ARIC Manuscript Proposal # 3909 (revised)

PC Reviewed: 10/12/21	**Status:**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Identification of Plasma Proteins for the early detection of cancer in ARIC

b. Abbreviated Title (Length 26 characters): Proteins & cancer detection

2. Writing Group: Ru, Platz, Douville, Cohen, Kinzler, Papadopoulos, Vogelstein, Coresh, Joshu, Lu, Prizment, Butler, other interested ARIC investigators
Will invite: Couper.

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. MR (ARIC author) **[please confirm with your initials electronically or in writing]**

First author: **Meng Ru, MS**
Address: Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St.
Baltimore, MD 21205
Phone: 347-574-0811
E-mail: menaru@jhu.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: **Elizabeth A. Platz, ScD, MPH**
Address: Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St., Room E6132
Baltimore, MD 21205
Phone: 410-614-9674
E-mail: eplatz1@jhu.edu

3. Timeline: Expect to complete this work by December 2022

4. Rationale:

The prognosis of some cancer can be greatly improved if these cancers are detected early. A recent study combined genetic alterations and a small number of candidate protein biomarkers into a multi-analyte blood-based test called CancerSEEK to identify the presence of cancer[1,2]. The test has a median sensitivity of 70% in 8 common cancer types (breast, colorectum, esophagus, lung, ovary, pancreas, stomach, liver, lung) while maintaining a high specificity (>99%) in the non-cancer samples. The utility of this test when coupled with PET-CT was

documented in the DETECT-A trial in women[1]. However, a more informative set of proteins markers beyond the preliminary candidates may improve the sensitivity of detection of early-stage cancers (e.g., pre-clinical) while optimizing specificity for population use. Additional proteins may extend the types of cancers identified beyond the 8 cancer types, including those without currently available screening tools.

We propose a case-control study[3] to identify plasma proteins that differ in level between cancer cases and controls using the SomaLogic protein data from approximately 10,539 participants of the Atherosclerosis Risk in Communities (ARIC) study[4]. The ARIC study has already measured the protein data at three time points and has ascertained cancer incidence over nearly 30 years, making it a unique resource for discovering protein biomarkers for cancer screening. We will identify cancer cases as those with cancer diagnosis following their biospecimen collection, and restrict the controls to persons who never had a cancer history before or after the time that the sample used for protein measurement was collected. This study differs in concept and approach from one addressing proteins in the etiology or risk of cancer incidence in that we aim to identify proteins that are produced by tumors or in response to tumors (rather than proteins that mark exposures or response to exposures), and accordingly, we hope to identify proteins in the ARIC interval blood draw that are produced by tumors or in response to tumors that precede initial cancer diagnosis, and can discriminate subsequent cancer cases and controls well. As shown in previous literature[5–7], markers with etiological associations might not necessarily be good candidates for classifiers.

Participants have ~5,000 proteins measured by SomaScan v.4 (SomaLogic, Boulder, CO), an aptamer-based technology, at visit 2, 3 and visit 5 in the ARIC study[8]. Of ARIC participants diagnosed with cancer, we will select those diagnosed with cancer after their ARIC visit. Cases will be defined as those who had a cancer diagnosis within pre-specified windows of time after the sample used for protein measurement was collected (e.g., 2, 3, 5 years) at visit 2, 3, or 5. Since visits 2 and 3 were 3 years apart, for cases diagnosed within >0 to 7 years from visit 3 sample collection, we will also use their visit 2 protein as part of the repeated measurements for the discovery analysis. Controls will be defined as those without a cancer (invasive) history (except non-melanoma skin, for which ascertainment is not done by cancer registries) throughout the entire ARIC follow-up and did not die of cancer. We will include all ARIC participants that meet the eligibility criteria for “control”; matching will not be done. Hori et al. showed that tumors could start to grow 10 years before becoming detectable by biomarker preclinical phase[9]. Thus, we will explore a range of time windows for detection for the purpose of fully discovering any possible protein patterns in the early phase of cancer’s natural history or the body’s early response to a tumor, at a time before the cancer becomes detectable in screening tests in routine use or the onset of symptoms. We will compare the performance of discovery and prediction at different time windows, but we expect the best predictive capacity would be maximized in the shorter time windows from blood draw to cancer diagnosis.

At the time, we are not aware of other prospective cohort studies with repeated SomaScan measures. Thus, we will use the following strategy in this study: We will divide the eligible participants data into two sets (70% vs 30%) for the discovery and prediction model building, and for the model performance assessment.
In the discovery analysis and model building phase using the 70% set:

- First, we will scan for proteins that differ between the cases and controls irrespective of cancer site or any other potential confounders. To account for the interactive nature of proteins, we will scan proteins for differences between cases and controls first individually, then in clusters or networks of proteins.
- Second, we will examine if proteins that differ between the cases and controls are related to 1) demographics (age, sex, race, field center), socioeconomic factors (life course SES, access to and uptake of healthcare), cancer site, stage at cancer diagnosis, tumor histology, 2) then integrating with major cancer risk and protective factors (smoking, obesity, alcohol drinking, diabetes, family history) adjusted by propensity scores using the inverse probability of weighting method.
- Next, in an exploratory step, we will repeat the prior two steps within the subgroups of:
    - common cancer sites either individually (lung, breast, prostate, colorectal, pancreatic, ovarian) or in subgroups with expected common damage or activated/inactivated pathways (e.g., obesity associated, smoking associated, gastrointestinal, hormone). For prostate cancer, controls will be restricted to men (presumed persons with a prostate). For breast (female post-menopausal), ovarian, and endometrial cancer, controls will be restricted to women (presumed persons with these organs)
    - sex/gender (for cancers of non-sex specific organs), race (e.g., due to receptor status prevalence differences, such as triple negative breast cancer), stage at diagnosis (e.g., the goal is to detect cancers while they are localized and not symptomatic), histology (e.g., adenocarcinoma), cancer risk or protective factors that may be useful for risk stratification for screening (e.g., already know that smoking associated cancers are more likely to be found in smokers – smoking history is incorporated into lung cancer screening guidelines), or that affect the likelihood of cancer diagnosis in ARIC (e.g., access to and uptake of care).
- For model building, we will use statistical models suitable for high-dimensional data, such as linear mixed methods with lasso or non-linear methods (e.g., support vector machine, random forest model, and neural net) to identify a reduced number of protein markers with optimal classification performance. the leave-one-out cross-validation method to select the model with a group of proteins with optimal classification performance. The ideal panel of proteins should have a satisfactory sensitivity as well as a very high specificity for population screening.

In the predictive modeling analysis phase using the 30% set:
We will use the remaining sample to internally validate the prediction model for early cancer classification. We will evaluate the top performers based on model fitness, discrimination statistics (ROC/AUC, sensitivity and specificity, especially true negative rates on non-cancer samples). We will determine a threshold in this dataset that maximizes specificity needed in a cancer screening setting.

We are aware that circulating proteins can differ within and between person due to genetic polymorphisms, pseudogenes, and post-translational modifications, and different technologies may have differing sensitivities for their detection. Thus, after sets of proteins are identified in this study in ARIC, the Papadopoulos/Kinzler/Vogelstein laboratory will first calculate the correlation between the measurements of proteins from SomaScan we discovere and those from the BioPlex platform which CancerSEEK used, and then conduct an independent protein identity confirmation analysis using non-ARIC samples. This step is necessary for the development of screening tools. Though variation was seen in absolute measurements of proteins from the two platforms[10], relative changes (fold change or trend) were reported to hold between different platforms[11]. Therefore, we hypothesize the proteins we identified with elevated or lowered levels should be recognized in both platforms.

Key points:

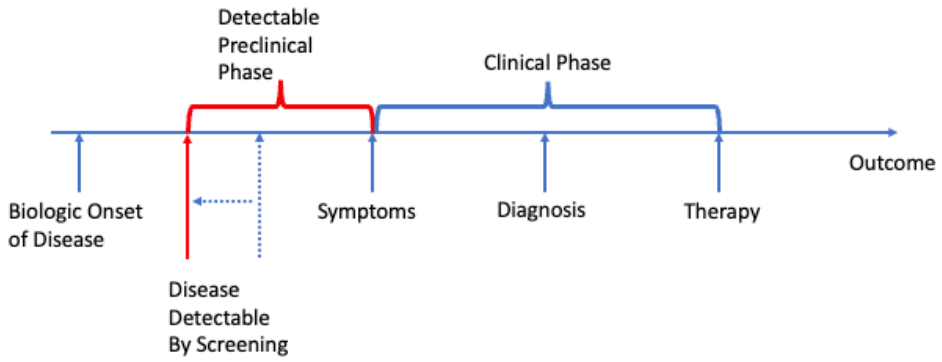
In this study, our primary goal is to identify proteins or patterns of proteins across cancer sites in participants with a cancer diagnosis within 2-5 years that differ in plasma level from participants without a cancer history. We aim to (1) discover the unique proteins or protein clusters that differ among cancer sites, as well as signals of proteins specific to cancer sites or other demographic factors, and (2) build a prediction model for early cancer classification based on the panels of proteins selected from the discovery analysis.

5. Main Hypothesis/Study Questions:

Our overarching hypothesis is that before a cancer becomes detectable using current screening tools or before the onset of symptoms, plasma proteins directly secreted from tumor cells or a result of the body's response to a growing tumor can be detectable. Such proteins individually or in combination, might have utility for early detection.

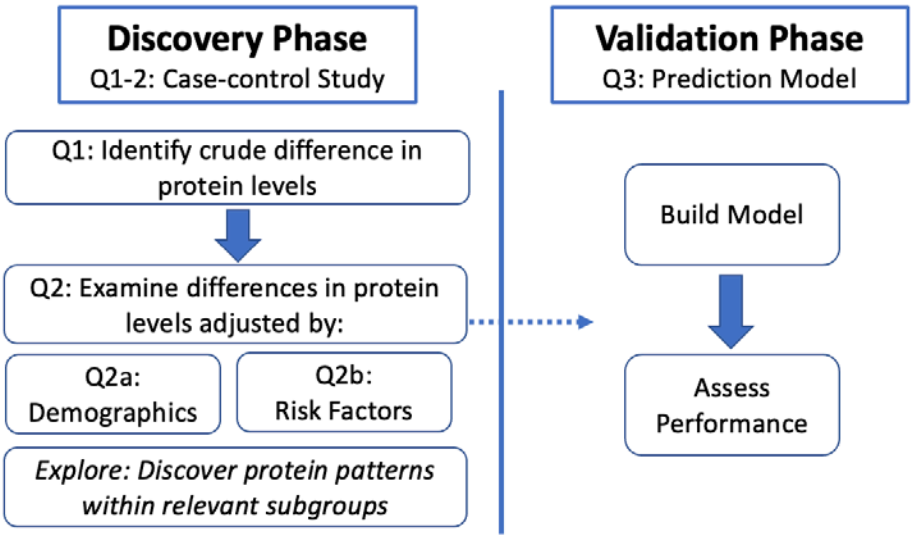
Therefore, we propose this study to discover proteins that may have used for cancer screening in currently asymptomatic populations. Our goal is to identify a panel of protein markers that can (1) either maximize the window of the detectable preclinical phase, in other words, enable cancer to be detected at an even earlier moment than current screening tools, which was, in theory, possible due to the high sensitivity and reliability of SomaScan for individual protein measurements, and our strategy to discover panels of proteins accounting for their interactive nature, (2) increase the sensitivity/specificity of the screening in the same time window of preclinical phase for cancers with current screening tools, or (3) detect cancers that do not currently have a routinely used screening tool (Figure 1). We also expect that we will detect aggressive cancers over indolent cancers (important for common cancers like breast and prostate). We will further compare proteins between cases that later did and did not die of their cancer.

Figure 1. Screening timeline[12].



Among ARIC participants diagnosed with a first primary (invasive) cancer within windows of time after protein collection (e.g., 2, 3, 5 years) at visit 2, 3 or 5, and participants without a cancer history, evaluate (Fig 2):

Figure 2. Flow chart of study design and primary objectives.



Q1: What proteins are statistically different between cancer cases and controls?

Q2a: What proteins are statistically different between cancer cases and controls after adjusting for propensity score using inverse probability of weighting estimated based on demographics (age, sex, race, field center) and socioeconomic factors (life course SES, access to and uptake of healthcare)?

Q2b: What proteins are statistically different between cancer cases and controls after adjusting for propensity score and integrating major cancer risk and protective factors (smoking, obesity, alcohol drinking, diabetes, family history) with covariates mentioned in Q2a?

Exploratory analysis:
We will repeat the Q1-2 within the subgroups of cancer sites either individually or in subgroups with expected common damage or activated/inactivated pathways; sex/gender, race, stage at diagnosis, and histology.

Q3a: Can the set of proteins above (Q1-2) predict subsequent near-term cancer status and

Q3b: What is the predictive model's performance on the withheld samples?

6. Design and analysis (study design, inclusion/exclusion, outcome, and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Study design:

Q1 and Q2a-b – Identify a subset of proteins using a case-control set.

Q3 – Develop a prediction model using the subset of proteins in the case-control study and assess of performance on a withheld set of samples.

Analytic population:
Q1 and Q2: Men and women who have Visit 2, 3, or 5 protein scan data that passed quality control checks, who consented to studies on chronic diseases including cancer are eligible. We will exclude participants who are not White or Black and participants who are Black from the Washington County and suburban Minneapolis (small numbers).

From these participants, we will select all eligible cancer cases, and all eligible participants without a cancer history, using these criteria (Figure 3-4, Table 1):

- Subsequent cancer cases - participants who had one or more first primary invasive cancers (other than non-melanoma skin) or bladder in situ (state cancer registries require reporting) diagnosed within windows of time after protein collection (e.g., 2, 3, 5 years) at visit 2, 3, or 5. Participants with a non-melanoma skin cancer or a pre-malignant tumor who develop an invasive cancer of any site after visit 2, 3, or 5 are eligible to be selected as a case.
    - For cancer patients diagnosed within >0 to 7 years from visit 3, their visit 2 protein data will also be used (repeated measures) to expand the detection window.
- Participants without a cancer history - participants who never had a cancer diagnosis by Visit 5 (current end of cancer follow up) and did not die of cancer (2018 is the current end of cancer end of follow up). Participants with a diagnosis of non-melanoma skin cancer or in situ malignancies are eligible to be controls (aside from bladder in situ (state cancer registries require reporting).

Figure 3. Participants to be included and excluded from the analysis

Study population	Status at Visit 2, 3, or 5	Status at 2, 3, 5 years after Visit 2, 3,5	Status in the analysis
Row 1: Included	No cancer	First primary cancer(s)	Case
Row 2: Included	No cancer	No cancer	Control
Row 3: Excluded	Cancer(s)	No cancer	-
Row 4: Excluded	Cancer(s)	Cancer(s)	-

Figure 4. Case Definition

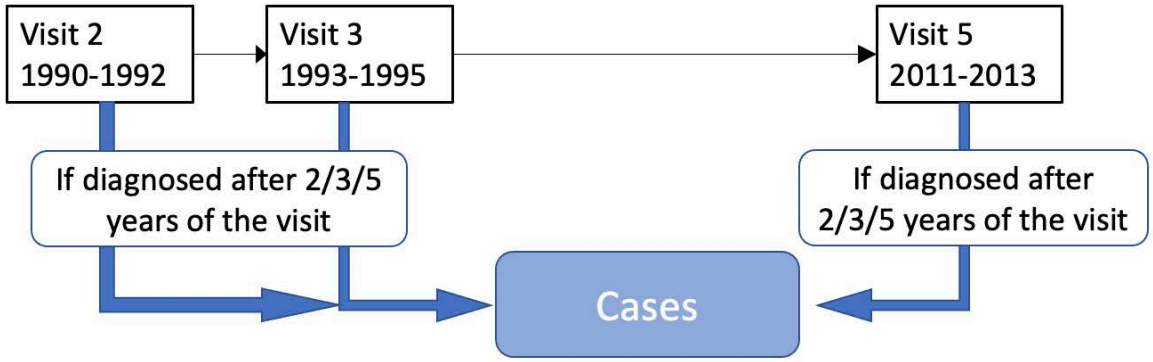


Table 1. Preliminary* sample size for cancer cases and controls.

Participants who developed cancer within 5 years after Visit 2/3/5 (Row 1 in Figure)	1,100
Participants without a history of cancer through 12/31/2015 and who did not die of cancer through 12/31/2018 (Row 2 in Figure)	10,054
Cancer cases after Visit 2	255
>0 to 2 Years	119
>0 to 3 Years	219
>0 to 5 Years	255
Median time from Visit 3 to cancer dx [IQR]	2.10 [1.34 – 2.73] years
Cancer cases after Visit 3	527
>0 to 2 Years	192
>0 to 3 Years	307
>0 to 5 Years	527
Median time from Visit 3 to cancer dx [IQR]	2.61 [1.39, 3.88] years
Cancer cases after Visit 5 (through the max follow up of 12/31/2015)	318
>0 to 2 Years	183
>0 to 3 Years	259
>0 to <5 Years	318
Median time from Visit 5 to cancer dx [IQR]	1.72 [0.89, 2.76] years

* The sample size here might be lower after utilizing the annual follow-up or semi-annual follow-up responses to identify self-reported cancer diagnosis to refine the control group between the end of cancer follow up in 12/31/2015 and end of death follow up in 12/31/2018 (e.g., diagnosed with a cancer in that interval, but did not die of it) and other possible exclusion.

Table 2. Sample size for cancer site by time to diagnosis after specified.

		Time to Diagnosis		
Cancer Site	**Total**	**>0 to <5 Years After**	**>0 to <5 Years After**	**>0 to <5 Years**

		Visit 2 (N=255)	Visit 3 (N=527)	After Visit 5 (N=318)
Head & Neck	23	1	15	7
Colon	81	15	41	25
Rectal	12	1	7	4
Liver	5	0	1	4
Pancreatic	49	13	16	20
Stomach	20	7	9	4
Other digestive	25	6	12	7
Lung/Bronchus	182	63	78	41
Other Respiratory	9	2	7	0
Hematopoietic and Lymphatic	104	24	37	43
Melanoma skin	32	2	14	16
Breast	147	31	74	42
Cervical	7	1	4	2
Endometrial	21	2	10	9
Ovarian	16	5	8	3
Prostate	210	44	128	38
Bladder	48	5	20	23
Kidney	27	6	13	8
Brain	21	10	9	2
Thyroid	4	0	3	1
Other	42	9	16	17
Unknown	23	10	8	5

* Highlighted: cancer sites with >40 cases, which we will focus on cancer site subgroup analysis.

Measurements:

Protein scan – We will use protein levels, measured as relative fluorescence units [RFU] for a standard plasma volume per participant using aptamer-based profiling in blood collected at Visit 2, 3 and 5.

Some proteins were FLAGGED on one or more plates. Several options are available for handling such proteins: exclude protein (conservative), include protein (not conservative), exclude only if the number of plates with the flagged protein is larger than expected by chance alone (optimal, but need to define an arbitrary expected)

Cancer – Existing 2015 ARIC cancer case file, which captures cases diagnosed between Visit 1 and 12/31/2015; we will also utilize the annual follow-up or semi-annual follow-up responses to identify self-reported cancer diagnosis to refine the control group between the end of cancer follow up in 12/31/2015 and end of death follow up in 12/31/2018 (e.g., diagnosed with a cancer in that interval, but did not die of it).

Additional variables (either covariates or for use in sensitivity analyses):

Demographics: age at Visit with proteins, age at diagnosis, sex (Visit 1), race (Visit 1), field center (Visit 1)

Socioeconomic factors: Lifecourse SES (Visit 4), neighborhood income, health insurance status, frequency of routine physical examination, having a dentist, frequency of routine dental visit, last time of dental visit (Visit 4)

Other covariates: Body mass index (Visits 2, 3, 5), waist circumference, height (Visit 1), blood volume (derived from height and weight), cigarette smoking status and pack years (accumulated to Visit 2, 3, 5), DNA methylation predicted smoking pack years, alcohol consumption (Visit 5), diabetes status (accumulated Visit 2, 3, 5, also included undiagnosed diabetes, uncontrolled diabetes), hypertension medication use, cholesterol medication use, aspirin use (Visit 2, 3, 5), statin drug use (Visit 2, 3, 5), hormone replacement therapy use (women, Visit 2, 3, 5)

Factors that influence plasma protein levels or are indicative of acutely distorted protein levels (for sensitivity analyses): In a subanalysis, we will exclude participants using a diuretic at Visits 2, 3, 5. Key proteins that mark kidney and liver function and acute phase inflammation were measured currently in the protein scan. In the sensitivity analysis, we will exclude individuals at the extremes of the distributions of these: cystatin-C; aspartate aminotransferase [AST] and alanine aminotransferase [AAT]; and C-reactive protein, respectively (Visits 2, 3, 5). As an alternative, we could use indicators previously measured, albeit, not concurrently with the protein scan, that have clinically relevant cutpoints: eGFRcr-cys (15 mL/min/1.73m^2 [stage 5]), hsC-reactive protein concentration (acute inflammation, 10 mg/L), liver function (>3 times the reference ranges: ALT 7-56 IU/L, AST 0-35 IU/L).

Statistical analysis

In the discovery and prediction model building phase (Q1-3a), 70% of data from visit 2, 3 and visit 5 will be used.

Q1: Assess the crude difference in plasma protein levels between cancer cases and controls
In this step, we aim to identify any pre-diagnostic plasma protein levels that differ between the cancer cases and controls using linear regression models for each protein. The model will only contain a fixed factor for the binary indictor of being a cancer case or a control. This differential analysis will be conducted without adjustment to reflect the feasibility and applicability of using plasma protein levels for cancer detection in the general population.

Q2a: Assess the difference in protein levels between cases and controls with propensity score weights estimated by demographics (age, sex, race, field center) and socioeconomic factors (life course SES, access to and uptake of healthcare) using linear regression models with the same specifications as in Q1.

Q2b: Assess the difference in protein levels between cases and controls with propensity score weights estimated by major cancer risk and protective factors (smoking, obesity, alcohol drinking, diabetes, family history) as well as factors used in Q2a using the same linear regression models.

Exploratory analysis: Analysis in Q1-2 will be repeated with the same method in the following subgroups to allow for discovery of different sets of proteins: major cancer sites (prostate, breast, lung, colorectal), sex/gender, age, race, study site, stage at cancer diagnosis, cancer histology, major cancer risk factors (smoking, obesity, alcohol drinking, diabetes, and family history of cancer)

Since there were only 3 years apart from visit 2 and 3, it's likely that some cancer cases diagnosed after visit 3 can still fall under the early detection time window for visit 2. In that case, we plan to conduct a sensitivity analysis using a linear mixed model to examine if time-updated protein levels can enhance the discovery.

We will also determine whether the proteins newly identified enrich certain pathways (e.g., using propriety software such as Ingenuity https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/, or use publicly available databases such as STRING https://string-db.org/cgi/input.pl?sessionId=MsFSlNgwAW97&input_page_show_search=on, which includes KEGG https://www.genome.jp/kegg/brite.html#gene and Reactome https://reactome.org/PathwayBrowser/ databases. We will also examine the Spearman correlations among these proteins in the cases and in the controls.

Q3a: Prediction model building.

Based on the findings from the discovery phase of analysis, we will use leave-one-out cross-validation method to build the prediction model with the most informative protein sets. More sophisticated statistical models, such as linear mixed methods with lasso or non-linear methods (e.g., support vector machine, random forest model, and neural net), will be adopted as necessary to reduce the dimensions and capture the combinatorial protein patterns. In addition to building one model predicting all near-term cancer diagnoses, we will also explore the potential of combining multiple models built for each major cancer types to enhance power and account for the heterogeneity among cancer types.

The remaining 30% of the participants will be used for performance assessment of the prediction model built in Q3a.

Q3b: Prediction model evaluation.

We will select the top performers based on model fitness, discrimination statistics (ROC/AUC, sensitivity and specificity, especially true negative rates on non-cancer samples). We will determine a threshold in this dataset that maximizes specificity needed in a cancer screening setting.

We will also evaluate the performance of prediction in each time window, that is, cancer diagnosis within 2, 3, 5 years after visit 2, 3, 5 using a time-to-event analysis (Cox regression) model to generate AUC to reflect the timing feature of the cancer diagnosis.

We will apply our model using the pre-determined cutoff for positivity to a completely independent external validation set from the CancerSEEK study. The CancerSEEK samples offer a large number of cancers and non-cancers to determine how robust our estimates for sensitivity and specificity are.

Propensity score

For Q2a, we will model the association of age (continuous), sex, race, field center, lifecourse SES, neighborhood income, health insurance status, frequency of routine physical examination, having a dentist, frequency of routine dental visit, last time of dental visit with cancer status using logistic regression to predict the propensity score for each participant[13]. We will confirm the positivity assumption of the use of the propensity score (i.e., no/negligible number of participants have a probability of 0 or 1 of being a cancer case) and determine whether the overlap in scores between cancer cases and those without a cancer history is satisfactory. Inverse probability weighting of propensity score will be used to avoid potential mismatch on propensity scores between the cancer cases and controls on both ends of the distributions of propensity scores. For Q2b, we will repeat these steps with an expanded set of variables with cancer risk/protective factors, which include smoking, obesity, inactivity, alcohol drinking, diabetes, family history.

An additional sensitivity analysis will be performed based on matched propensity scores on overlapped cases and controls to test the robustness of results.

* Including common cancer risk/protective factors via a propensity score will allow us to investigate the incremental value of plasma protein levels on cancer detection. Moreover, it could provide new insights on risk stratification for cancer screening in future practice, such that a specific group of smokers with certain protein patterns would carry higher risk of certain cancer and should be the target population for cancer screening.

We will generally follow statistical analysis approaches described previously for proteomics data in epidemiologic studies[14–16].

Multiple testing Adjustment

In the individual protein discovery stage, we will consider false discovery rate (FDR) and Bonferroni correction to deal with multiple testing. In the stage of prediction modeling with combinations of proteins, we aim to find a minimum set of proteins that will maximize the prediction together.

Minimum detectable association

Based on conventional p-values corrected for multiple testing (alpha = 0.05/5000 = 0.00001) and currently available sample size estimates, 490 cases from the 2-year window and 10,000

controls, we estimate the minimum detectable effect size using a 2-sided t-test with a power of 80%. Utilizing the pwr.t2n.test function from the R package 'pwr', we were able to obtain the Cohen's D effect size as 0.24 between cases and controls, which is considered a small to medium effect size. Cohen's D effect size is defined as mean difference between two groups divided by pooled standard deviation[17].

Points of clarification, challenges, and potential pitfalls and solutions:

- As some cancers can be detected and diagnosed early on through active screening, such as breast cancer and prostate cancer, the diagnosis of these cancers is highly dependent on the access to healthcare. Therefore, when estimating the propensity score, access to care is an essential factor to consider. However, given the age of the participants in the ARIC cohort, many of them would be eligible for Medicare. Thus, there would be less variability on health insurance coverage, which makes it a suboptimal indicator for access to care. Thus, we included dental insurance coverage, frequency, and usage of dental care to better address this confounder.

- In case there is no satisfactory overlap in propensity scores estimated by the factors mentioned previously, we choose to perform weighted analyses with propensity scores in our main analyses. We will run propensity score matching on mostly overlapped samples between cases and controls as a sensitivity analysis.

- In the ARIC cohort, we only know the timing of a cancer diagnosis, not the earliest date of symptom onset. Therefore, it's possible to have cancer patients already manifesting symptoms when their samples were drawn at the study visit. This means the signals in their blood samples might have passed beyond the time window for screening interventions. Therefore, we might need to expand our detection window (e.g. up to 10 years) to reduce the chance that they were already symptomatic but undiagnosed at blood draw. We will also utilize the annual follow-up or semi-annual follow-up to identify any symptoms or hospitalizations between visits that might indicate cancer.

- Non-melanoma skin cancers (basal, squamous cell) are not routinely collected by US cancer registries. While we will not be able to exclude participants with non-melanoma skin cancer from the controls, we will handle the cases in a parallel manner, if their medical records or self-report history indicate a non-melanoma skin cancer; if the skin cancer is the first diagnosis, and a second primary cancer occurs, then these participants would be eligible for selection as a first primary cancer based on the second primary cancer diagnosed. This strategy is appropriate in that we do not aim to identify proteins for the early detection of non-melanoma skin cancer, which has an indolent course.

  Similarly, we will not exclude from either the cases or controls participants with pre-malignant lesions because these are not systematically detected in populations and are not routinely collected in ARIC (e.g., cervix pre-malignant lesions, colorectal adenoma). Exceptions are those that are routinely collected by ARIC because cancer registries systematically collect them (non-invasive bladder cancer).

- For controls in the study, we did not specify that the participant must have intact each organ as aligned with the array of cancer sites detected with the cases in ARIC. In the US, hysterectomy with or without oophorectomy is a common procedure for the treatment of several non-cancer conditions. Also, women who carry BRCA1/2 mutations often have prophylactic mastectomy and oophorectomy. Including women who had such procedures in the control group would distort the comparison of proteins between cases (have the organ intake) and controls (do not have the organ). Thus, we will determine whether ARIC has sufficient information available to perform sensitivity analyses excluding participant who had organs that are common cancer sites surgically removed for reasons other than to treat cancer at any point in follow up.

- There is variability in proteomic profiles measured by different proteomics platforms. To investigate the reproducibility of our findings, we will also validate our discovery findings in proteins measured by another proteomics technology (e.g., ELISA) in the Papadopoulos/Kinzler/Vogelstein laboratory using non-ARIC samples. We also note, our detection capability of proteins might depend on the genetic polymorphisms, pseudogenes and post-translational modifications they have whether within or between person, potentially under the influence of the tumor.

- Given the large number of proteins in the scan, we recognize the potential statistical issue of inflating type I errors brought by the multiple tests we will perform on the proteins. We will use Bonferroni correction and false discovery rate to address this issue. However, since the proteins might be very likely correlated, the multiple tests we will perform will not likely to be independent, thus we prioritize finding groups of protein that are related to cancer diagnosis. We plan to further reduce the dimensions by using penalized regression models or feature selection algorithm or collapsing individual proteins to enriched pathways or groups with homogenous patterns identified by the models. We will also explore the approach of creating a composite score consisting of all the informative proteins identified and selecting the best set with the highest score. A later goal will be to test the performance of such score in an external sample.

- We recognize that the sample size can be small for ‘omics research for specific cancer site. Our main goal is to identify potential protein markers that capture the differences between those with and without a cancer history at a pan-cancer level. As the sample size allows, we will target on discovering proteins that provide an early warning for a specific cancer before the occurrence of symptoms, since different cancers can have substantially heterogenous etiology or mechanisms. We intend this study to provide preliminary data for future studies to further examine the role of novel candidate proteins in early detection in a bigger sample.

- We recognize that we might not know the proteins we will identify in this study will be coupled with appropriate treatment that will prevent cancer death. A randomized clinical trial like NLST (https://www.cancer.gov/types/lung/research/nlst)) will need to be performed as was done for spiral CT as the tool for lung cancer early detection .

7.a. Will the data be used for non-CVD analysis in this manuscript? __X__ Yes ____ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? ___X_ Yes ____ No
(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ____ Yes _X___ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ____ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status.
ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

___X___ Yes _____ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

MS # 3057: Repeatability and Longitudinal Variability of the Plasma Proteome (Tin, Coresh et al.)

MS # 3415: Characterizing prostate specific antigen (PSA) measured by SOMAscan and its correlates in men and women in ARIC (Platz, Tin, Coresh et al.)

Ancillary study: 2019.06 Systemic Inflammation, Aging phenotypes and Mortality in Cancer survivors and associated manuscript proposals to be submitted (Ugoji et al.)

MS # 3327 A proteomic analysis of incident dementia: The ARIC Study (Walker et al.)

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __X__ Yes ____ No

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number*
2017.27 Proteomic longitudinal ARIC study: SOMAscan of multiple visits
2011.07 Enhancing ARIC Infrastructure to Yield a New Cancer Epidemiology Cohort
1995.04 Cancer Study

___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes __X__ No.

References

1. Lennon AM, Buchanan AH, Kinde I, et al. Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science*. 2020;369(6499). doi:10.1126/science.abb9601

2. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018;359(6378):926-930. doi:10.1126/science.aar3247

3. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: Statistical guidelines for nested case-control studies. *BMC Med Res Methodol*. 2002;2(1):4. doi:10.1186/1471-2288-2-4

4. Joshu CE, Barber JR, Coresh J, et al. Enhancing the Infrastructure of the Atherosclerosis Risk in Communities (ARIC) Study for Cancer Epidemiology Research: ARIC Cancer. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2018;27(3):295-305. doi:10.1158/1055-9965.EPI-17-0696

5. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *Am J Epidemiol*. 2004;159(9):882-890. doi:10.1093/aje/kwh101

6. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. *JNCI J Natl Cancer Inst*. 2008;100(20):1432-1438. doi:10.1093/jnci/djn326

7. Kattan MW. Judging New Markers by Their Ability to Improve Predictive Accuracy. *JNCI J Natl Cancer Inst*. 2003;95(9):634-635. doi:10.1093/jnci/95.9.634

8. Gold L, Ayers D, Bertino J, et al. Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLOS ONE*. 2010;5(12):e15004. doi:10.1371/journal.pone.0015004

9. Hori SS, Gambhir SS. Mathematical Model Identifies Blood Biomarker–Based Early Cancer Detection Strategies and Limitations. *Sci Transl Med*. 2011;3(109):109ra116-109ra116. doi:10.1126/scitranslmed.3003110

10. Lim SY, Lee JH, Welsh SJ, et al. Evaluation of two high-throughput proteomic technologies for plasma biomarker discovery in immunotherapy-treated melanoma patients. *Biomark Res*. 2017;5(1):32. doi:10.1186/s40364-017-0112-9

11. Christiansson L, Mustjoki S, Simonsson B, Olsson-Strömberg U, Loskog ASI, Mangsbo SM. The use of multiplex platforms for absolute and relative protein quantification of clinical material. *EuPA Open Proteomics*. 2014;3:37-47. doi:10.1016/j.euprot.2014.02.002

12. Gordis L. *Epidemiology - 5th Edition*. Saunders and Company; 2013.

13. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8 Pt 2):757-763. doi:10.7326/0003-4819-127-8_part_2-199710151-00064

14. Ko D, Benson MD, Ngo D, et al. Proteomics Profiling and Risk of New-Onset Atrial Fibrillation: Framingham Heart Study. *J Am Heart Assoc*. 2019;8(6):e010976. doi:10.1161/JAHA.118.010976

15. Sampson DL, Parker TJ, Upton Z, Hurst CP. A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. *PloS One*. 2011;6(9):e24973. doi:10.1371/journal.pone.0024973

16. Mischak H, Critselis E, Hanash S, Gallagher WM, Vlahou A, Ioannidis JPA. Epidemiologic Design and Analysis for Proteomic Studies: A Primer on -Omic Technologies. *Am J Epidemiol*. 2015;181(9):635-647. doi:10.1093/aje/kwu462

17. Sawilowsky S. New Effect Size Rules of Thumb. *J Mod Appl Stat Methods*. 2009;8(2). doi:10.22237/jmasm/1257035100