Malignant neoplasm of main bronchus
	C34.00	Malignant neoplasm of unspecified main bronchus
	C34.01	Malignant neoplasm of right main bronchus
	C34.02	Malignant neoplasm of left main bronchus
	C34.1	Malignant neoplasm of upper lobe, bronchus or lung
	C34.10	Malignant neoplasm of upper lobe, unspecified bronchus or lung
	C34.11	Malignant neoplasm of upper lobe, right bronchus or lung
	C34.12	Malignant neoplasm of upper lobe, left bronchus or lung
	C34.2	Malignant neoplasm of middle lobe, bronchus or lung
	C34.3	Malignant neoplasm of lower lobe, bronchus or lung
	C34.30	Malignant neoplasm of lower lobe, unspecified bronchus or lung
	C34.31	Malignant neoplasm of lower lobe, right bronchus or lung
	C34.32	Malignant neoplasm of lower lobe, left bronchus or lung
	C34.8	Malignant neoplasm of overlapping sites of bronchus and lung
	C34.80	Malignant neoplasm of overlapping sites of unspecified bronchus and lung
	C34.81	Malignant neoplasm of overlapping sites of right bronchus and lung
	C34.82	Malignant neoplasm of overlapping sites of left bronchus and lung
	C34.9	Malignant neoplasm of unspecified part of bronchus or lung
	C34.90	Malignant neoplasm of unspecified part of unspecified bronchus or lung
	C34.91	Malignant neoplasm of unspecified part of right bronchus or lung
	C34.92	Malignant neoplasm of unspecified part of left bronchus or lung
ICD-9-CM	162.2	Malignant neoplasm of main bronchus
	162.3	Malignant neoplasm of upper lobe, bronchus or lung
	162.4	Malignant neoplasm of middle lobe, bronchus or lung
	162.5	Malignant neoplasm of lower lobe, bronchus or lung
	162.8	Malignant neoplasm of other parts of bronchus or lung
	162.9	Malignant neoplasm of bronchus and lung, unspecified

Table A.13: ICD codes for ovarian cancer

System	Code	Description
ICD-10-CM	C56.1	Malignant neoplasm of right ovary
	C56.2	Malignant neoplasm of left ovary
	C56.3	Malignant neoplasm of bilateral ovaries
	C56.9	Malignant neoplasm of unspecified ovary
ICD-9-CM	183.0	Malignant neoplasm of ovary

Appendix B

Figures



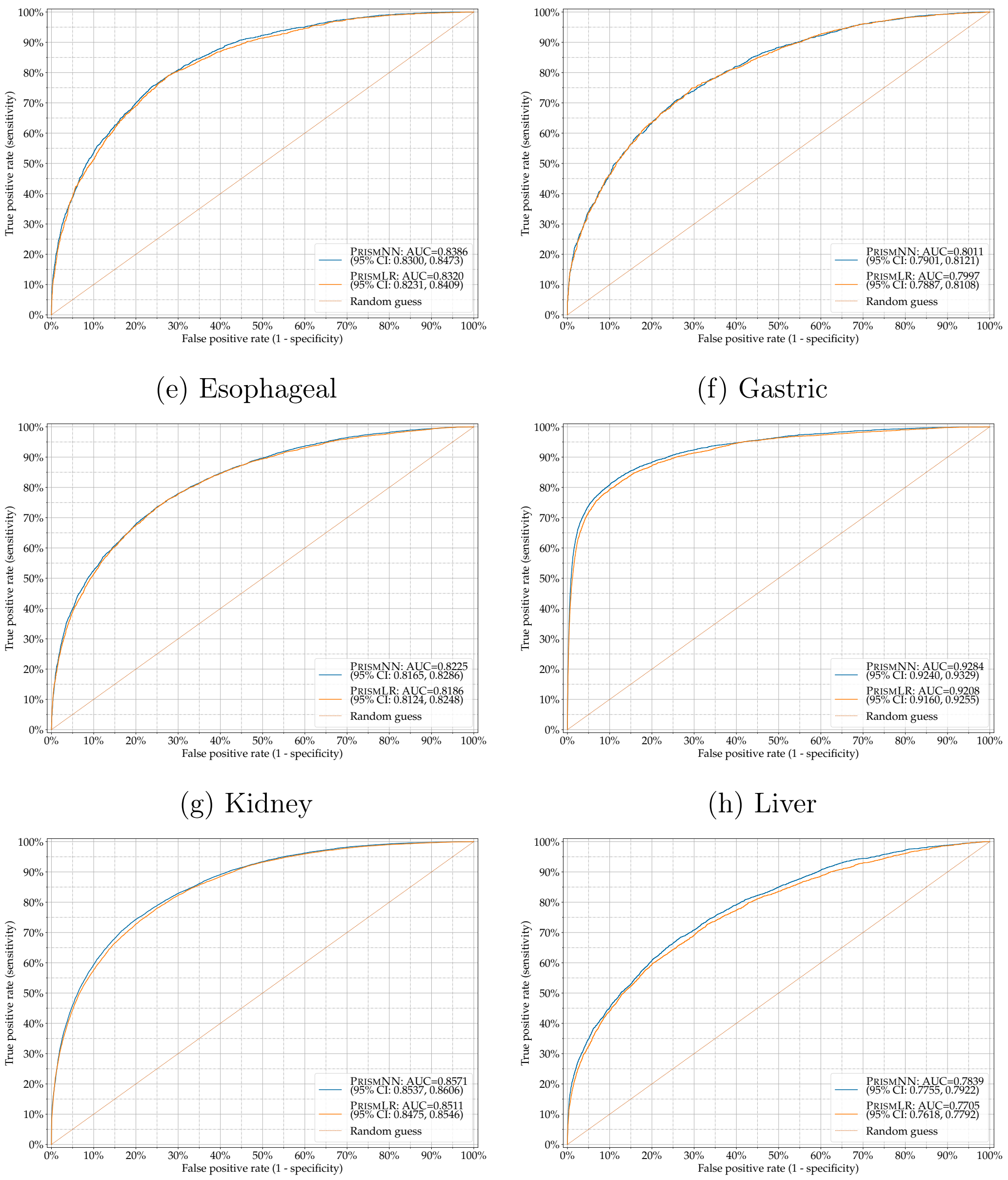
(a) Biliary Tract

(b) Brain

(c) Breast (Female)

(d) Colon

Figure B-1: ROC curves in singular models



(e) Esophageal

(f) Gastric

(g) Kidney

(h) Liver

(i) Lung

(j) Ovarian

Figure B-1: ROC curves in singular models (continued)

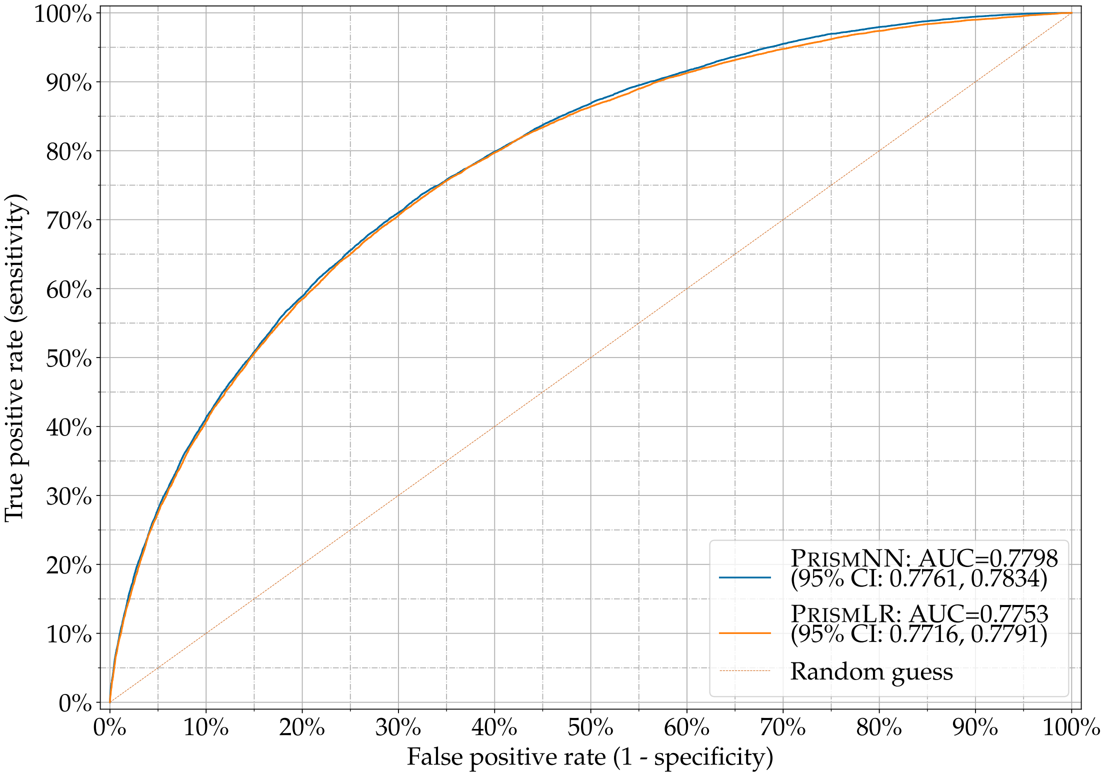
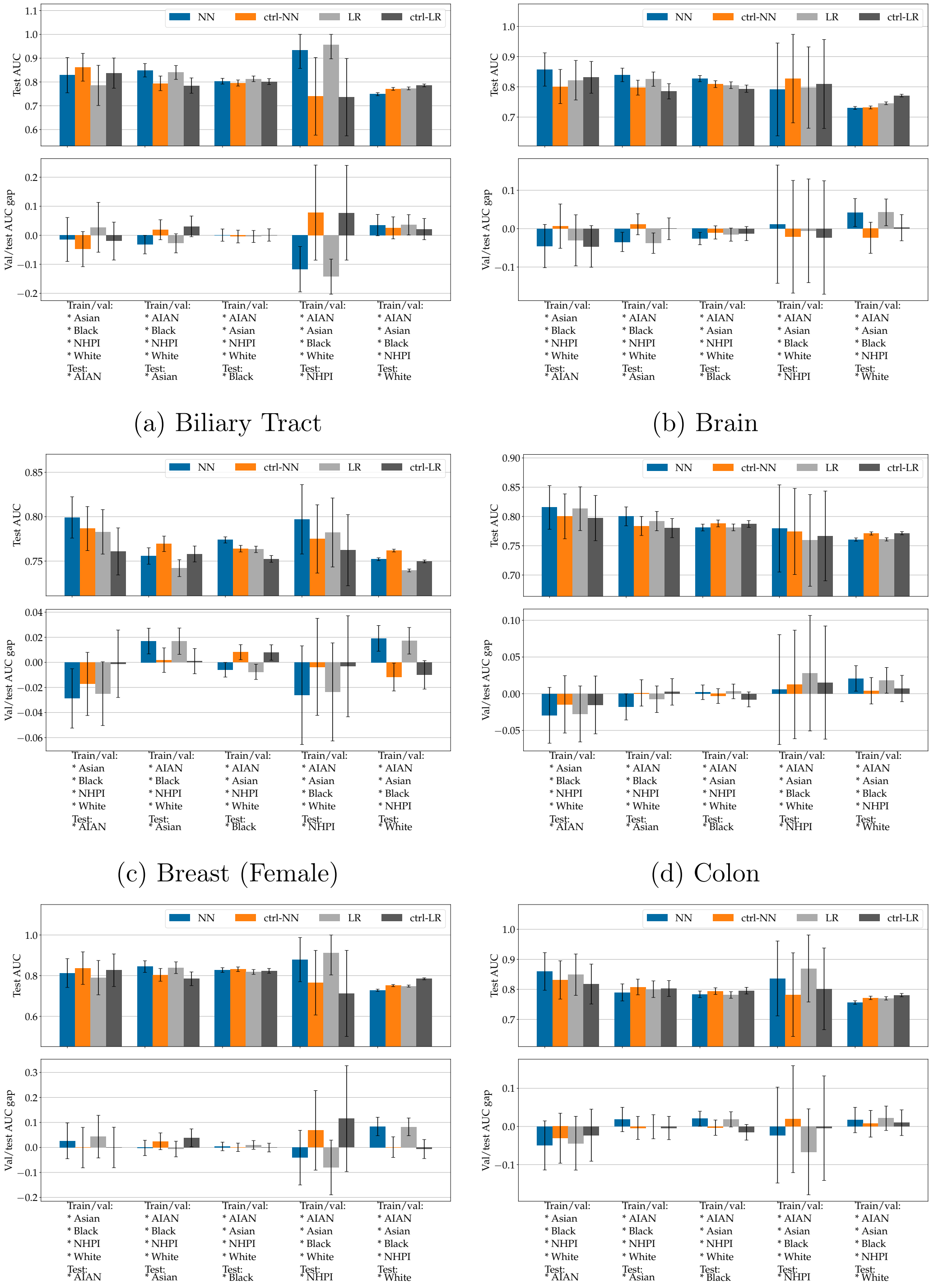


Figure B-2: ROC curves in unified models



(a) Biliary Tract

(b) Brain

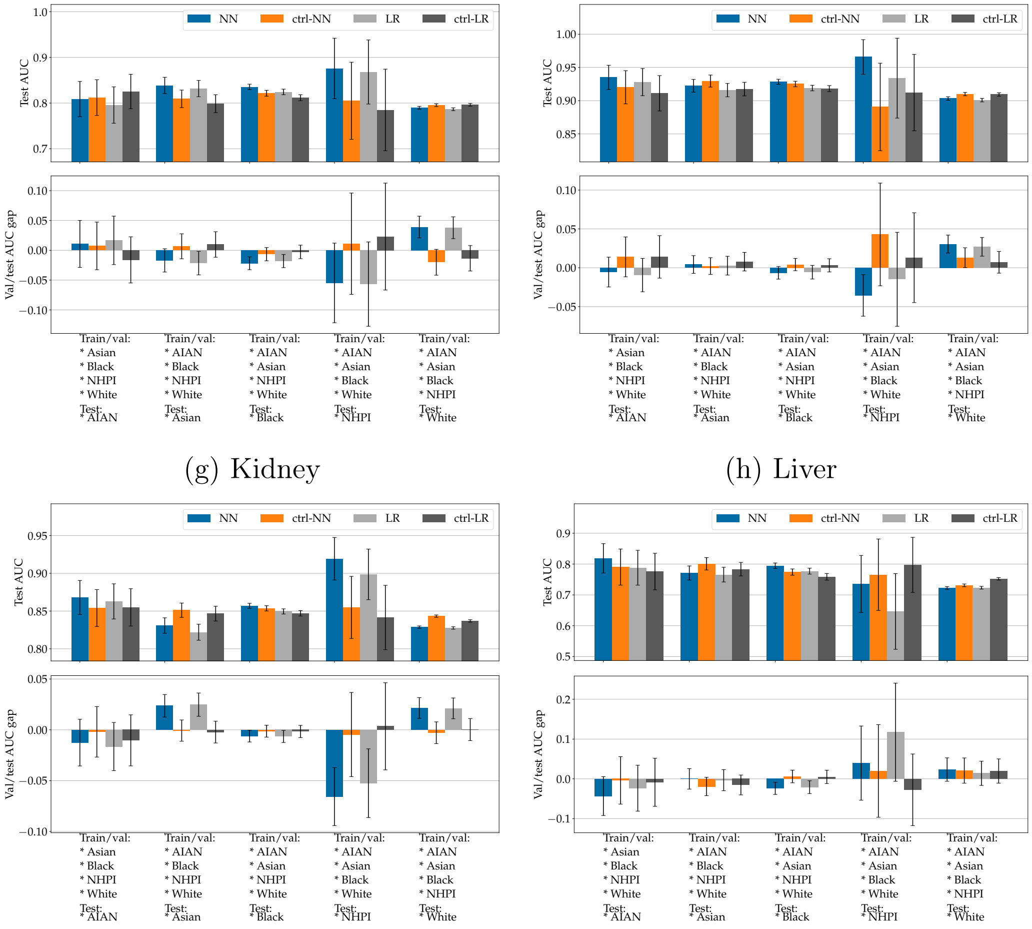
(c) Breast (Female)

(d) Colon

(e) Esophageal

(f) Gastric

Figure B-3: External validation by race for singular models



(g) Kidney

(h) Liver

(i) Lung

(j) Ovarian

Figure B-3: External validation by race for singular models (continued)

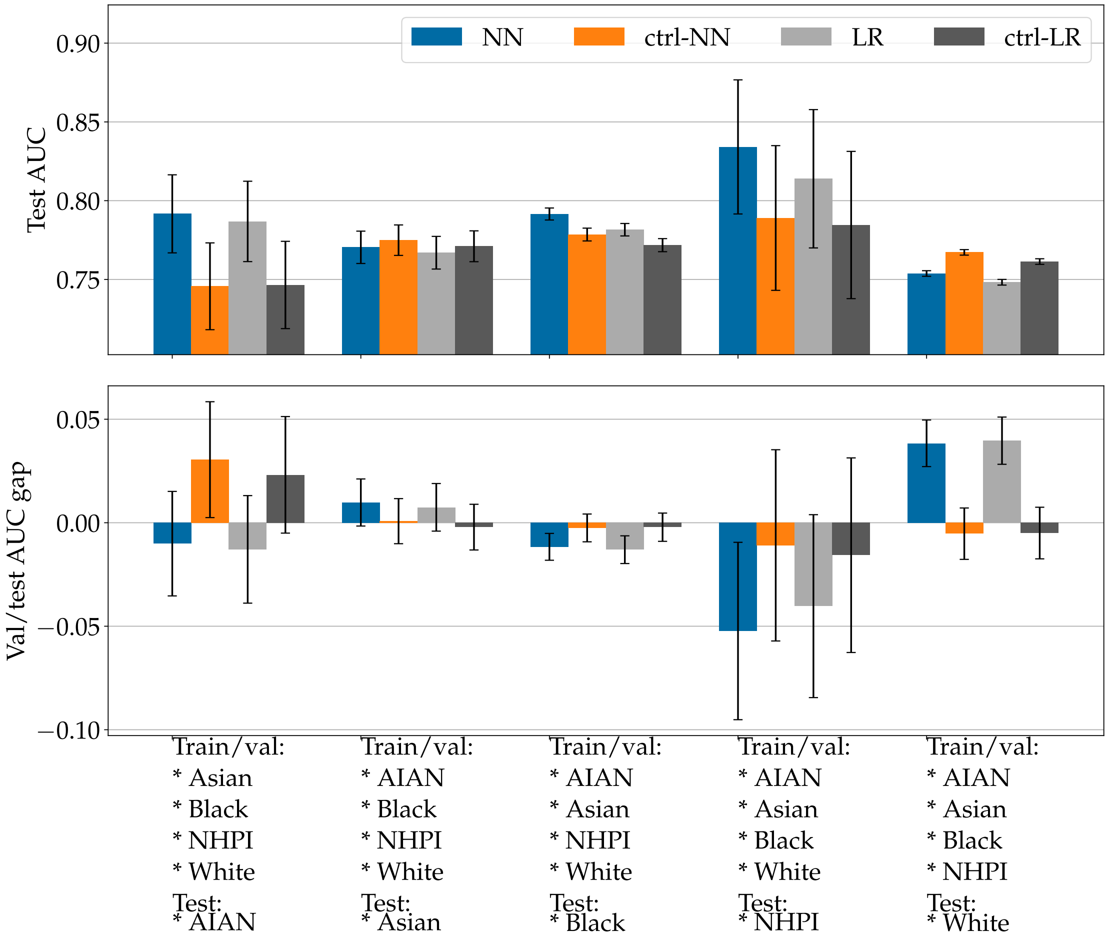
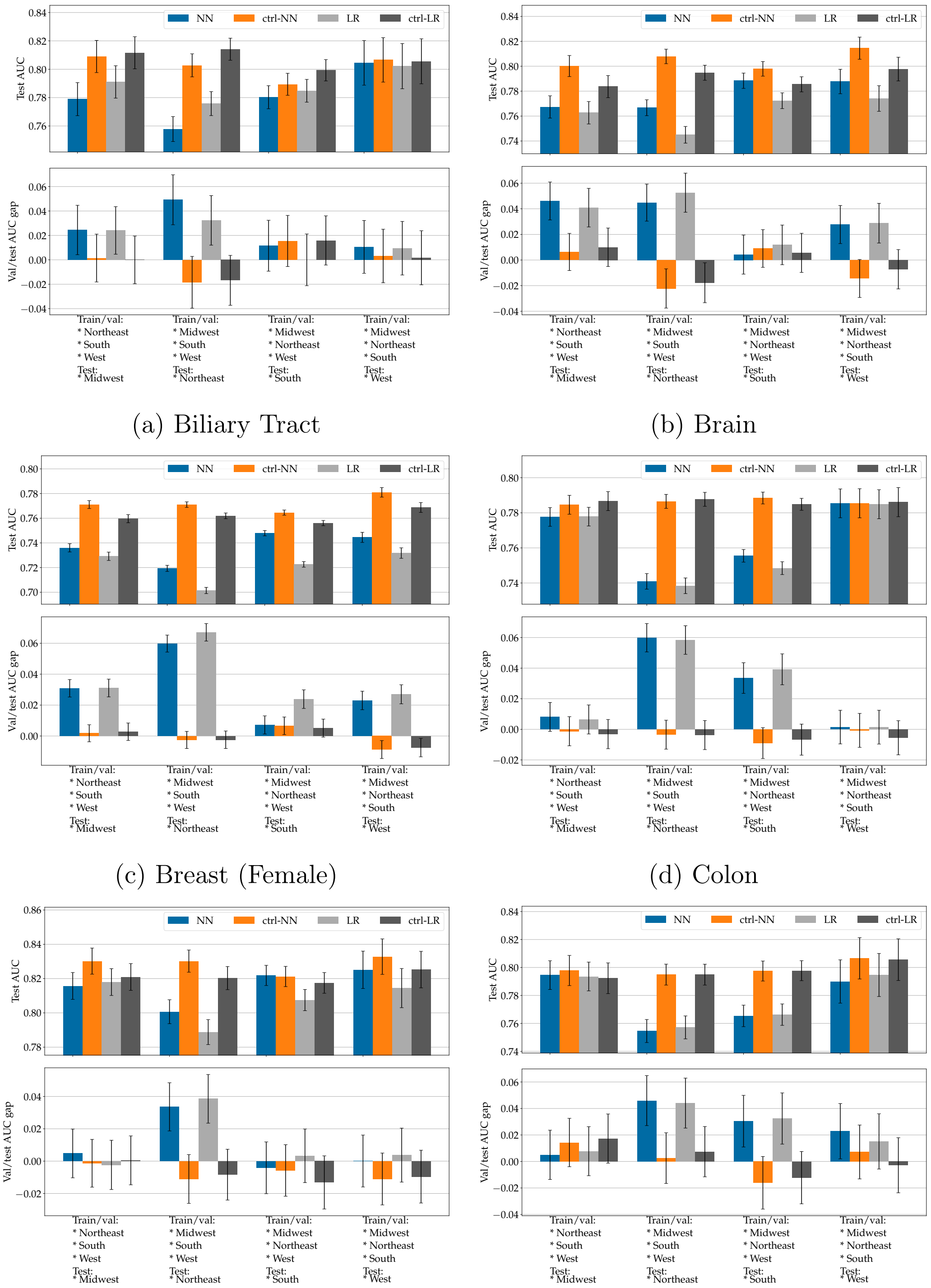


Figure B-4: External validation by race for unified models



(a) Biliary Tract

(b) Brain

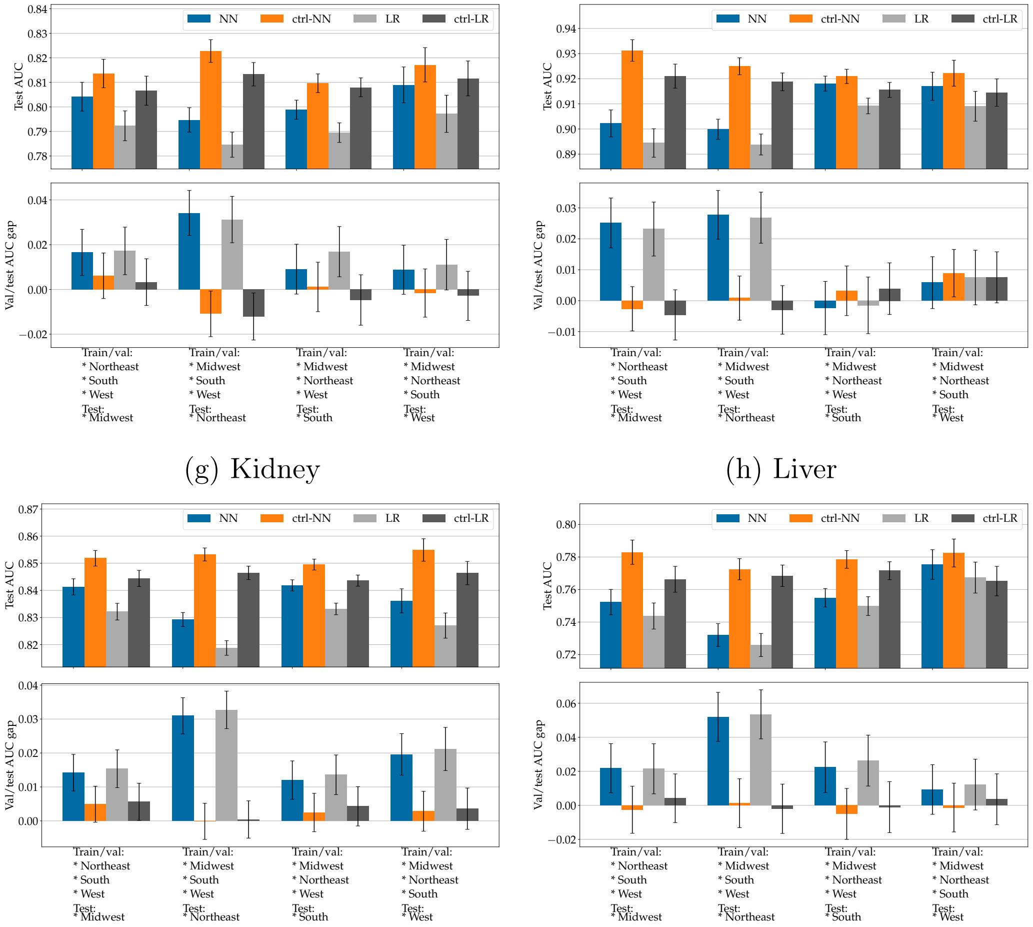
(c) Breast (Female)

(d) Colon

(e) Esophageal

(f) Gastric

Figure B-5: External validation by location for singular models



(g) Kidney

(h) Liver

(i) Lung

(j) Ovarian

Figure B-5: External validation by location for singular models (continued)

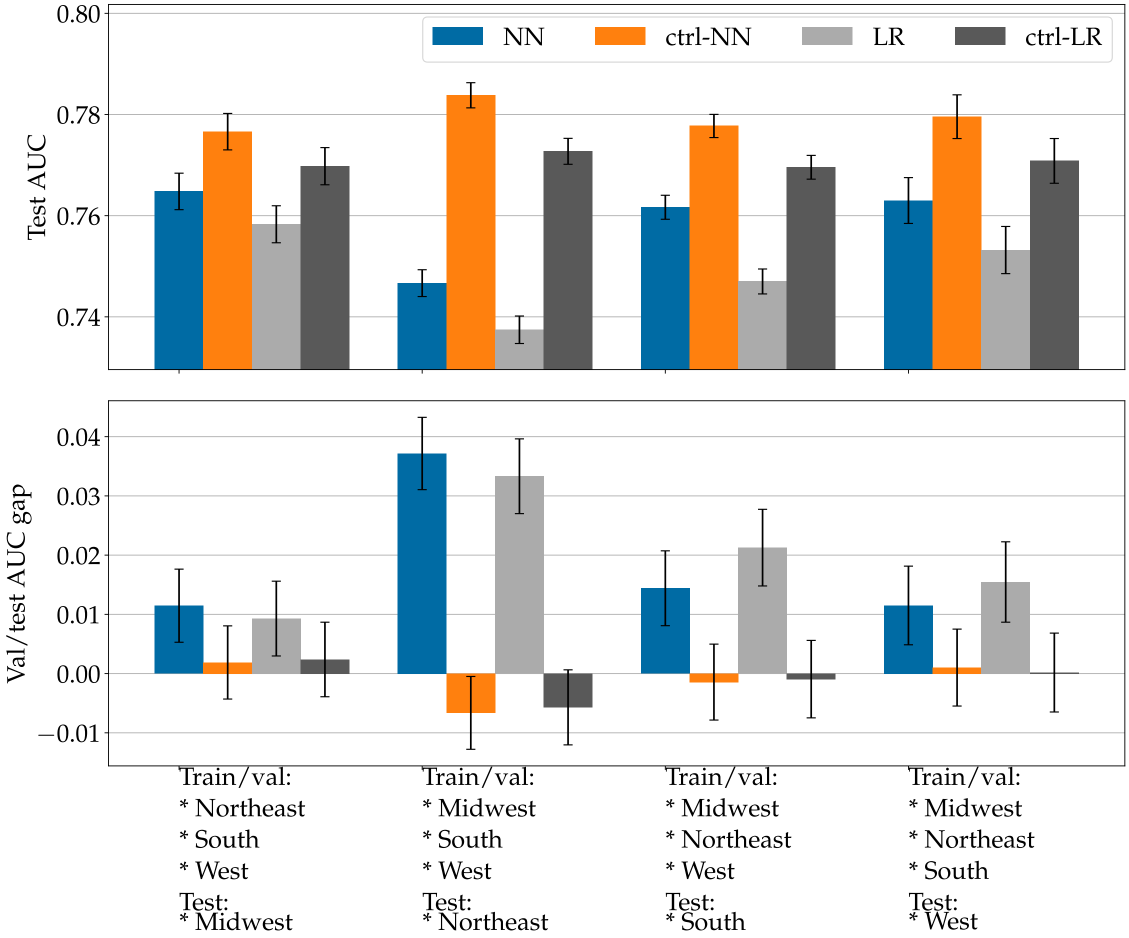
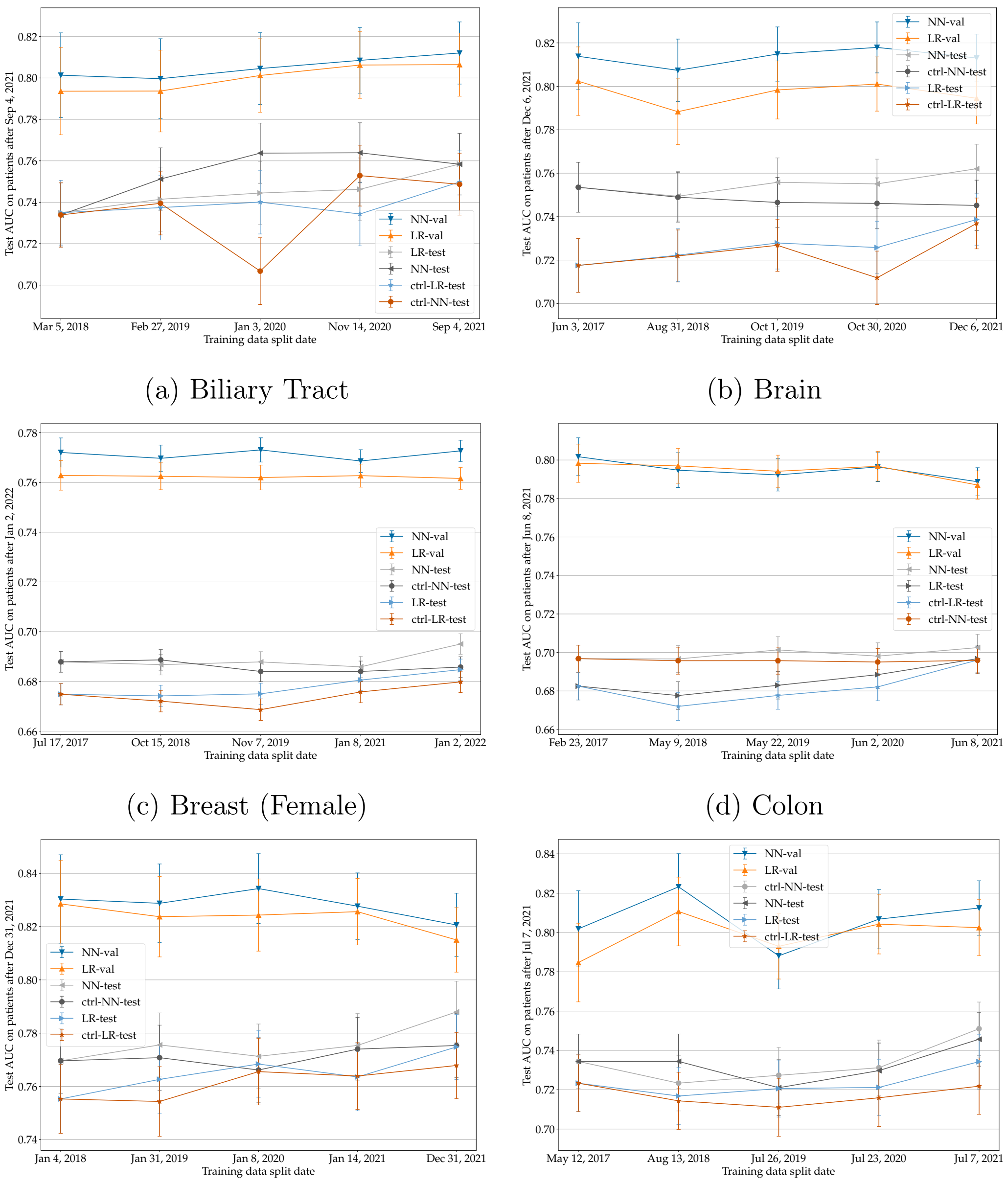


Figure B-6: External validation by race for unified models



(a) Biliary Tract

(b) Brain

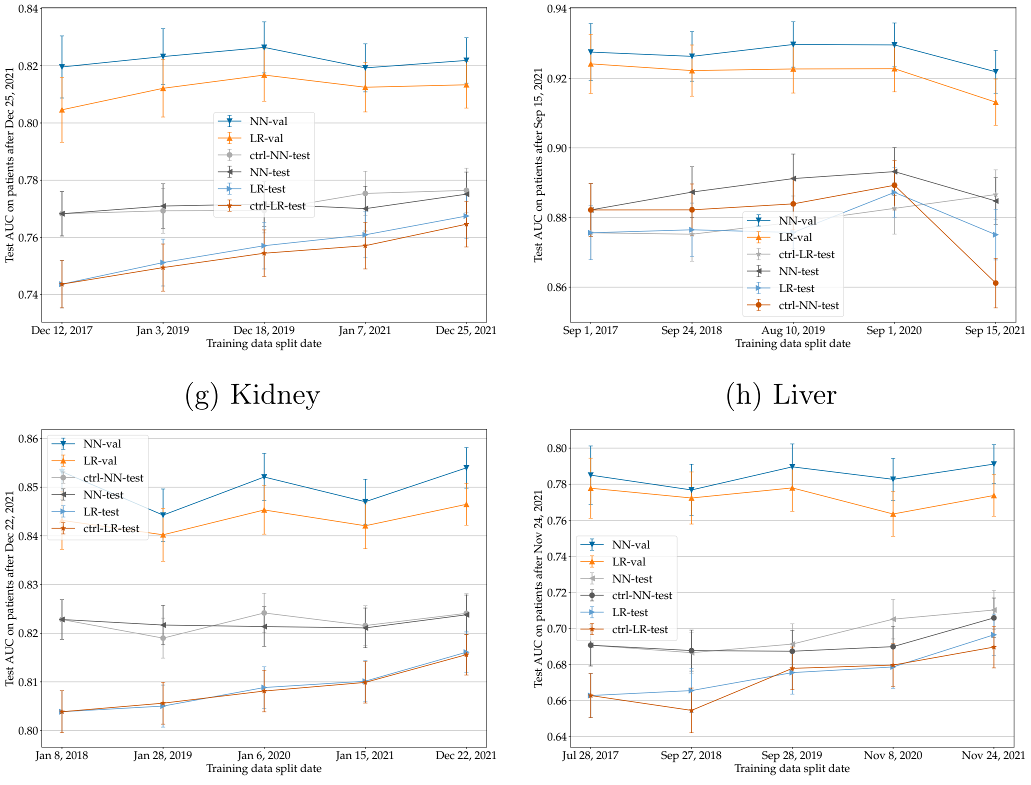
(c) Breast (Female)

(d) Colon

(e) Esophageal

(f) Gastric

Figure B-7: Temporal validation for singular models



(g) Kidney

(h) Liver

(i) Lung

(j) Ovarian

Figure B-7: Temporal validation for singular models (continued)

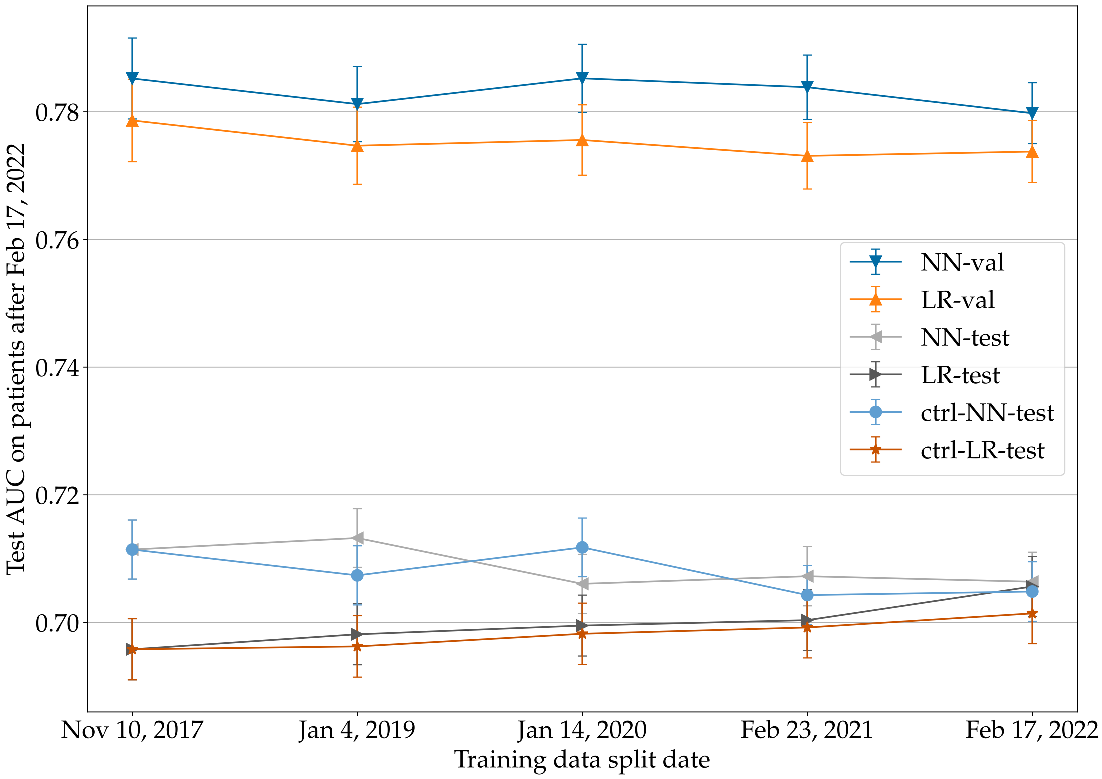
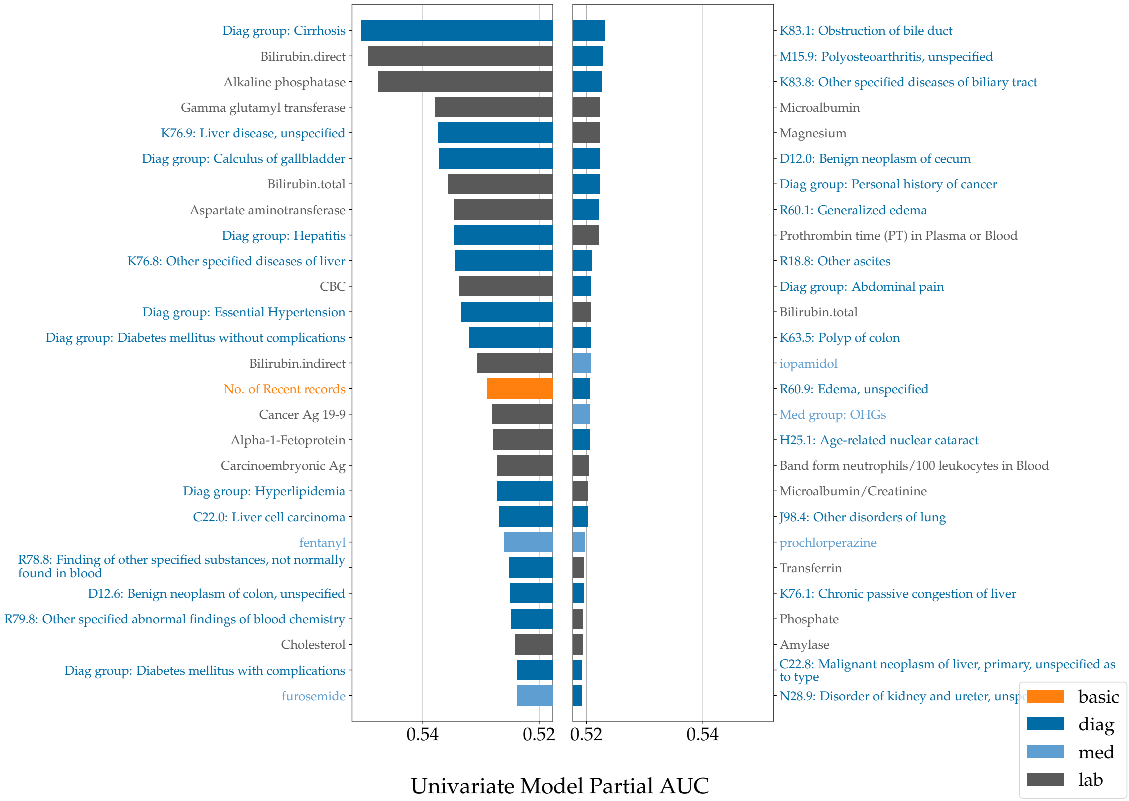
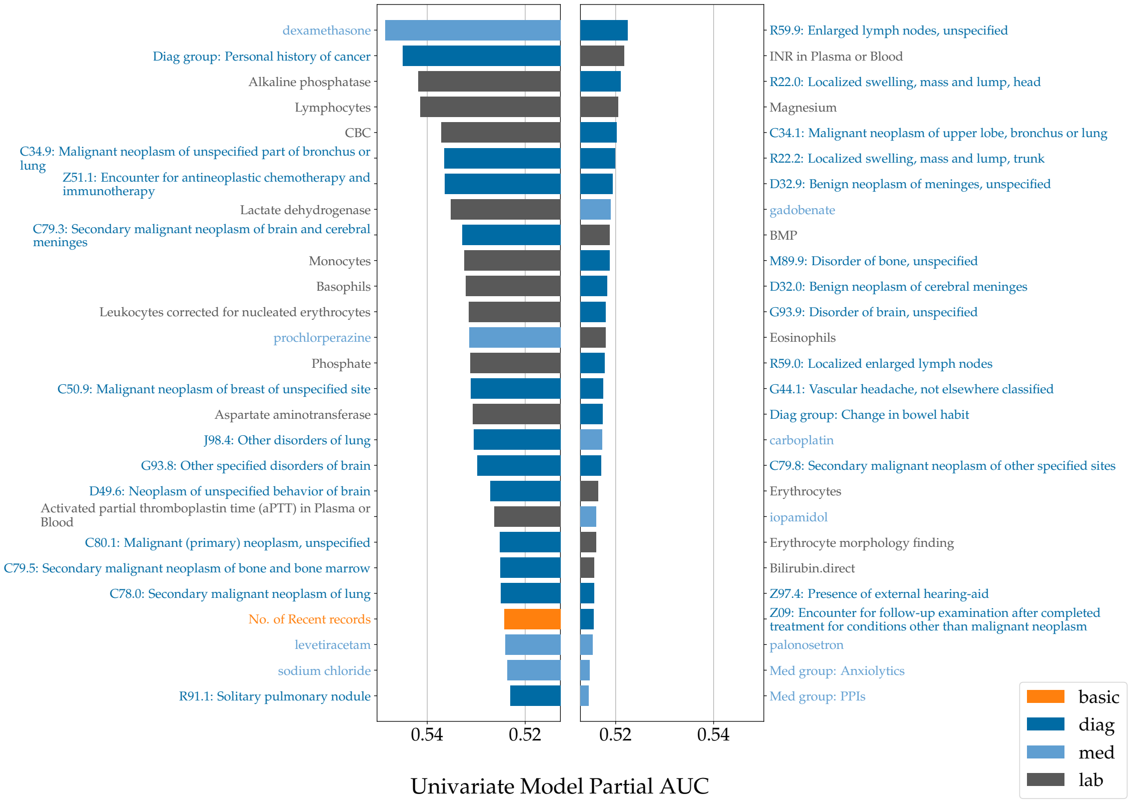


Figure B-8: Temporal validation for unified models



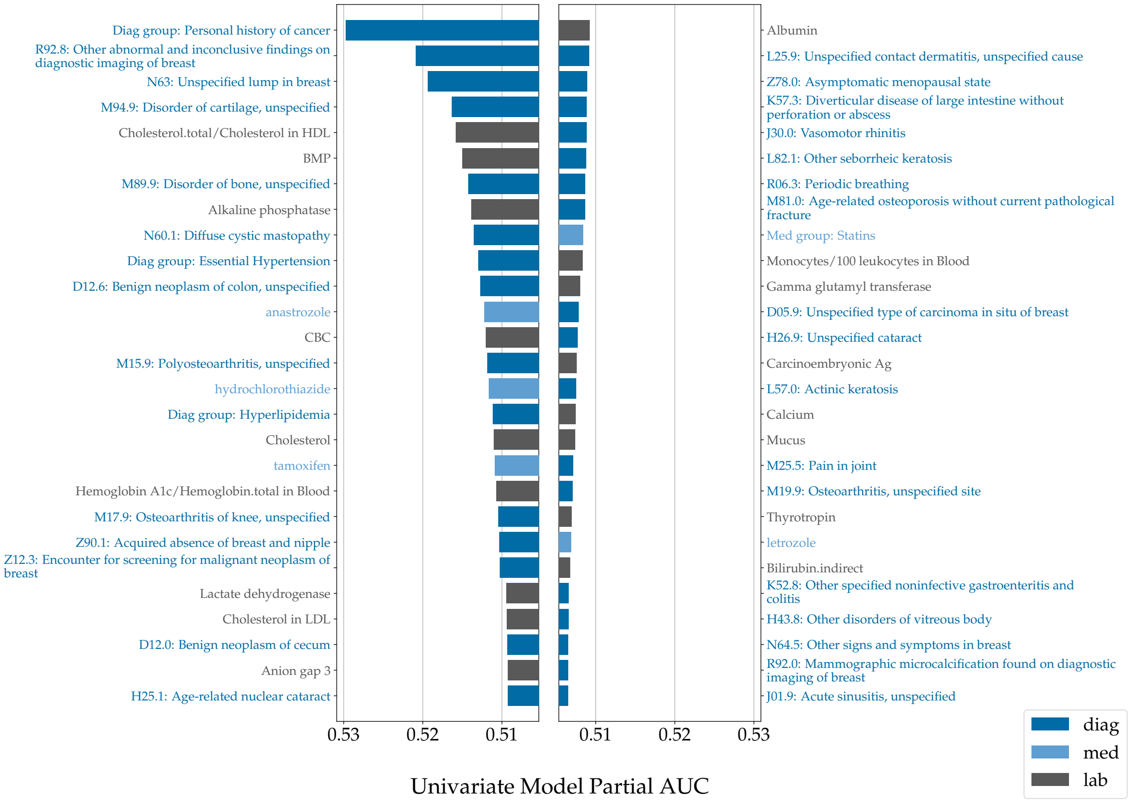
(a) Biliary Tract

Figure B-9: Top predictive features for singular NN models



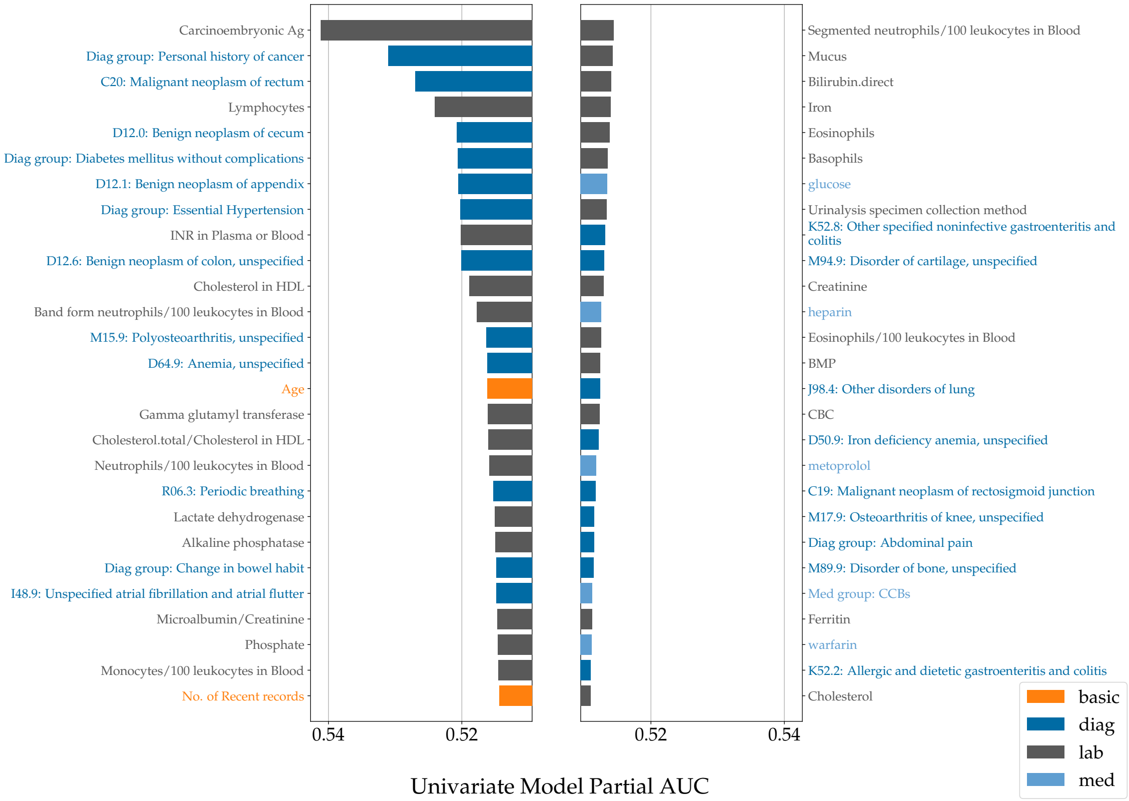
(b) Brain

Figure B-9: Top predictive features for singular NN models (continued)



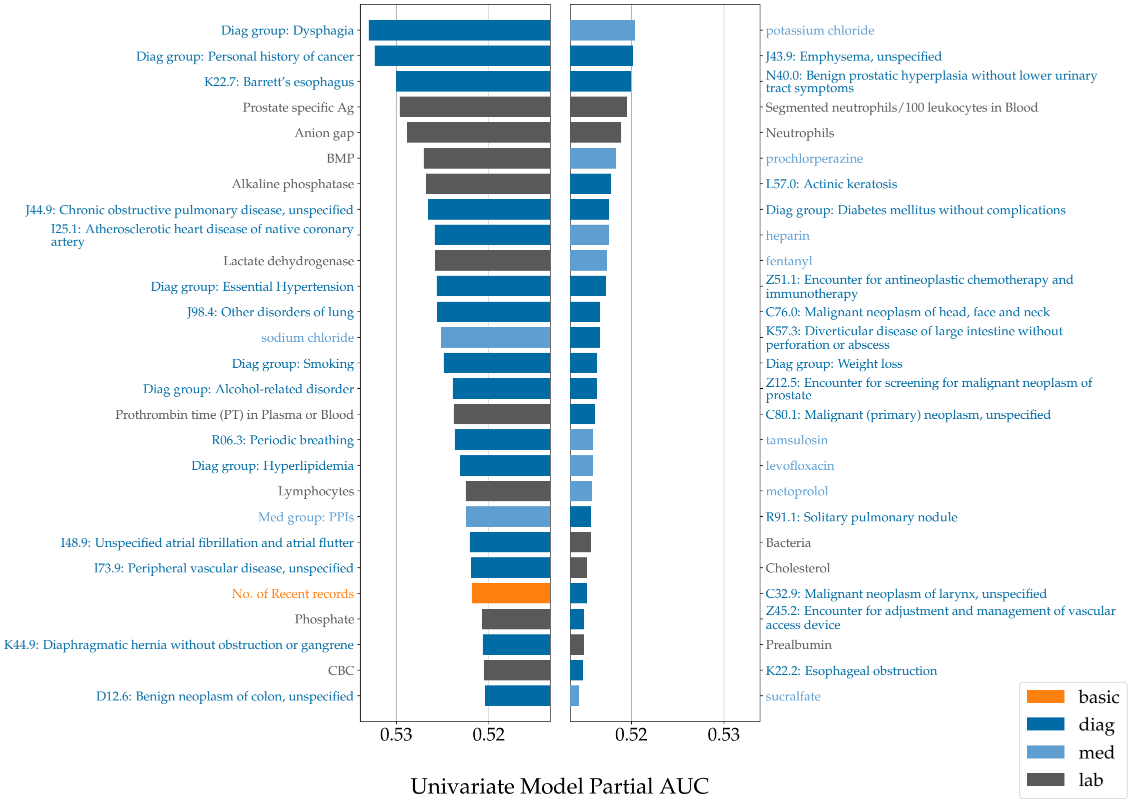
(c) Breast (Female)

Figure B-9: Top predictive features for singular NN models (continued)



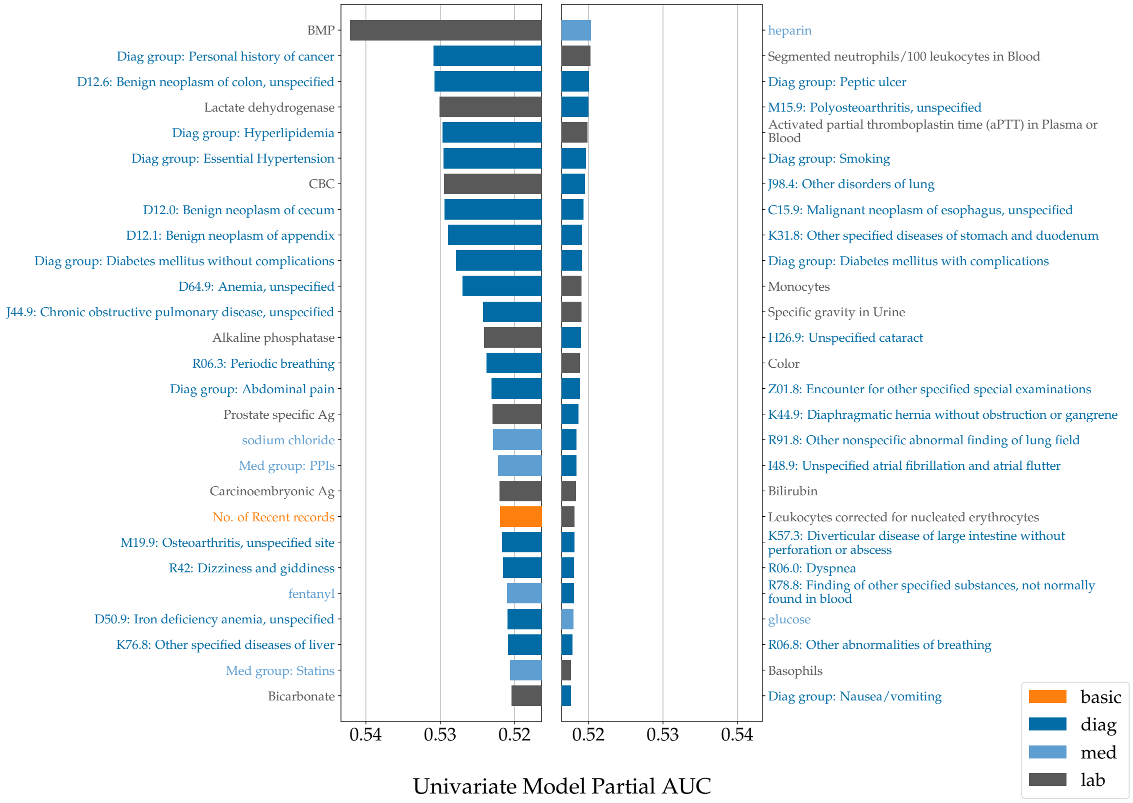
(d) Colon

Figure B-9: Top predictive features for singular NN models (continued)



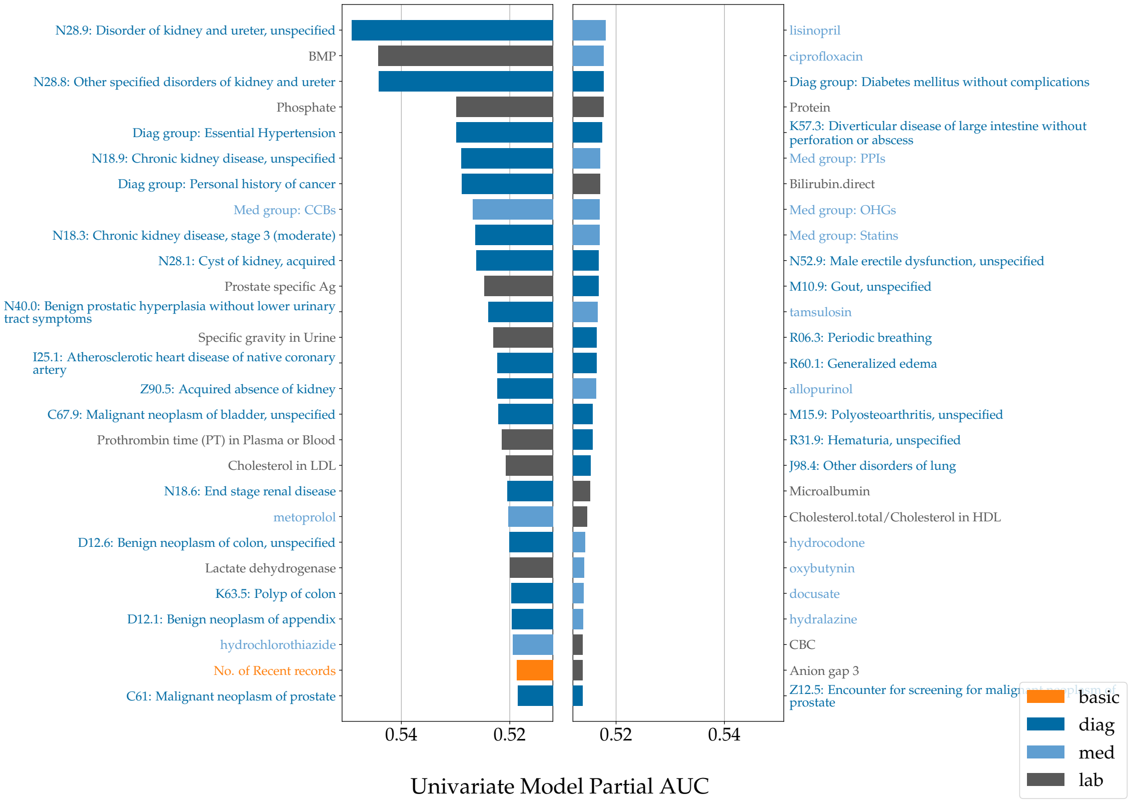
(e) Esophageal

Figure B-9: Top predictive features for singular NN models (continued)



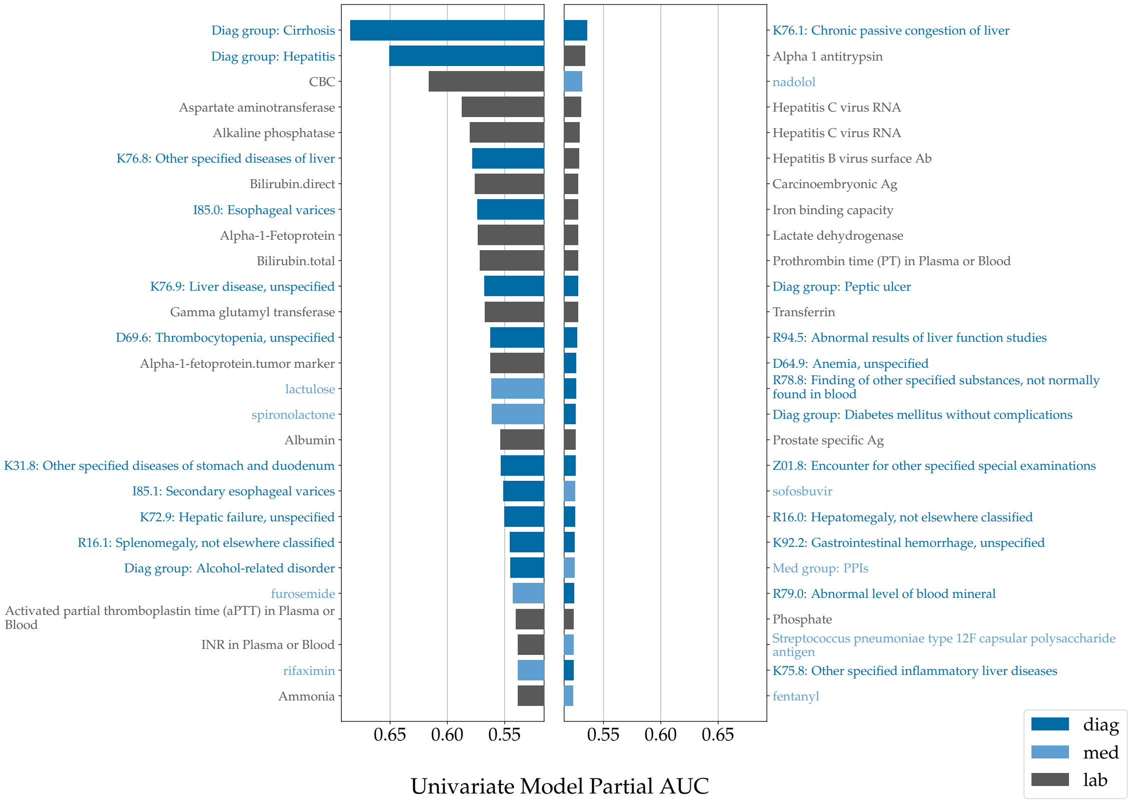
(f) Gastric

Figure B-9: Top predictive features for singular NN models (continued)



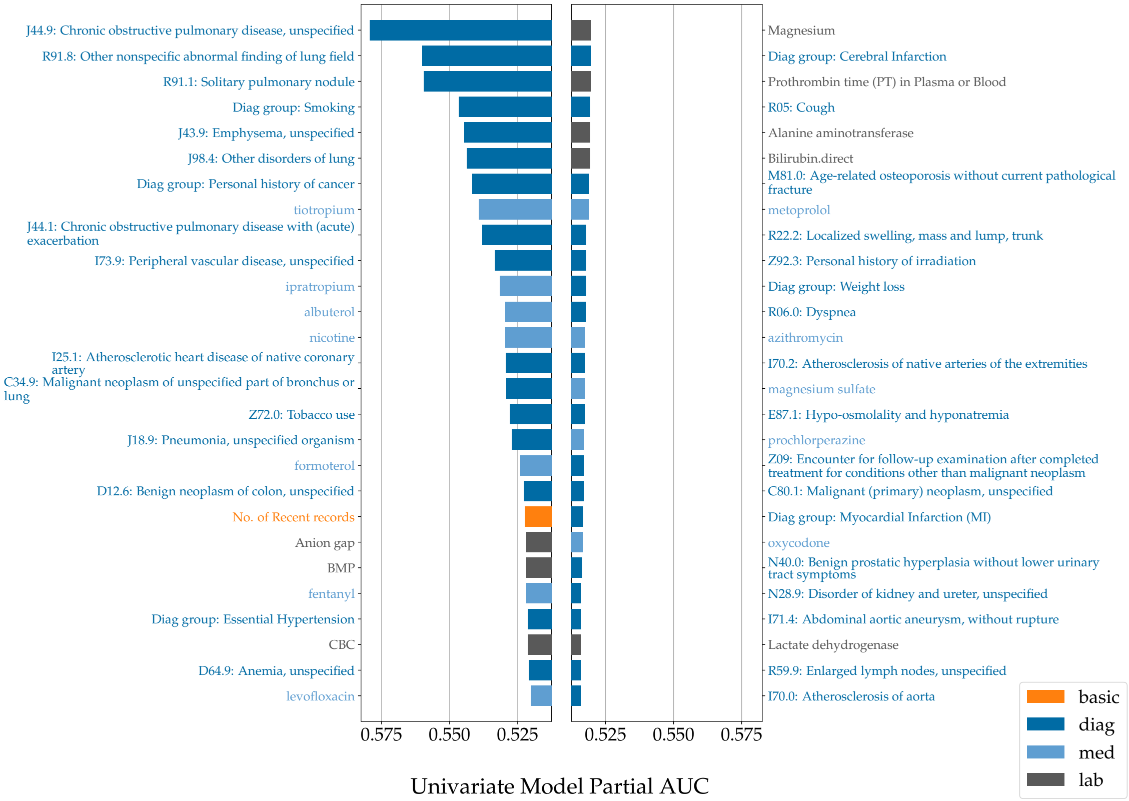
(g) Kidney

Figure B-9: Top predictive features for singular NN models (continued)



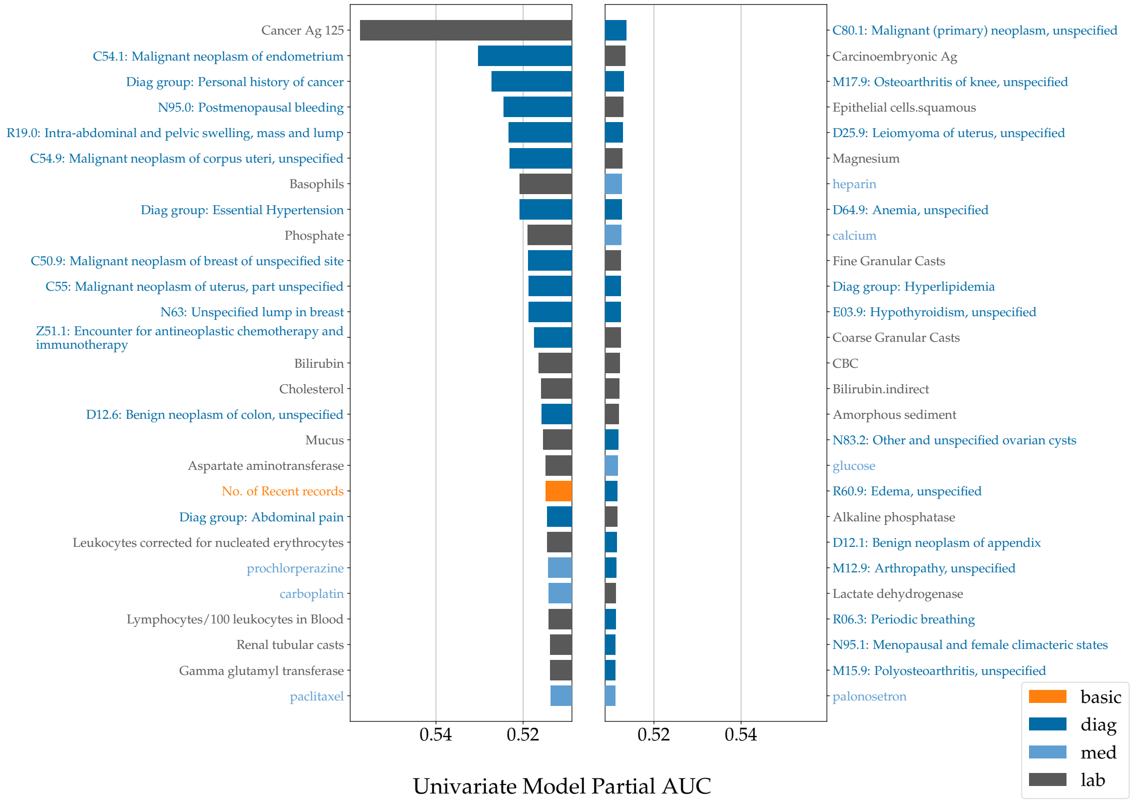
(h) Liver

Figure B-9: Top predictive features for singular NN models (continued)



(i) Lung

Figure B-9: Top predictive features for singular NN models (continued)



(j) Ovarian

Figure B-9: Top predictive features for singular NN models (continued)

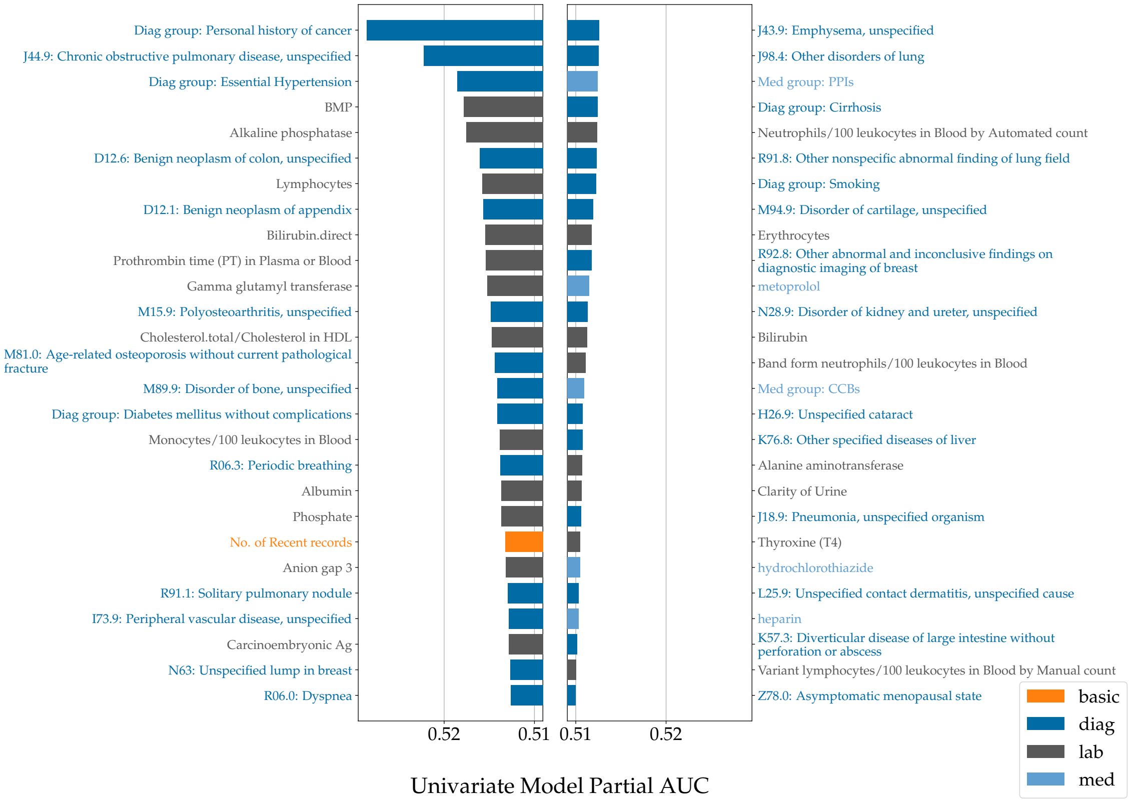


Figure B-10: Top predictive features for the unified NN model

Bibliography

[1] Jennifer Anne Cooper, Ronan Ryan, Nick Parsons, Chris Stinton, Tom Marshall, and Sian Taylor-Phillips. The use of electronic healthcare records for colorectal cancer screening referral decisions and risk prediction model development. *BMC Gastroenterology*, 20(1), March 2020.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives, 2014.

[3] Giovanna Fattovich, Tommaso Stroffolini, Irene Zagni, and Francesco Donato. Hepatocellular carcinoma in cirrhosis: Incidence and risk factors, Nov 2004.

[4] Yuting Han, Xia Zhu, Yizhen Hu, Canqing Yu, Yu Guo, Dong Hang, Yuanjie Pang, Pei Pei, Hongxia Ma, Dianjianyi Sun, Ling Yang, Yiping Chen, Huaidong Du, Min Yu, Junshi Chen, Zhengming Chen, Dezheng Huo, Guangfu Jin, Jun Lv, Zhibin Hu, Hongbing Shen, and Liming Li. Electronic health record–based absolute risk prediction model for esophageal cancer in the chinese population: Model development and external validation. *JMIR Public Health and Surveillance*, 9:e43725, March 2023.

[5] Timothy P Hanna, Will D King, Stephane Thibodeau, Matthew Jalink, Gregory A Paulin, Elizabeth Harvey-Jones, Dylan E O'Sullivan, Christopher M Booth, Richard Sullivan, and Ajay Aggarwal. Mortality due to cancer treatment delay: systematic review and meta-analysis. *BMJ*, 371, 2020.

[6] Robert J. Huang, Nicole Sung-Eun Kwon, Yutaka Tomizawa, Alyssa Y. Choi, Tina Hernandez-Boussard, and Joo Ha Hwang. A comparison of logistic regression against machine learning algorithms for gastric cancer risk prediction within real-world clinical data streams. *JCO Clinical Cancer Informatics*, (6), June 2022.

[7] Kai Jia, Steven Kundrot, Matvey Palchuk, Jeff Warnick, Kathryn Haapala, Irving Kaplan, Martin Rinard, and Limor Appelbaum. Developing and validating a pancreatic cancer risk model for the general population using multi-institutional electronic health records from a federated network. *medRxiv*, 2023.

[8] Kai Jia and Martin C. Rinard. Efficient exact verification of binarized neural networks. *CoRR*, abs/2005.03597, 2020.

[9] Chia-Wei Liang, Hsuan-Chia Yang, Md Mohaimenul Islam, Phung Anh Alex Nguyen, Yi-Ting Feng, Ze Yu Hou, Chih-Wei Huang, Tahmina Nasrin Poly, and Yu-Chuan Jack Li. Predicting hepatocellular carcinoma with minimal features from electronic health records: Development of a deep learning model. *JMIR Cancer*, 7(4):e19812, October 2021.

[10] Davide Placido, Bo Yuan, Jessica X. Hjaltelin, Chunlei Zheng, Amalie D. Haue, Piotr J. Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, Alexander Chowdhury, Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S. Marks, Aviv Regev, Siamack Ayandeh, Mary T. Brophy, Nhan V. Do, Peter Kraft, Brian M. Wolpin, Michael H. Rosenthal, Nathanael R. Fillmore, Søren Brunak, and Chris Sander. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories, May 2023.

[11] Yirong Wu, Elizabeth S. Burnside, Jennifer Cox, Jun Fan, Ming Yuan, Jie Yin, Peggy Peissig, Alexander Cobian, David Page, and Mark Craven. Breast cancer risk prediction using electronic health records. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 224–228, 2017.

[12] Marvin Chia-Han Yeh, Yu-Hsiang Wang, Hsuan-Chia Yang, Kuan-Jen Bai, Hsiao-Han Wang, and Yu-Chuan Jack Li. Artificial intelligence–based prediction of lung cancer risk using nonimaging electronic medical records: Deep learning approach. *Journal of Medical Internet Research*, 23(8):e26256, August 2021.