Development and validation of proteomic models to assess susceptibility for high-risk cancers associated with military duty

Rosalynn Gill[1], Sama Shrestha[1], Hannah Biegel[1], Jessica Chadwick[1], Jess Kuzma[2], Clare Paterson[1], In collaboration with the EPIC – Proteomic investigators
[1]Illumina, San Diego, California, USA. [2]SomaLogic, Boulder, Colorado, USA.

Introduction

- Military personnel and veterans experience elevated incidence of several cancers compared with the general U.S. population, likely reflecting a combination of occupational exposures, environmental hazards, and screening practices.
- Breast cancer is the most common cancer diagnosed in active-duty women, and deployed female veterans have a 13% higher breast cancer risk than non-deployed veterans.
- Prostate cancer occurs at elevated rates among service members and military aviators, with studies reporting 16–20% higher incidence than the general U.S. population after adjustment for demographic factors.
- Lung cancer remains a major health burden for veterans, with nearly 8,000 new diagnoses annually and approximately 5,000 veteran deaths each year, making it the deadliest cancer in this population
- Routine screenings can miss high-risk cancers associated with military duty due to gaps in understanding the impact of military environmental exposures.
- Improved risk stratification tools are needed to identify cancers associated with military environmental and occupational exposures.

Objectives

This study investigated whether high-throughput proteomics could be used to develop lung, breast, and prostate cancer susceptibility models that may help stratify individual risk, direct screening procedures, and monitor changing risk over time.

Methods

- Using the SomaScan™ Assay v4.1 modified-aptamer proteomics technology, we assessed ~7,000 proteins in 14,787 citrate plasma samples from cancer-free individuals aged 35-76 from the European Prospective Investigation into Cancer and Nutrition study, totaling >103 million protein measurements.
- Cases and controls were split into training, verification, and validation sets for each cancer subtype.
- Machine learning techniques were used to identify separate survival models that predict risk of a malignant cancer diagnosis (lung, breast, or prostate) within 5 years of blood draw.
- Models were fit using the entire 20-year follow-up with predictive performance optimized and evaluated at 5 years. The final models were each assessed on their respective hold-out validation sets.
- Discrimination was assessed via C-index and 5-year AUC.

Table 1. Study Demographics.

Covariate	Measure	Total	Malignant Lung Cancer Diagnosis within 5 years		
			Yes	No	Censored
Sample Size	N	14322	98	12645	1579
Tobacco Use	Current	3922 (27.4%)	55 (56.1%)	3487 (27.6%)	380 (24.1%)
	Never	6324 (44.2%)	8 (8.2%)	5638 (44.6%)	678 (42.9%)
	Former	4076 (28.5%)	35 (35.7%)	3520 (27.8%)	521 (33%)
Sex	Female	7962 (55.6%)	38 (38.8%)	7046 (55.7%)	878 (55.6%)
	Male	6360 (44.4%)	60 (61.2%)	5599 (44.3%)	701 (44.4%)
Age (years)	Median	56	57	56	59
	Range	35 - 75	42 - 74	35 - 75	35 - 75
Time to Event or Censoring (years)	Median	13.60	2.86	14.27	2.84
	Range	0.01 - 20.42	0.12 - 4.89	5.00 - 20.42	0.014 – 5.00

Covariate	Measure	Total	Malignant Breast Cancer Diagnosis within 5 years		
			Yes	No	Censored
Sample Size	N	7922	178	7010	734
Tobacco Use	Current	1685 (21.3%)	30 (16.9%)	1486 (21.2%)	169 (23%)
	Never	4611 (58.2%)	106 (59.6%)	4107 (58.6%)	398 (54.2%)
	Former	1626 (20.5%)	42 (23.6%)	1417 (20.2%)	167 (22.8%)
Sex	Female	7922 (100%)	178 (100%)	7010 (100%)	734 (100%)
Age (years)	Median	56	54	56	59
	Range	35 - 75	35 - 75	35 - 75	35 – 75
Time-to Event or Censoring (years)	Median	13.76	2.96	14.35	2.84
	Range	0.01 - 20.42	0.04 - 5.00	5.00 - 20.42	0.014 – 5.00

Covariate	Measure	Total	Malignant Prostate Cancer Diagnosis within 5 years		
			Yes	No	Censored
Sample Size	N	6395	117	5629	649
Tobacco Use	Current	2246 (35.1%)	20 (17.1%)	2011 (35.7%)	215 (33.1%)
	Never	1699 (26.6%)	38 (32.5%)	1519 (27%)	142 (21.9%)
	Former	2450 (38.3%)	59 (50.4%)	2099 (37.3%)	292 (45%)
Sex	Male	6395 (100%)	117 (100%)	5629 (100%)	649 (100%)
Age (years)	Median	56	62	56	59
	Range	35 - 75	43 - 75	35 - 75	36 - 75
Time-to Event or Censoring (years)	Median	13.36	3.33	14.16	2.75
	Range	0.02 - 19.91	0.03 – 5.00	5.00 - 19.91	0.02 – 5.00

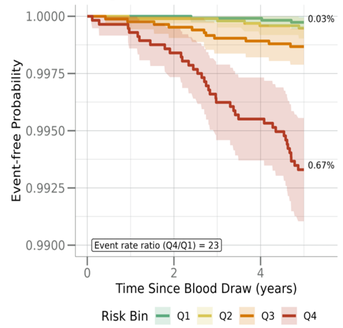
Results

- The best performing model for each of the 3 cancers was an accelerated failure time model with Weibull distribution:
  - Final lung cancer risk model included 13 features with a 5-year AUC of 0.74 and C-Index of 0.61 in a hold-out validation dataset.
  - Final breast cancer risk model included 10 features with a 5-year AUC of 0.66 and C-Index of 0.55 in a hold-out validation dataset.
  - Final prostate cancer risk model included 13 features with a 5-year AUC of 0.84 and C-Index of 0.48 in a hold-out validation dataset.
- Performance of these protein models was higher than the average performance of Polygenic Risk Score-based or clinical models, and age-only models.
- Proteomic models also displayed strong separation between highest and lowest proteomic-risk score-based quartiles.

Table 2. Discrimination metrics for proteomic-based models across cancer types in training and validation dataset at 5 years post blood-draw. C-index is time independent. Sensitivity and Specificity calculated using cutoff that maximized Youden's J on the training dataset.

Cancer Model	Dataset	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	C-index (95% CI)
Lung	Training	0.780 (0.720, 0.830)	0.804 (0.696, 0.909)	0.702 (0.690, 0.713)	0.790 (0.749, 0.828)
	Validation	0.739 (0.627, 0.841)	0.733 (0.583, 0.895)	0.682 (0.670, 0.703)	0.607 (0.473, 0.819)
Breast	Training	0.718 (0.672, 0.766)	0.728 (0.654, 0.813)	0.626 (0.607, 0.639)	0.652 (0.625, 0.680)
	Validation	0.657 (0.577, 0.737)	0.641 (0.513, 0.784)	0.620 (0.593, 0.645)	0.550 (0.498, 0.607)
Prostate	Training	0.876 (0.841, 0.910)	0.863 (0.790, 0.935)	0.757 (0.740, 0.771)	0.688 (0.648, 0.728)
	Validation	0.837 (0.764, 0.893)	0.767 (0.600, 0.900)	0.737 (0.717, 0.769)	0.484 (0.411, 0.727)

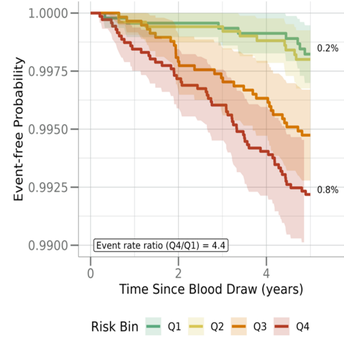
Lung Cancer



Model Proteins

HE4 | PGCB | sICAM-5 | PSMG4 | KIN17 | PSA | CTSA | IDLC | SCF | IGF1 | INHBA | LPLC1 | PPBN

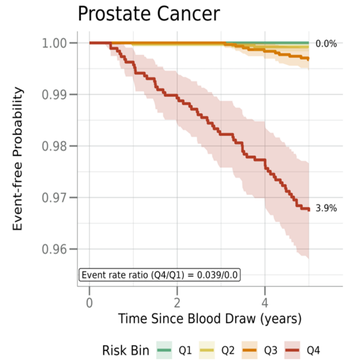
Breast Cancer



Model Proteins

AGR3 | PCSK9 | PIANP | KLK5 | SCG3 | OBP2B | MAG | PSMA | LIPK | GIP

Prostate Cancer



Model Proteins

PSA* | BPSA | IDE | CDON | DCE1 | PSMP | SLIT2 | Mn SOD | T106B | CLEC3B | PTN

*3 PSA aptamers are included in the final model

Conclusions

- We successfully developed proteomic models to predict 5-year risk of the three high-risk cancers associated with military duty.
- Proteomic signatures outperformed traditional risk stratification methods, highlighting their potential for early detection and personalized screening strategies.

SomaLogic Operating Co., Inc.
2945 Wilderness Pl. Boulder, CO 80301
somalogic.com