Possibilities of 5 primary cancers

Objective

- 건강 보험 공단의 건강검진코호트 DB (NHIS-50) 를 활용하여, 5대 암(stomach, liver, colorectal, breast, lung)에 대한 risk prediction model

- 예측 시점으로부터 과거 10년 이내 건강 검진 데이터 사용

- 과거 특정 질병의 발병 여부 고려

Existing studies - Liver cancer

- https://www.sciencedirect.com/science/article/pii/S0091743512002174?via%3Dihub
  - 10-year risk prediction model for HCC
  - 17,654 Japanese aged 40 to 69 years (1993-2006)
  - Cox proportional hazards regression
  - age, sex, alcohol consumption, BMI, diabetes, coffee consumption, and hepatitis B and C virus infection
  - Developed a simple risk scoring system (score range: − 1 to 19)
- https://www.ncbi.nlm.nih.gov/pubmed/23073549
  - 428 584 subjects from a private health screening firm in Taiwan (1994-2008)
  - age, sex, health history-related variables; HBV or HCV infection-related variables; serum levels of alanine transaminase (ALT), aspartate transaminase (AST), and alfa-fetoprotein (AFP), as well as other variables of routine blood panels for liver function
  - medical history (such as diabetes, hypertension, stroke, heart diseases)
  - Diagnosed with diabetes or currently taking diabetes medication were defined as having diabetes
  - Smoking was classified by the number of pack-years (ie, daily cigarette quantity × duration in years)
  - “regular drinkers” (those who consumed ≥2 drinks/day on ≥3 days/week) and “occasional drinkers” (those who consumed <2 drinks/day on <3 days/week)
  - Cox proportional hazards regression
  - **AUC = 0.933, 95% CI = 0.929 to 0.949 (해당 논문은 간기능 관련 여러 테스트 결과를 feature 로 이용)**

참고: https://epi.grants.cancer.gov/cancer_risk_prediction/

Existing studies - Breast cancer

- https://www.ncbi.nlm.nih.gov/pubmed/21562243
  - 589 women with breast cancer (case patients) and 952 women without breast cancer (control subjects) in the Asian American Breast Cancer Study
  - age, number of affected mothers, sisters, and daughters, and number of previous benign biopsies
- https://www.ncbi.nlm.nih.gov/pubmed/18042936
  - 1647 African American control subjects
  - logistic regression risk models
  - age, breast density, family history of breast cancer, and a prior breast procedure
  - age, breast density, race, ethnicity, family history of breast cancer, a prior breast procedure, body mass index, natural menopause, hormone therapy, and a prior false-positive mammogram
  - **C-statistics: 0.631**

Existing studies - Colorectal cancer

- https://www.ncbi.nlm.nih.gov/pubmed/24533067
  - Gender specific five-year risk prediction models
  - 846,559 men and 479,449 women
  - Examinees were 30-80 years old and free of cancer in the baseline years of 1996 and 1997
  - Age, body mass index, serum cholesterol, family history of cancer, and alcohol consumption for MEN
  - age, height, and meat intake frequency for women
  - **C-statistics between 0.69 and 0.78**
- https://www.ncbi.nlm.nih.gov/pubmed/24653621
  - Totally 905 cases
  - age, gender, coronary heart disease, egg intake and stool frequency
  - **AUC: 0.75**
- https://www.ncbi.nlm.nih.gov/pubmed/24385598
  - age, sex, family history of colorectal cancer, cigarette smoking ($p<0.001$ for these four factors), and Body Mass Index ($p=0.033$)
  - **c-statistic 0.62**

Dataset: 건강검진 코호트 (NHIS-50)

- 코호트 설명
  - 2002년 자격유지자 중 2002~2003년 40~79세 일반건강검진 수검자, 약 51만명 데이터
  - 이 중 자격, 건강검진 그리고 진료 DB 명세서 테이블을 참조
  - DB 매뉴얼: https://nhiss.nhis.or.kr/bd/ab/bdaba006cv.do
- 진단 코드 맵핑 (참고: http://www.koicd.kr/2016/main.do)

질병	진단 코드
liver cancer	C22
stomach cancer	C16
colorectal cancer	C18, C19, C20
breast cancer	C50
lung cancer	C34
cervical cancer	C53
pancreatic cancer	C25

Data preparation

- Subject 선정 과정
  - 2002년 자격유지자 중 2002~2003년 40~79세 일반건강검진 수검자
    - 514,866명
  - 2002년 부터 2013년 또는 사망일 때까지 자격이 유지된 사람
    - 506,496명
  - 건강 검진 기록이 있는 사람
    - 155,912 명
  - 2002-03년 동안 (wash-out period) 고려되는 암의 진단을 받은 환자는 제외
    - 152,560 명

실험 대상에서 연도별 암 진단 기록 수

Year	LIVER	STOMACH	COLORECTAL	BREAST	LUNG	CERVICAL	PANCREATIC
2002	218	602	375	183	261	118	35
2003	364	872	568	216	472	141	79
2004	399	961	663	247	564	129	78
2005	456	1101	825	286	619	137	107
2006	446	1170	901	308	630	135	101
2007	465	1279	1033	351	673	135	123
2008	494	1423	1128	380	657	134	149
2009	491	1473	1229	417	681	134	140
2010	507	1556	1264	457	749	137	154
2011	485	1587	1281	461	744	122	130
2012	519	1601	1309	480	797	107	160
2013	565	1569	1327	475	845	112	151

선정된 **subject**에서 암 환자 수

	LIVER	STOMACH	COLORECTAL	BREAST	LUNG	CERVICAL	PANCREATIC
#Free	150612	148941	149344	73746	149394	74164	151786
#Diagnosed	1948	3619	3216	693	3166	275	774

건강 검진DB로부터 29개의 features 추출

Index	항목	설명	특이사항
0	AGE	나이	자격 DB와 조인해서, 검진 받을 당시 나이로 변환
1	SEX	성별	(1: 남자, 2: 여자)
2	BMI	Body Mass Index	100만 코흐트에는 존재 안함 없는 경우 HEIGHT, WEIGHT 컬럼으로 부터 계산
3	BP_HIGH	수축기혈압	
4	BP_LWST	이완기혈압	
5	BLDS	식전혈당(공복혈당)	
6	TOT_CHOLE	총콜레스테롤	
7	GAMMA_GTP	감마지티피	
8	HMG	혈색소	
9	OLIG_PROTE_CD	요단백	
10	SGOT_AST	(혈청지오티)AST	
11	SGPT_ALT	(혈청지피티)ALT	

Index	항목	설명	특이사항
12	FMLY_APOP_PATIEN_YN	(가족력)뇌졸중유무	2002-08: (1: 미해당, 2: 해당) 2009-13: (0: 미해당, 1: 해당) (0: 미해당, 1: 해당) 방식으로 코드 통일
13	FMLY_CANCER_PATIEN_YN	(가족력)암유무	
14	FMLY_DIABML_PATIEN_YN	(가족력)당뇨병유무	
15	FMLY_HDISE_PATIEN_YN	(가족력)심장병유무	
16	FMLY_HPRTS_PATIEN_YN	(가족력)고혈압유무	
17	HCHK_APOP_PMH_YN	(본인)뇌졸중과거병력유무	2002-08: 총 9가지 질병에 대하여, 3가지 질병까지 기록 ('HCHK_PMH_CD1', 'HCHK_PMH_CD2', 'HCHK_PMH_CD3') 2009-13: 7가지 질환에 대한 발병 여부 (binary) 컬럼들이 존재 Integration 공통적으로 기록이 되는 6가지 질환에 대하여 발병 여부를 표현하는 컬럼으로 통합 즉, 02-08년도 데이터의 위 세컬럼에서 각 질병이 언급되어있는지 여부에 따라 binary 값 생성
18	HCHK_DIABML_PMH_YN	(본인)당뇨병과거병력유무	
19	HCHK_ETCDSE_PMH_YN	(본인)기타(암포함)질환 과거병력유무	
20	HCHK_HDISE_PMH_YN	(본인)심장병과거병력유무	
21	HCHK_HPRTS_PMH_YN	(본인)고혈압과거병력유무	
22	HCHK_PHSS_PMH_YN	(본인)폐결핵과거병력유무	

Index	항목	설명	특이사항
23	DRNK_HABIT_RSPS_CD	음주습관	2002:-08: 1 : (거의)마시지 않는다 2 : 월2~3회정도 마신다 3 : 일주일에 1~2회 마신다 4 : 일주일에 3~4회 마신다 5 : 거의 매일 마신다 2009-13: 일주일에 평균적으로 마시는 횟수 0-8 일 값 Integration 02-08년도 categorical 값들은 다음과 같이 주당 평균 횟수로 변환 (1: 0, 2: 0.5, 3: 1.5, 4: 3.5, 5: 6.5)
24	TM1_DRKQTY_RSPS_CD	1회 음주량	2002-08: 1 : 소주 반 병 이하 2 : 소주 한 병 3 : 소주 1병 반 4 : 소주 2병 이상 2009-13: "잔' 단위 numerical value Integration 02-08년도 categorical 값들은 다음과 같이 잔 수로 변환 (1: 3, 2: 7, 3: 10.5, 4: 15)

Index	항목	설명	특이사항
25	SMK_STAT_TYPE_RSPS_CD	흡연상태	1 : 피우지 않는다, 2 : 과거에 피웠으나 지금은 끊었다, 3 : 현재도 피운다
26	SMK_TERM_RSPS_CD	(과거,현재)흡연기간	2002-08: SMK_TERM_RSPS_CD 한개의 컬럼으로 categorical variable (1 : 5년 미만, 2 : 5~9년, 3 : 10~19년, 4 : 20~29년, 5 : 30년 이상) 2009-13: PAST_DSQTY_RSPS_CD (과거 흡연기간), CUR_SMK_TERM_RSPS_CD (현재 흡연기간) 두개의 컬럼으로 존재하며, "년" 단위 numerical variable Integration 1) 09-13 데이터의 경우 SMK_STAT_TYPE_RSPS_CD 컬럼을 이용하여 SMK_TERM_RSPS_CD 하나로 통합 2) 02-08 데이터의 categorical 값들은 09-13 과 같은 "년" 단위 numerical 값으로 변경. 09-13 년도 데이터로부터 각 카테고리별 mean, std를 구하여 normal distribution 에서 radom sampling
27	CUR_DSQTY_RSPS_CD	(현재)하루흡연량	2002-08: (1 : 반갑미만, 2 : 반갑~한갑미만, 3 : 한갑~두갑미만, 4 : 두갑이상) 2009-13: (개피) 개피 형식으로 통일. 09-13 년도 데이터로부터 각 카테고리별 mean, std를 구하여 normal distribution 에서 radom sampling

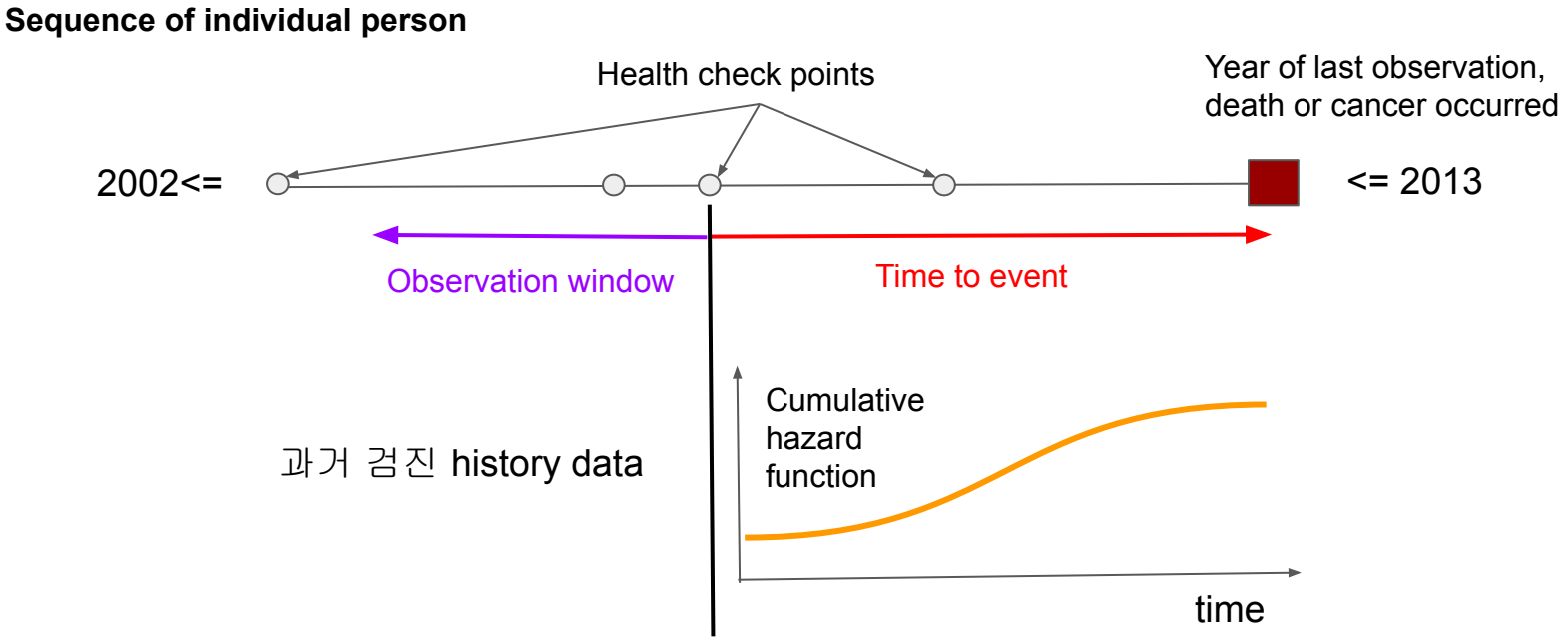
Index	항목	설명	특이사항
28	EXERCI_FREQ_RSPS_CD	1주 운동횟수	2002-08: categorical variable (1 : 안한다, 2 : 1~2회, 3 : 3~4회, 4 : 5~6회, 5 : 거의 매일) 2009-13: 다음 세 컬럼에 대하여 "일" 단위의 numerical variable 1주_20분이상 격렬한 운동(MOV20_WEK_FREQ_ID) 1주_30분이상 중간정도 운동(MOV30_WEK_FREQ_ID) 1주_총30분이상 걷기 운동(WLK30_WEK_FREQ_ID) Integration 09-13 데이터를 02-08과 같은 형식으로 변환. 세 컬럼 합이 0 -> 1 두컬럼 ('MOV20_WEK_FREQ_ID','MOV30_WEK_FREQ_ID') 합이 1-2 -> 2 3-4 -> 3 5-6 -> 4 7이상 -> 5

Approaches

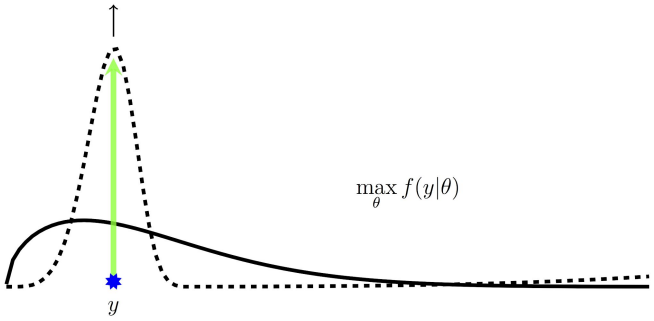
- Survival regression
  - Cox’s proportional hazard model
    - https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html#cox-s-proportional-hazard-model
  - RNN-based model using Weibull distribution
    - https://github.com/ragulpr/wtte-rnn
- Binary classification
  - Logistic regression
  - Random forest
  - Tree-based gradient boosting (LightGBM)
  - RNN-based model

Survival regression

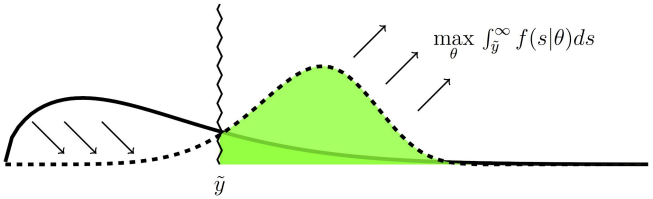
- 예측 시점으로 부터 과거 검진 기록을 이용하여, 시간에 따른 암의 발병 위험도 (hazard function)를 예측



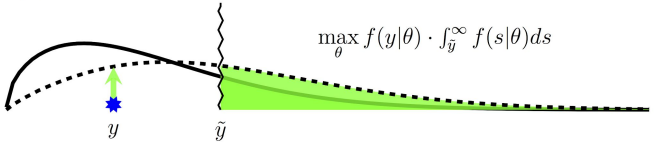
Censored data를 이용한 학습



(a) Uncensored observation: Push the pdf up at event



(b) Censored observation: Push mass over the point of censoring



(c) Uncensored and censored observation: Compromise

(a) 암이 발병한 sequence의 경우엔, 암이 발병한 시점에 risk 가 최대값을 같도록 학습

(b) 마지막 관찰 시점까지 암 발병과 연관이 없는 경우, 마지막 관찰 시점 이후에 risk 가 높게 분포되도록 학습

(c) Objective function

$$\sum_{n=1}^{N}\sum_{t=0}^{T_n} u_t^n \cdot \log\left[\Pr(Y_t^n = y_t^n | x_{0:t}^n)\right] + (1-u_t^n) \cdot \log\left[\Pr(Y_t^n > y_t^n | x_{0:t}^n)\right]$$

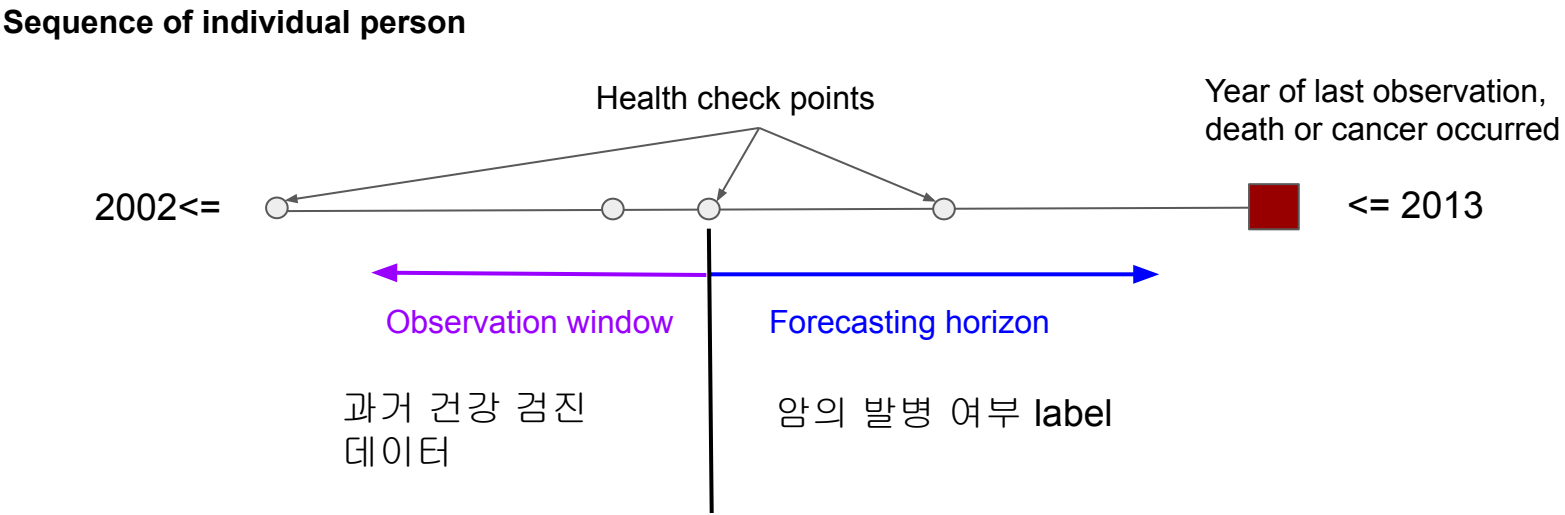
https://ragulpr.github.io/2016/12/22/WTTE-RNN-Hackless-churn-modeling/

Performance of survival regression models

		Stomach	Liver	Colorectal	Breast	Lung
Survival regression (C-statistics)	Cox	0.6985	0.7612	**0.6939**	**0.6113**	0.7629
	RNN (window=5)	0.6901	**0.7849**	0.6794	0.5679	**0.7671**
	RNN (window=10)	**0.7685**	0.6902	0.6409	0.4105	0.7658

Binary classification

- 예측 시점으로 부터 과거 검진 기록을 이용하여, 향후 N년 동안 (forecasting horizon) 암의 발병 여부 (binary label)을 예측



Performance of binary classification models

		Stomach	Liver	Colorectal	Breast	Lung
Binary classification (ROC AUC)	Logistic	0.7550	0.8288	0.7565	**0.6760**	0.8064
	LGB	**0.8000**	**0.8649**	**0.7978**	0.6661	**0.8437**
	RF	0.7300	0.7779	0.7463	0.6558	0.7687
	RNN	0.7997	0.7755	0.7764	0.6145	0.8231

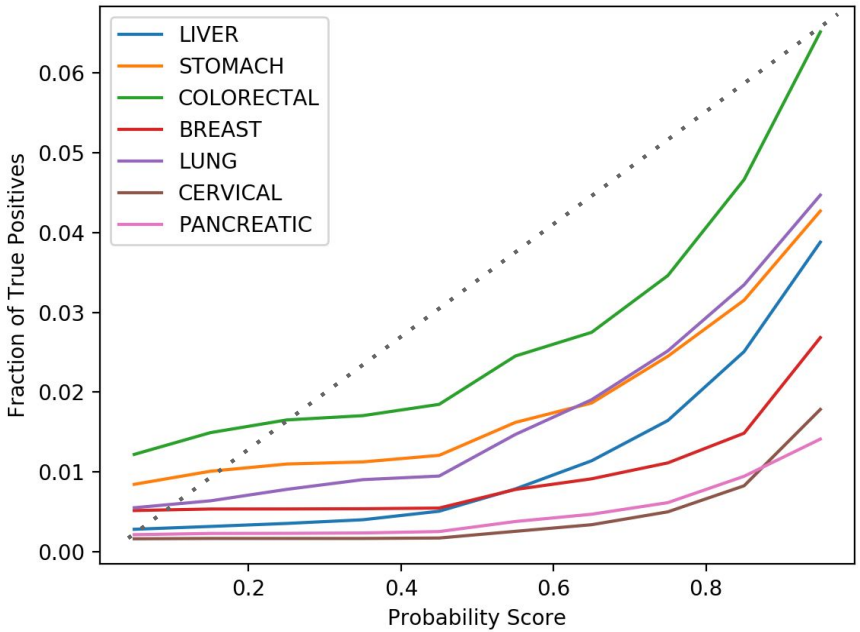
Observation window & Attention mechanism

	observation window	**Stomach**	**Liver**	**Colorectal**	**Breast**	**Lung**
RNN	3	0.7708	0.7565	0.7655	0.5850	0.8028
	5	0.7848	0.7740	0.7667	0.5004	0.7848
	10	**0.7997**	0.7755	0.7764	**0.6145**	**0.8231**
RNN+Attention	10	0.7612	**0.7833**	**0.7795**	0.5211	0.8190

진료 DB로부터 features 추출

- 검진 DB 외에 추가로 진료 DB를 활용
- 암 진단을 받은 subject들 (cancer group)과 암이 발병하지 않은 subject들 (free group)을 나누어 암 발병 전에 자주 진단되는 질병을 risk factor 로 이용
    - 예) 대장암(colorectal cancer) 의 경우 결장 폴립 관련 진단 (D120-129, K635)이 암 발병하기 전에 자주 진단되는 것으로 알려짐.
  - Cancer group 내 1% 이상의 환자들에게서 진단된 진단 코드중, free group에 비해 진단 비율이 높은 상위 20개의 진단 코드를 학습에 이용

Model의 Probability Score 와 Precision 의 관계

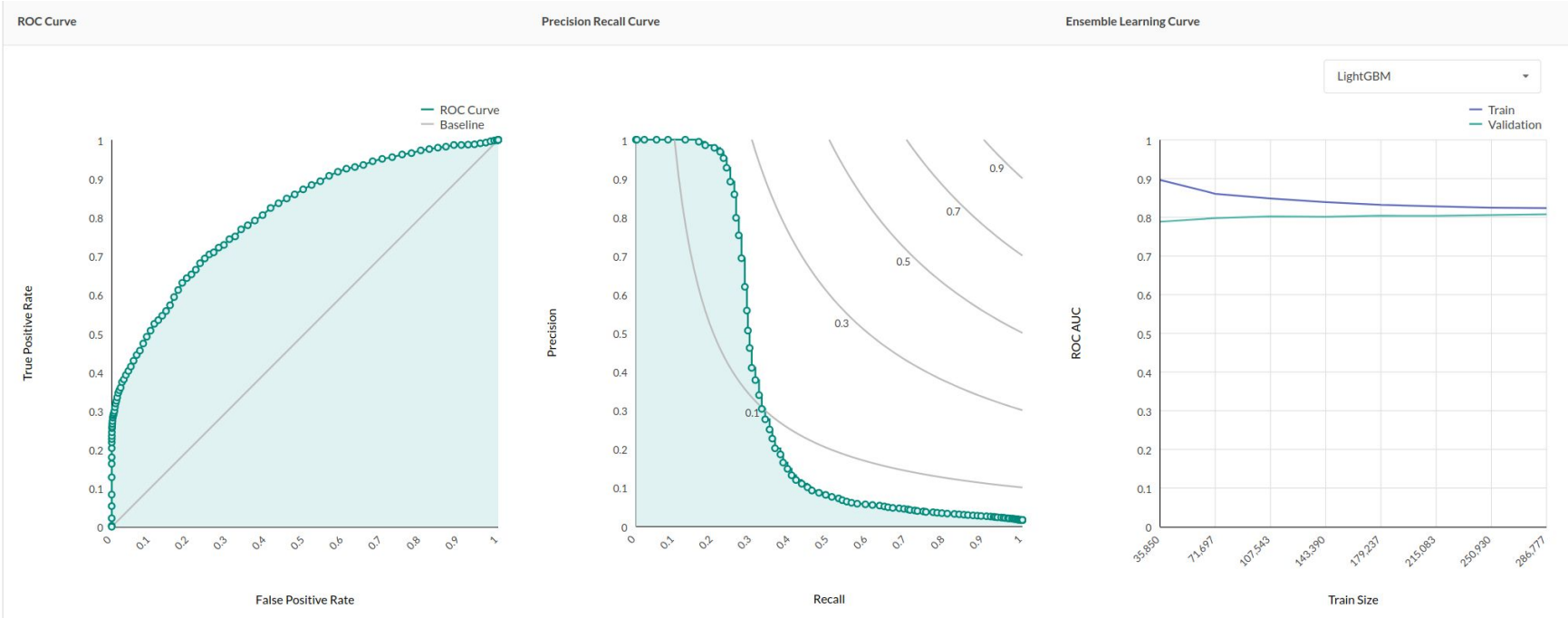


- 현실적으로 precision 이 낮은 모델이기 때문에, risk score 가 실제 cancer 발병 확률과 맵핑 되기는 어려움
- Score calibration 을 통해서 score 와 positive fraction이 linear relationship을 갖도록 score calibration 가능

Summary

- Survival regression 보다는 classification 으로의 모델링이 좀 더 효과적
- 시계열의 검진 데이터를 효과적으로 학습하기 위하여 RNN-based model 을 시도해보았으나, 성능이 LightGBM에 비해 비슷하거나 낮음
- 진료 DB 로 부터 암 발병과 연관성이 있어보이는 질병 feature를 활용해 보았으나, 성능에 큰 차이를 볼 수 없었음
- 가족력 / 본인 과거 병력을 제외한 기본 feature set으로 학습하여도 성능이 크게 떨어지지 않음
- 전반적으로, 기존의 유사 연구 결과들과 비슷하거나 좀 더 나은 성능을 보여줌

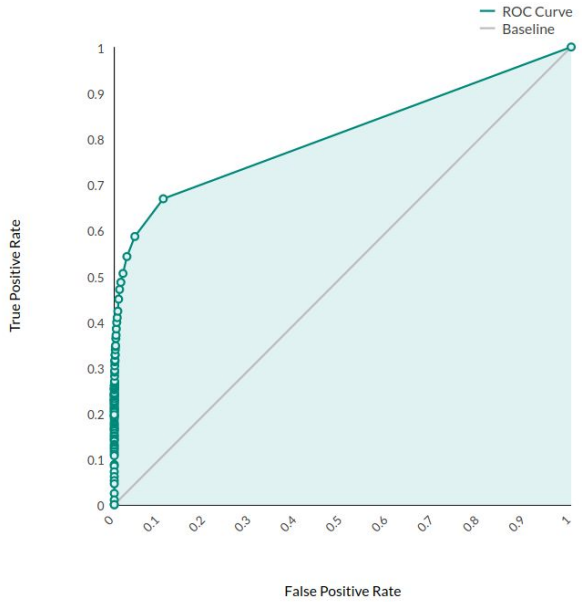
Stomach cancer



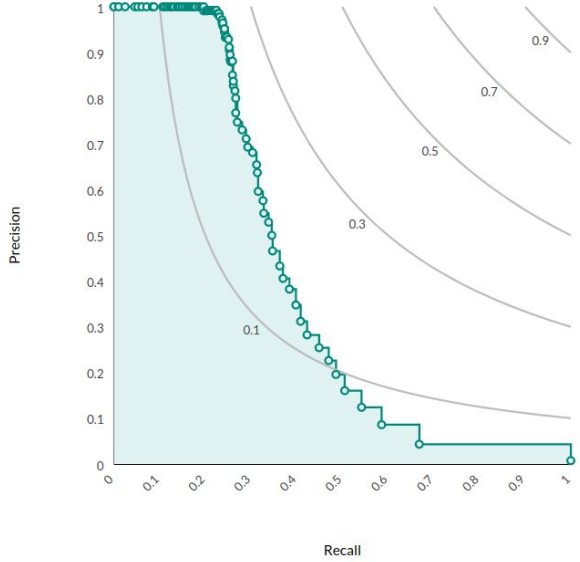
Note: Precision-Recall 차트에서 등고선은 F1 score 을 나타냄.

Liver cancer

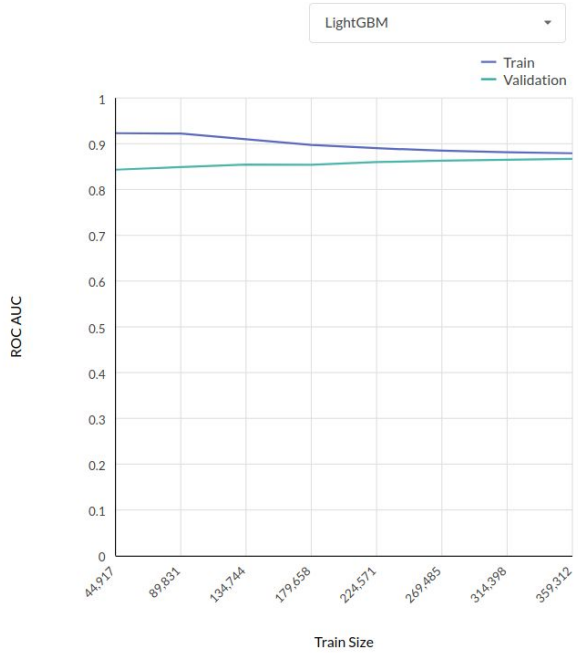
ROC Curve



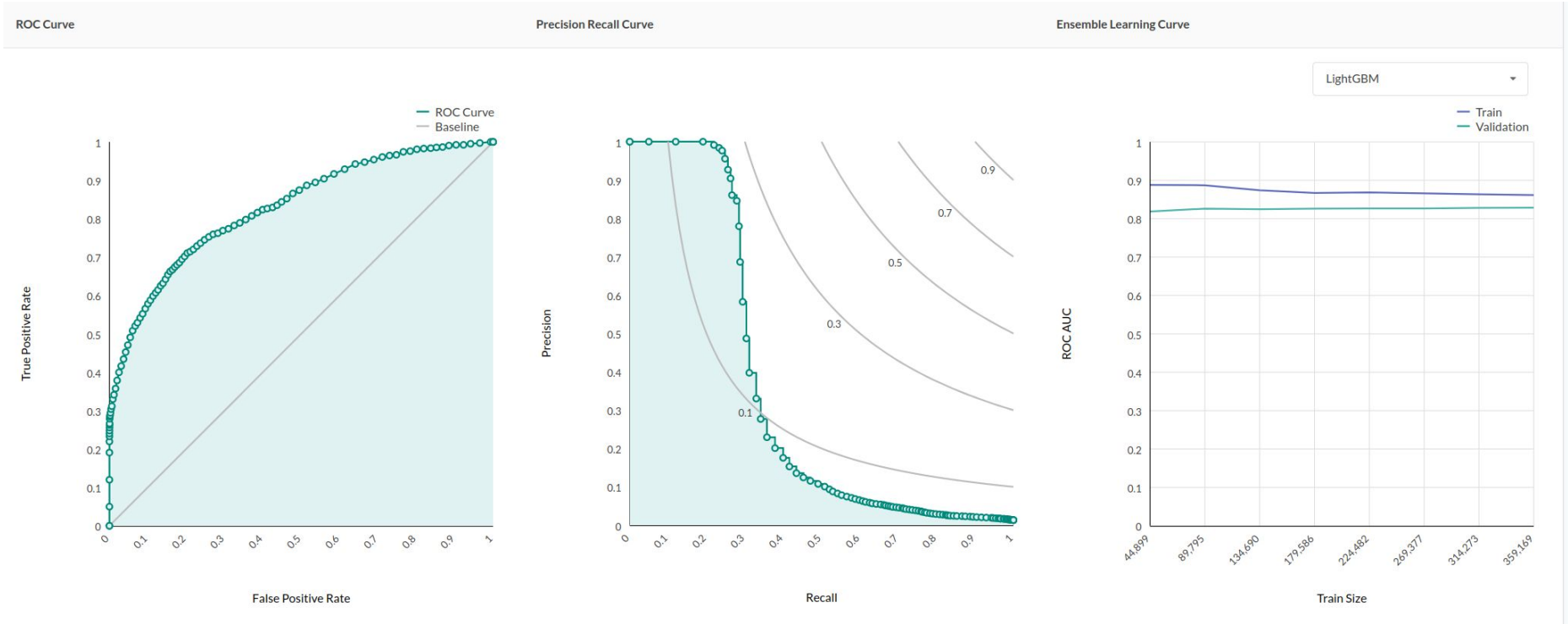
Precision Recall Curve



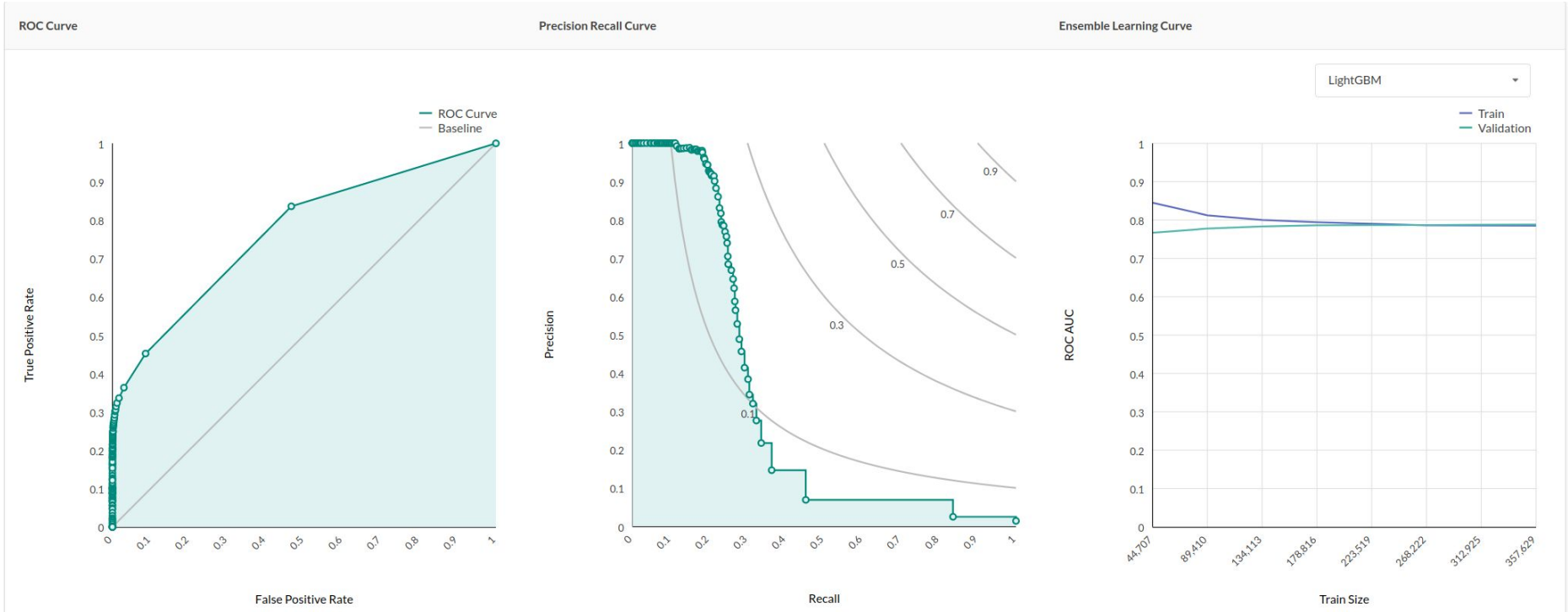
Ensemble Learning Curve



Lung cancer


ROC Curve
Precision Recall Curve
Ensemble Learning Curve
ROC Curve
Baseline
True Positive Rate
False Positive Rate
Precision
Recall
0.9
0.7
0.5
0.3
0.1
LightGBM
Train
Validation
ROC AUC
Train Size
44,899
89,795
134,690
179,586
224,482
269,377
314,273
359,169

Colorectal cancer



Breast cancer