Abbreviated Title: Connect for Cancer Prevention

Protocol #: 000034

Version Date: 07/21/2021 (v1.5)

Title: Connect for Cancer Prevention: A Prospective Cohort Study within Integrated Healthcare Systems in the US

Principal Investigator*:*

Dr. Montserrat Garcia-Closas
Division of Cancer Epidemiology and Genetics
National Cancer Institute
National Institutes of Health
9609 Medical Center Drive, Room 7E-404
Rockville, MD 20850-9774
Phone: 240-276 7648
E-mail: montserrat.garcia-closas@nih.gov

Coordinating Center: Division of Cancer Epidemiology and Genetics, National Cancer Institute

TABLE OF CONTENTS

STATEMENT OF COMPLIANCE

The protocol will be carried out in accordance with International Conference on Harmonisation Good Clinical Practice (ICH GCP) and the following:

- United States (US) Code of Federal Regulations (CFR) applicable to clinical studies (45 CFR Part 46, 21 CFR Part 50, 21 CFR Part 56, 21 CFR Part 312, and/or 21 CFR Part 812)

National Institutes of Health (NIH)-funded investigators and trial site staff who are responsible for the conduct, management, or oversight of NIH-funded trials have completed Human Subjects Protection and ICH GCP Training.

The protocol, informed consent form(s), recruitment materials, and all participant materials will be submitted to the Institutional Review Board (IRB) for review and approval. Approval of both the protocol and the consent form must be obtained before any participant is enrolled. Any amendment to the protocol will require review and approval by the IRB before the changes are implemented to the study. In addition, all changes to the consent form will be IRB-approved; a determination will be made regarding whether a new consent needs to be obtained from participants who provided consent, using a previously approved consent form.

1 PROTOCOL SUMMARY

1.1 SYNOPSIS

Title:	Connect for Cancer Prevention: a prospective cohort study within integrated healthcare systems in the US.
Study Description:	*Connect* is an observational cohort study that will prospectively collect exposure information and biospecimens and follow participants for cancer and other endpoints. Participants will be free of cancer at study invitation. Because most cancers develop over long periods of time and biological, behavioral and environmental factors can influence cancer development and change over time, participants will be asked to provide repeated exposure information and biological specimens. The study protocol takes advantage of developments in digital technologies as well as exposure and biomarker assessment tools to investigate suspected and emerging factors that can influence cancer development. The study data system will be built within an efficient, flexible and integrated cloud-hosted infrastructure that leverages modern interoperability standards in order to be an accessible research resource for current and future generations of scientists.
Objectives:	The primary objective is to build a state-of-the-art cohort in the US using new technologies and methods to provide a diverse and comprehensive research resource for the scientific community to study: • cancer etiology • precursor to tumor transformation • cancer risk assessment • early detection of cancer • second cancer development and survivorship after a cancer diagnosis The secondary objective is to establish a rich database connected to a biorepository for general research use.
Endpoints:	The primary endpoints are the continuum of cancer incidence, detection, progression, and survival. Secondary endpoints are numerous and could include methodological research, human biology, ancestry, evolution or health-related outcomes.
Study Population:	200,000 adults aged 40 to 65 years, no personal history of invasive cancer, and patients or members of participating integrated health care systems are eligible to participate.
Description of Sites/Facilities Enrolling Participants:	Nine US integrated health care systems will host enrollment.
Study Duration:	Recruitment will last at least five years. Participants will be followed for life. Analysis of the data is expected to continue for the foreseeable future.

Participant Duration:	Completion of the baseline study activities will take participants approximately four hours. Subsequent activities are expected to take one to four hours each.

1.2 SCHEDULE OF ACTIVITIES (SOA)

Study participant activities are expected to last through a participant's lifetime. Table 1 shows an example of a hypothetical participant's experience in the study from enrollment through active activities in the first ten years of follow-up. The frequency of research specimens collected directly from participants (i.e., blood, urine, saliva) could vary depending on age, risk of cancer, or other factors. The calendar shows retrieval of electronic health records (EHR) as an example of the passive activities (e.g., data linkages to regional or national databases) that will be performed at regular intervals during follow-up. This protocol details baseline activities at enrollment. Follow-up activities will be detailed in subsequent amendments.

[redacted]Example of Study Calendar for Data and Specimen Collections

Activities	**Enrollment and Retention Activities (years)**										
	Base-line	1	2	3	4	5	6	7	8	9	10
Study Eligibility Screener	X										
Enrollment	X										
Electronic Questionnaires and Personal Wearables											
Baseline questionnaire	X										
Serial assessments		X	X	X	X	X	X	X	X	X	X
Personal Wearables		X	X	X	X	X	X	X	X	X	X
Electronic Health Records (EHR)											
EHR retrieval	X	X	X	X	X	X	X	X	X	X	X
Biological Specimens											
Blood	X			X			X			X	
Urine	X			X			X			X	
Saliva	X						X				
Tissue, precursor lesions*	X	X	X	X	X	X	X	X	X	X	X
Tissue, cancer*		X	X	X	X	X	X	X	X	X	X
Fecal		X					X				
Hair, toenails		X					X				
Clinical discard**		?	?	?	?	?	?	?	?	?	?

*Collection of tissue from clinically diagnosed cancer precursor lesions at baseline and follow-up and cancers during follow-up, if diagnosed (see Section 7.3 Biospecimen Evaluations);
**Discard specimens from clinical collections, as permitted (see Section 7.3 Biospecimen Evaluations)

2 INTRODUCTION

2.1 STUDY RATIONALE

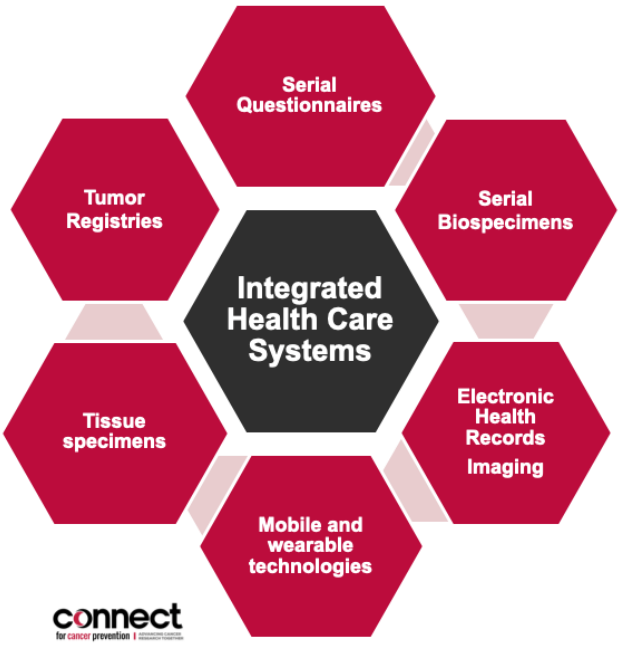
Because of the changing nature of most cancer-related exposures over time, **a prospective cohort study design** (i.e. an observational longitudinal study of individuals free of cancer at invitation) to determine exposures and how carcinogenic changes unfold over time (e.g., through changes in biomarkers and detection of early lesions), is essential to advance our understanding of the causes of cancer. Ascertainment of disease diagnoses, treatment courses, laboratory tests, and other endpoints in prospective cohort studies across the US is logistically very challenging and expensive because of the complexities of its health care system. **Integrated Health Care Systems** (IHCS) in the US, however, provide comprehensive healthcare to members through a range of coordinated health care facilities and services as well as storage of extensive healthcare information and biospecimens (e.g., clinical tissue specimens) on their members. Thus, this form of existing health system infrastructure provides a highly effective setting for the proposed **Connect for Cancer Prevention** cohort. Major advantages of IHCSs include the availability of comprehensive EHRs, facilitating a passive follow-up system (i.e., observing outcomes for patients undergoing usual standards of care within the IHCSs) for research that is cost effective with high levels of completeness, existing clinical infrastructure for specimen collection, and the long-term membership stability of their patient population. Indeed, based on the long history of successful collaborations between the National Cancer Institute (NCI) and IHCSs, this setting promises to offer an efficient and cost-effective platform to study a wide range of questions related to cancer etiology and other general research questions.

The central focus of Connect for Cancer Prevention is to understand the etiology and natural history of cancer to inform new approaches in precision prevention and early detection of cancer (1). This prospective cohort will incorporate recent developments in digital technologies, biomarkers and exposure assessments to advance the study of suspected and emerging factors that influence cancer development. These new developments include:

- Widespread use of electronic devices and applications in the digital age that facilitate engagement of study participants to provide and receive health information
- Opportunities for high-quality exposure assessment using innovative technologies such as wearable devices to measure behavior and environment
- Increased availability and quality of EHR and medical imaging for assessment of health conditions and medications
- New, cost-effective methodologies to interrogate the genome, epigenome, transcriptome, proteome, metabolome, microbiome and other biological processes
- Increased availability and quality of databases for data linkages, including pollution monitoring, pharmaceutical records and cancer registries
- Advances in molecular profiling of tumors and precursor lesions to study the natural history of cancer and etiologic heterogeneity across tumor subtypes

These technologies coupled with new methods in complex analytics of integrated high-dimensional data provide powerful tools to address the primary objectives of Connect for Cancer Prevention (see Section 3.1 Primary Objectives).

Key design features of Connect for Cancer Prevention to address a broad range of scientific questions related to cancer include (see Figure 1 Study Design Features for Connect):



- Large sample size to identify a sufficient number of participants expected to develop incident cancer.
- Serial exposure assessment and biological specimen collections to study changing exposures/biomarkers and evaluate critical exposure windows prior to disease diagnosis.
- Collection and storage protocols for biological specimens that are optimal for new assays and technologies.
- Longitudinal access to information from EHRs (both retrospectively for information prior to enrollment and prospectively after enrollment) to obtain information on medical conditions, screening and diagnoses of cancer and other diseases as well as prescribed and self-reported medications.

Figure 1. Study Design Features for Connect for Cancer Prevention

- Longitudinal (retrospective and prospective to enrollment) acquisition of non-cancerous, precursor or tumor tissue specimens to study the natural history and molecular subtypes of cancer as well as identify new prognostic and predictive biomarkers.

This state-of-the-art cohort will be built using an efficient, flexible, and integrated cloud-based infrastructure that utilizes modern interoperability standards to serve as a research workhorse for future generations of scientists. Although the primary endpoints of interest for this cohort are cancer incidence and mortality, the infrastructure will be designed to also enable collaborative studies of general research use and ancillary enhancements.

2.2 BACKGROUND

Over 1.8 million people in the US are expected to be diagnosed with cancer in 2020 and over 600,000 people will die of cancer-related causes, making it the second leading cause of chronic disease death, after cardiovascular diseases (2,3). Although treatments for cancer are improving

and there are promising advances through precision medicine, the total number of cancer diagnoses is expected to increase substantially over the next decade, in large part due to the aging of the population and behavioral/lifestyle changes (2,3). The risk of developing cancer depends on the complex interplay of multiple factors, including genes, age and gender, behavioral factors (e.g., diet, energy balance, physical activity, tobacco and alcohol), endogenous factors (e.g., hormones and growth factors), medication and drug use, infectious agents, and environmental factors (4). Research on the causes of primary and second cancers as well as prevention strategies are critical to reducing the burden of cancer in the population. Existing prospective cohort studies have greatly contributed to our understanding of cancer etiology and will continue to do so However, since most of them started many years ago, generally they are not based on a state-of-the-art data platforms, contemporary exposure information and repeated collection of biospecimens to provide full opportunity to address new questions using new approaches. The proposed cohort will address these limitations.

2.3 RISK/BENEFIT ASSESSMENT

2.3.1 Known Potential Risks

Connect for Cancer Prevention is an observational study that does not involve providing medications or medical interventions to participants. The direct physical or psychological risk to participants is minimal and limited to risks related to providing a biospecimen and disclosing sensitive data.

- **Biospecimen collection risks.** The most common physical risks associated with providing a blood sample is brief pain or bruising at the puncture site. The amount of blood drawn for research purposes at each visit, up to 48 ml, should not have adverse physiological effects nor lead to long-term distress and is allowable by the IRB. With venipuncture, less than 2% of people are expected to feel faint, nauseous, or dizzy, about 1 in 100 may experience a blood clot forming in their vein, and about 1 in 1,000 may have infection or significant blood loss (5). Risks for blood-borne pathogens from accidental needle sticks and during sample processing exist but are very rare.
- **Loss of privacy/confidentiality.** The main risk to participants, albeit very small though not zero, is a breach of privacy or a loss of confidentiality of a participant's personal identifying information (PII) and personal health information (PHI). During the baseline activities, a participant will be asked to provide identifying information including but not limited to name, social security number, and date of birth. This information is needed for tracking participants, particularly if they leave the health plan for passive linkage to national databases such cancer or mortality. We will collect data from the participant using Connect participant application ("app"). There is a risk to privacy whenever a participant uses an app or other online communication tools. Participating in the study poses no additional risk than use of other internet tools, particularly since Connect's backend is developed and audited in FedRAMP/FISMA compliant cloud infrastructure facilitated by NIH STRIDES program (datascience.nih.gov/strides). This oversight involves NCI IT team (CBIIT), which has ample experience with oversight of software engineering for field studies and clinical trials. Over

time the risk of re-identification becomes greater as there are more data sources to triangulate a participant's identity. During recruitment and active follow-up, potential and enrolled participants will have access to a help desk, administered by a support service contractor (currently National Opinion Research Center at the University of Chicago, NORC; FWA: 00000142), that will have access to participant PII and PHI.

- **Unknown risks.** Participants will be informed that the study could include risks that are currently unknown. When possible, we will inform the participant if new risks are identified that could affect their decision to participate.
- **Risks related to return of results.** From the baseline blood collection, we will offer return of genetic ancestry results using a graphical map display. We will inform the participants that this information has no medical value. We will also tell participants that because this sample is collected for research rather than clinical purposes, there is a risk of sample mix-ups, and that the results returned to the participant might not be their own. We expect the chances of this happening are small. Other non-medically actionable results, such as national recommendations on dietary and physical activity or results from biological assays (e.g., SARS-CoV-2 antibodies, cytokines, hormone levels), may also be returned to participants. With future blood collections, we will decide whether to return medically actionable results to participants . Before returning results, an amendment covering this activity will be submitted and reviewed by the IRB.
  Genetic and other biospecimen analyses and physical measurements included in the study could uncover abnormal values that may be medically actionable (see Section 7.3.2.3 Management of Results). Participants could experience stress as a direct result of receiving health measurements that could be indications of illness or be inconsistent with their understanding of their health status. Cost for emergency services and/or follow-up care associated with a value that is deemed to require medical attention will be the responsibility of participants. Connect for Cancer Prevention does not assume responsibility for fees associated with responding to any emergent or urgent situations for medical care or transportation associated with existing conditions uncovered during the course of participation in the research. However, we might suggest possible referrals at the cost of the participant.

2.3.2 Known Potential Benefits

Connect for Cancer Prevention has potential benefits to society as a research resource that will address important scientific questions that could inform preventive and early detection strategies to reduce the burden of cancer and other health conditions in the population. However, there are no direct immediate or long-term health benefits for participants. Return of any clinically relevant results will follow the procedures laid out in Management of Results (see Section 7.3.2.3) and specific protocols for return of results will be developed, reviewed, and approved by the IRB prior to returning any results to study participants. It is possible that there is a benefit to a participant if a research finding leads to clinical procedures that improves the participant's health. Potential indirect benefits to participants in Connect for Cancer Prevention include opportunities to:

- Learn about health indicators and possibly access their own data.
- Contribute to research efforts to improve effective, personal/precision healthcare, and develop screening approaches to prevent cancer and improve the health of future generations.
- Ensure that their community is included in research studies that could lead to new understanding of cancer etiology.
- Participate in future studies involving Connect for Cancer Prevention.

Similar to how volunteering for charities (e.g. at a local foodbank or Habitat for Humanity) can make individuals feel empowered to contribute to addressing problems impacting their community, this study should have an indirect benefit in allowing study participants to feel empowered to contribute to research aimed at preventing suffering and mortality associated with cancer.

2.3.3 Assessment of Potential Risks and Benefits

Throughout the duration of the study, the staff will work closely with the NIH IRB to mitigate risk of participation, including submitting any major changes to study procedures and participant communications to the NIH IRB for approval. Furthermore, the study staff will take important steps to minimize risks. Overall, the scientific value of the research for the benefit of society in general outweighs the minimal risk to the study participants.

- **Biospecimen collection:** Trained program staff or clinical phlebotomists will use standard sanitary biological specimen collection safety protocols for collection and processing of samples (e.g., antiseptics, gloves and appropriate clothing). All objects that come in contact with bodily fluids will be disposed of in appropriate biohazard waste containers. Whenever possible, the study blood draw will occur simultaneously with a clinically necessary draw to avoid a separate venipuncture and possible risk of pain, bruising, fainting, or infection. Should adverse events occur from the blood draw, clinic staff are trained and prepared to handle such situations. The local IHCS staff will document adverse events and the Connect for Cancer Prevention Senior Scientist will review them for any patterns and opportunities to further minimize risks.
- **Loss of privacy/ confidentiality:** Processes and procedures for access to participant data and biospecimens will protect privacy and confidentiality (see Section 9.11 Publication and Data Sharing Policy) and Section 9.1.1 Consent and Assent Procedures and Documentation).
    - Most participants will complete the consent and questionnaires in a private location of their choosing. If participants are recruited during a clinical visit, they will not be required to complete study questionnaires on site to protect their privacy.
    - Initial or long-term participation will have no consequences on the individual's relationship with their health care insurer(s) or providers.
    - Health Insurance Portability and Accountability Act (HIPAA) identifiers will be separated from all research data in both paper and electronic forms whenever possible (see Section 9.1.1 Consent and Assent Procedures and Documentation) following

strict data security measures in a Federal Risk and Authorization Management Program (FedRAMP) compliant environment to minimize any possible breaches of confidentiality.

- Each study participant will be represented by unique ID for each research project proposal and no PII will be shared with collaborating researchers (see Section 9.11 Publication and Data Sharing Policy).
- Administrative activities (such as linkage with state or national cancer registries) will require sharing of PII (see Section 9.1.1 Consent and Assent Procedures and Documentation) using least privileged principles.
- Connect for Cancer Prevention is covered by the NCI Certificate of Confidentiality. As outlined in Section 2012 of the 21st Century Cures Act, the study staff will protect participants' privacy and use all available legal measures to oppose requests for disclosure of data and ensure that anyone using or in possession of copies of the data is prohibited from disclosing them. However, if data are disclosed for any reason, that information is inadmissible in any legal, administrative or other proceeding. Disclosure is permitted only when required by Federal, State or local laws and in compliance with applicable Federal regulations governing human subjects research. PII/PHI will only be released to third parties as outlined in NOT-OD-17-109. Decisions about the circumstances under which to release PII to third parties will be made by the IRB.
- The help desk, maintained by the support services contractor, will track requests using an electronic ticketing system. All computers and servers containing PII and PHI will be compliant with the latest regulations set forth by the US Government Configuration Baseline (USGCB) and/or other approved United States Department of Health & Human Services (HHS) Information Technology (IT) Security Configurations. Data in cloud-based applications will be stored in a FedRAMP compliant environment. The system will be compliant with the National Institute of Standards and Technology (NIST) Special Publication 800-53 (Recommended Security Controls for Federal Information Systems and Organizations, Moderate level) and has knowledge and experience in numerous laws, regulations and standards including, but not limited to: Office of Management and Budget (OMB) Circular A-130, OMB memorandum (M-05-24), and Federal Information Processing Standard (FIPS) Publication 201. It will also comply with the NIH configuration standards and policies for storage and transfer of HHS data, including the security to access information, encrypt data for transfer as needed, store sensitive data, and remove data once the project is complete.

- **Return of Medically-Actionable Results:** An IRB amendment will be submitted to detail return of medically actionable results for future biospecimen collections to participants (see Section 7.3.2.3 Management of Results). The plan will outline how to educate participants about the risk and benefits of medically-actionable results, the procedure to confirm results in

a Clinical Laboratory Improvement Amendments (CLIA) environment and appropriate quality assurance procedures, and other procedures.

3 OBJECTIVES AND ENDPOINTS

OBJECTIVES	ENDPOINTS	JUSTIFICATION FOR ENDPOINTS
Primary		
The **overarching objective** is to recruit 200,000 adults aged 40-65 years into a prospective, multi-center study within IHCSs in the US using a powerful digital platform. The **primary objective** is to study associations of endpoints in the cancer continuum with multilevel factors, including behavioral, biological, environmental, medical and social factors.	Cancer transformation, incidence, progression, and mortality endpoints could include, but are not limited to: • early biological effects (e.g., inflammation or metabolomic markers) related to cancer • intermediate biomarkers • cancer precursors • cancer incidence • cause of death • cancer survival, risk of second cancers and survivorship	These endpoints will facilitate inferences on **cancer etiology and natural history** to provide insights into carcinogenic processes and inform new approaches **in risk assessment** and **early detection of cancer** (see below for additional detail). Studying the cancer continuum will improve translation of epidemiologic findings to public health and clinical practice.
Secondary		
The **secondary objective** is to create a publicly available resource of deidentified data and biological sample repository to registered users for general research use.	The endpoints are numerous and could include methodological research, human biology, ancestry, evolution or health-related outcomes.	The Connect for Cancer Prevention is a rich resource that will benefit the progress of science for years to come.
Tertiary/Exploratory		
Not applicable		

3.1 PRIMARY OBJECTIVES

Etiology of Cancer: This prospective cohort will enable studies to identify and characterize biological, behavioral and (broadly defined) environmental risk factors, and the interactions among them associated with the incidence of different cancers. Specifically, serial exposure and biomarker assessments in individuals prior to any cancer diagnosis will allow the study of how changes over many years could influence cancer development and to identify critical exposure periods. Examples of research topics in this area include:

- Evaluate current and emerging behavioral and environmental factors (e.g., vaping, physical activity patterns) in relation to cancer risk and progression.
- Evaluate possible links between commonly used pharmaceuticals and medical procedures (e.g., opioid use, gastric bypass surgery) with subsequent cancer risk.

- Investigate the relationship between exposure to outdoor air pollution (including at the home and during the commute) and cancer risk, based on linkage of the residential address to modeled estimates of air pollutants derived from Environmental Protection Agency's air quality monitoring network.
- Investigate the relationship between ingested drinking water contaminants and cancer risk, based on linkage of the residential address and self-reported public drinking water utility to regulatory monitoring data for water contaminants.
- Characterize biomarkers of exposure (e.g., pollutants) and early biological effects (e.g., inflammation markers, metabolomic markers) to determine distribution in the population, sources of variation and changes over time.
- Identify biomarkers of susceptibility and early biological effects in tumor initiation to characterize biological processes including the genome (e.g., germline genetic susceptibility factors), epigenome, transcriptome, proteome, metabolome, microbiome, and immune response as well as their interactions with behavioral, contextual and environmental factors.
- Identify molecular signatures in cancers and precursors associated with behavioral, contextual and environmental factors as well as germline-somatic associations to provide insights into mechanisms of carcinogenesis.
- Characterize etiologic heterogeneity for different tumor subtypes using state-of-the-art imaging and molecular characterization of tumor and relevant tissues.

Natural history from precursor to tumor transformation: While cancer initiation is typically not observable, precursor lesions detected through population-based screening programs or clinical symptoms have been identified for certain cancer types (e.g., breast, cervical, colorectal cancers). Some cancer precursors are typically removed at detection to prevent progression to cancer (e.g., cervical cancer precursors and colorectal polyps). Others (e.g., hematological cancer precursors, atypical hyperplasia of the breast or the uterus or anal cancer precursors) can be left and observed through clinical surveillance. This allows for a direct observation of the possible transition from precursor to early stage cancer that can be detected and treated with improved prognosis. Clinical imaging data (e.g., ultrasound of the breast from screening programs or symptomatic conditions) can also detect abnormalities that represent cancer precursors. For many cancer sites, precursors are not well-characterized and typically are not detected widely.

- Study biomarkers of tumor transformation and progression, including early postzygotic mutational events (e.g., clonal mosaicism, clonal hematopoiesis), high-risk molecular changes in precursor lesions (e.g., benign breast disease, colon polyps, cervical cancer precursors, lung nodules, esophageal metaplasia/ dysplasia, endometrial hyperplasia, urothelial proliferation/dysplasia).
- Identify precursors for cancers with suspected or unknown precursors (e.g., monoclonal gammopathy of undetermined significance (MGUS) and hematological cancers).

- Identify determinants of cancer precursors (e.g., behavioral, contextual, environmental and genetic) to inform natural history studies and models for predicting clinically relevant cancer precursor.

Cancer Risk Assessment: Determining individual risk of cancer is important for public health (e.g., population-based cancer screening programs) and individual counseling for cancer prevention and early detection strategies. During the last decade, major progress has been made in the development of statistical methods and the application of risk prediction models of various cancer sites. Cancer risk prediction models often include a core set of well-established epidemiological risk factors. Inclusion of additional risk factors, including biomarkers and clinical parameters (e.g., comorbid conditions, blood test results, imaging data), and accounting from time-varying exposures could improve their calibration and discriminatory accuracy. The field now is moving from risk models for individual cancers to developing joint models for multiple cancer endpoints. The Connect for Cancer Prevention study resource will be instrumental in the development and validation of risk prediction models for precision prevention through the integration of risk factors from multiple sources (e.g., questionnaires, electronic health records, imaging, personal monitors, germline genetics, and blood/urine/tissue biomarkers), accounting for changes over time and more precise definitions of endpoints (e.g., molecular subtypes of cancer). Integral to these risk prediction efforts, the study questionnaire has been developed to provide relevant risk factor information for many cancer endpoints. As new risk factors are identified, they can be integrated into follow-up questionnaires to allow for continuous updating and improvement of cancer risk prediction models.

Early Detection of Cancer: Early detection of cancer promises to improve cancer outcomes through detection of early-stage cancers and identification of precursors that can be treated to avoid development of invasive cancers. Currently, only very few early detection approaches (e.g., blood biomarkers or imaging) exist that have demonstrated improved cancer outcomes in clinical trials. Repeated biomarker measurements using specimens collected prior to diagnosis of cancer from study participants will provide extensive opportunities to identify biomarkers (e.g., mutations in pre-diagnostic circulating tumor deoxyribonucleic acid (DNA)) and develop algorithms for the early detection of cancer based on the trajectories of high-dimensional biomarkers. Studies of early detection marker development can be conducted in Connect for Cancer Prevention, including marker discovery, assessment of clinical performance for detection of disease, evaluation of detection windows and lead time for specific markers and cancer sites, as well as comparative studies of cancer detection for several early detection approaches.

Although Connect for Cancer Prevention will be able to address a wide range of scientific questions, as for any epidemiological study, replication of findings in independent study populations and pooling data with other studies will be essential to confirm findings or address questions that require larger sample sizes (e.g., for the study of rare cancers or exposures). To this end, the cohort has been designed to facilitate data sharing and future data pooling efforts.

3.2 SECONDARY AIMS

The Connect for Cancer Research cohort will be a large and comprehensive data resource and biospecimen repository for general research use.

4 STUDY DESIGN

Connect for Cancer Prevention is a multi-site, prospective cohort study with the goal of enrolling 200,000 adults (see Section 8.2. Sample Size Determination) free of cancer who are patients or members of participating IHCSs in the US (see Section 5.1 Inclusion Criteria). Due to the long natural history of cancer, we anticipate following participants for life. Participants will have consented to passive study activities and to participate in active study activities at baseline and regular intervals throughout follow-up (see Figure 2 Participant Workflow for Connect Cohort). During follow-up, other ancillary studies also might be proposed to participants. Study activities and communication with participants will be facilitated by a secure, web-enabled participant application ("participant app").

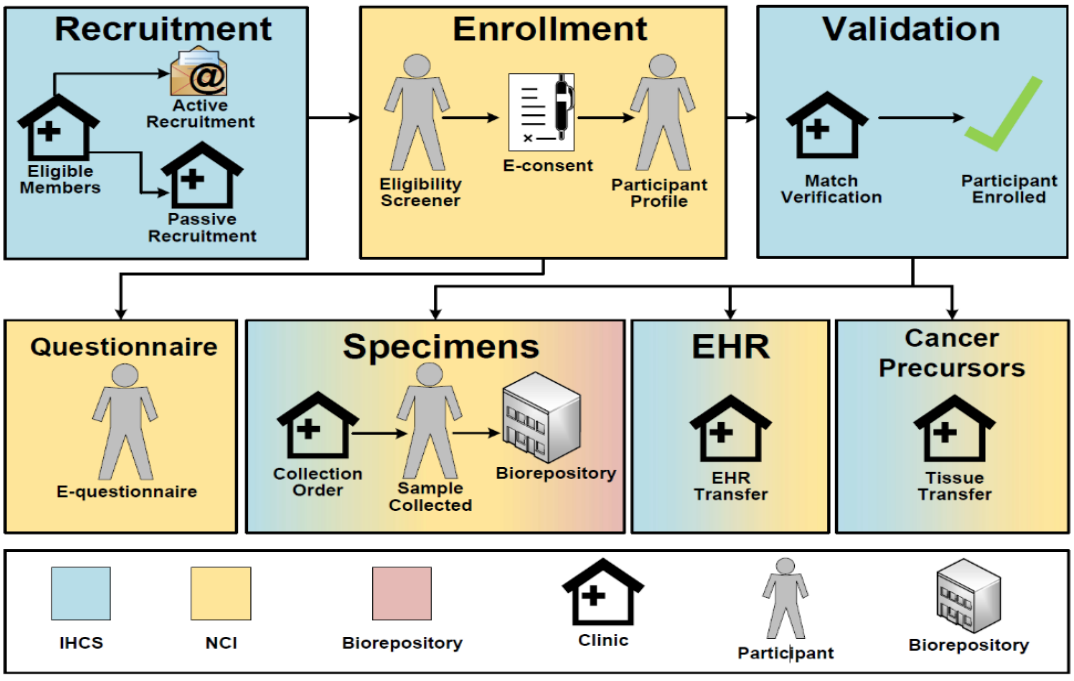


Figure 2. Participant Workflow for Connect Cohort

- **Baseline Activities**. After providing informed consent, the study participants will be asked to complete the following baseline study procedures:
    - Participant profile and match verification (see Section 5.6 Strategies for Recruitment and Retention)

  - Baseline questionnaire divided into four modules to capture behavioral, medical, residential history, social and identifying information (social security number will be requested but not required) (see Section 7.2 Clinical Evaluations)
  - Blood, urine and mouthwash specimen collection and associated questionnaires (see Section 7.3 Biospecimen Evaluations)
  - HIPAA authorization for the collection of EHR and other IHCS data (see Section 7.2 Clinical Evaluations) and acquisition of tissues for prevalent cancer precursors (see Section 7.3 Biospecimen Evaluations)
- **Passive Follow-up Activities**:
  - EHR updates – EHR data will be released by the IHCS and any other health care providers to NCI at least once a year as long as the participant remains in the IHCS (see Section 7.2 Clinical Evaluations). Cancer, mortality, and other health outcomes, as well as exposure information (e.g., weight), will be ascertained from the EHR.
  - Clinical tissue specimen collection – Precursor and cancerous lesions will be requested from the participating IHCS or affiliated facilities.
  - Data linkages – For all participants (including those that leave the IHCS), personal identifiable information, including name and SSN, will be linked to data from US state cancer registries and the National Death Index (NDI) to ascertain endpoints, including cancer and vital status. Linkage with state registries will be facilitated, where applicable, by the Virtual Pooled Registry Cancer Linkage System (VPR-CLS), an initiative from the North American Association of Central Cancer Registries (NAACCR) and the NCI Division of Cancer Control and Population Sciences to provide a central and streamlined application process for cancer surveillance data linkages. Other data linkages may include geocoding to environmental monitoring databases and demographic and health data, such as Census, Medicare, Human Immunodeficiency Virus (HIV) registries, pharmacies, and imaging centers.
- **Active follow-up activities**:
  - Serial questionnaires on changes in health status, endpoints as well as existing and new exposures, including quality of life and lifelong cancer screening behaviors (see Section 7.2 Clinical Evaluations).
  - Serial biospecimen collections from time to time as needed (see Section 7.3 Biospecimen Evaluations).
  - EHR updates – For participants who leave the IHCS, we will request that they provide the EHR data through the participant application.
  - Additional assessments – Participants will be asked to use their own devices and/or be provided with a device, such as wearable devices or applications for mobile devices. During follow-up other possible technology-assisted assessment tools may include those that measure heart rate, pulse oximetry, ambient air pollutants, and other assessments.

- **Ancillary activities.** During follow-up, ancillary efforts may include collection of biospecimens or measurements not specified in this protocol. Any ancillary studies will be submitted to the IRB for review and approval. Possible examples include:
    - Other biospecimen collections (e.g., cerebral spinal fluid, cervical swab)
    - Intensive sampling from a targeted subset of study participants
    - Physical measurements such as waist and hip circumference
    - Additional questionnaires (e.g., workplace and environmental exposures)

4.1 CONCOMITANT THERAPY

This section is not applicable to our observational study.

5 STUDY POPULATION

5.1 INCLUSION CRITERIA

Due to the minimal risk nature of this protocol, all individuals interested and able to participate in Connect for Cancer Prevention, who meet the eligibility criteria and are not specifically excluded, will be able to participate. In order to be eligible to participate in this study, an individual must meet all of the following criteria:

- Patients or members of participating IHCS at the time of enrollment
- Age between 40 and 65 years old at study invitation

5.2 EXCLUSION CRITERIA

An individual who meets any of the following criteria will be excluded from participation in this study:

- Individuals with a history of invasive cancer (other than non-melanoma skin cancer)
- Individuals with known cognitive impairment documented in their medical record

5.3 INCLUSION OF VULNERABLE PARTICIPANTS

This section is not applicable to our study. Children are not included in Connect for Cancer Prevention because cancer incidence is very rare.

5.3.1 Participation of Employees

NIH employees will not be specifically targeted for enrollment; this section is not applicable. The recruitment strategies of the IHCS may include recruitment of IHCS employees. If so the IHCS investigators will address specific procedures and safeguards for their recruitment in their site amendments.

5.4 LIFESTYLE CONSIDERATIONS

This section is not applicable to our study.

5.5 SCREEN FAILURES

This section is not applicable to our study.

5.6 STRATEGIES FOR RECRUITMENT AND RETENTION

- Enrollment in Connect for Cancer Prevention is voluntary and not time sensitive.
- **Enrollment sites**. Participants will be identified and recruited by participating IHCS and followed up by the IHCS and the Connect Coordinating Center at DCEG/NCI through the participant application (Section 5.2.1 Human Data Sharing Plan and Section 9.3 Confidentiality and Privacy). Although more sites may be added in the future, four initial awards have been established to set up recruitment at nine individual IHCSs (see Figure 3 Location of DCEG Coordinating Center, Participating IHCS and Support Services and Appendix 1 Study Governance Structure):
    - Henry Ford Health System with HealthPartners and Marshfield Clinic Health System
    - Kaiser Permanente Colorado with Kaiser Permanente Georgia, Kaiser Permanente Hawaii and Kaiser Permanente Northwest
    - Sanford Health
    - The University of Chicago
- The support service contract will assist with field activities, such as developing Standard Operating Procedures (SOPs) and the manual of operating procedures (MOOPs), as well as training research staff to implement the MOOPs. The support services contractor will also provide help desk support, conduct linkages and aid efforts to follow participants who leave the IHCS.

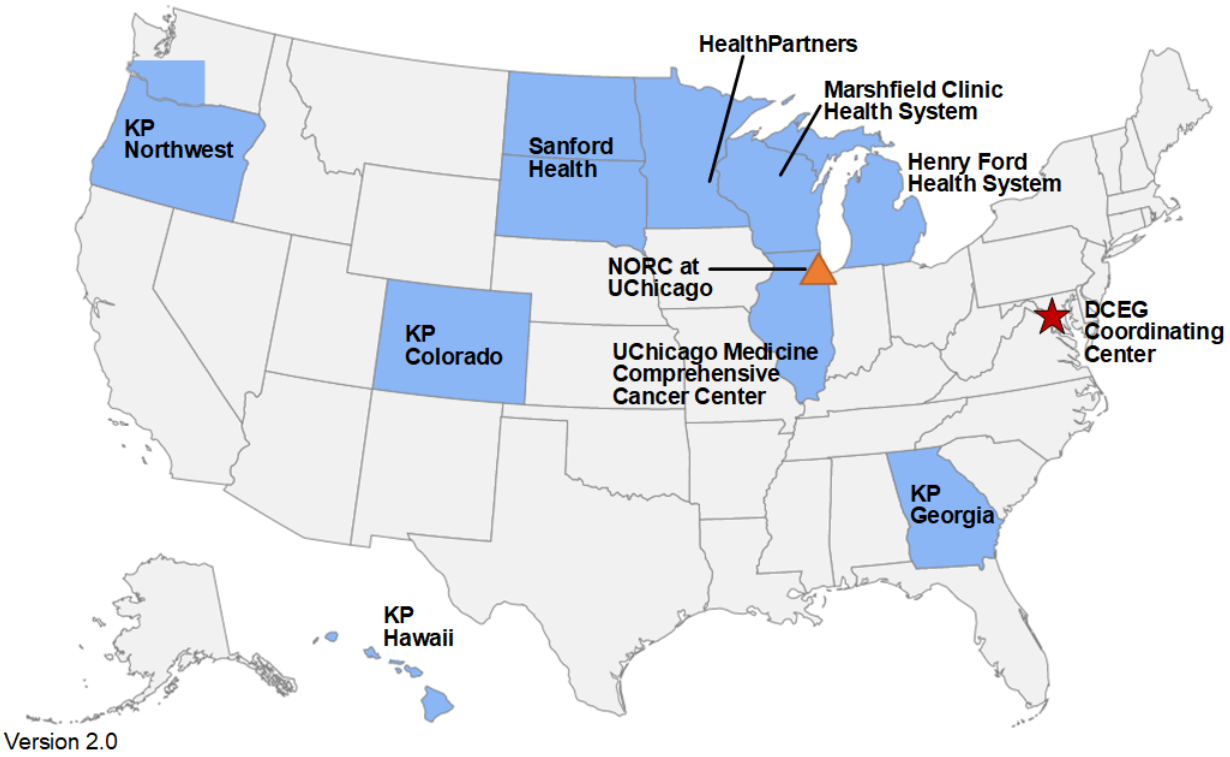


Figure 3. Location of DCEG Coordinating Center, Participating IHCS, and Support Services

- **Recruitment targets**. Participants (n=200,000) will be recruited on a rolling basis for at least five years, depending on recruitment proportions. It is expected that 10% of recruitment targets will be met in year 1 and the remaining individuals recruited evenly over the remaining four years.
- **Participant Application**. A secure, web-enabled progressive application is being developed by DCEG to provide study activities and communication to participants.
- **Study Websites**. The study will deploy a public-facing and participant-facing website that may include information (final content will be shared in an IRB amendment), such as:
    - Branding and videos about the study
    - Answers to frequently asked questions
    - Messages from NCI and IHCS study leadership
    - Testimonials from self-identified participants and community leaders
- **Recruitment materials**. The recruitment materials will convey a main value proposition, "Help prevent cancer for tomorrow", with defined variations on specific language and images to accommodate different target populations (e.g., male, rural participants). All recruitment materials will be submitted to the IRB for approval prior to dissemination. The Connect for Cancer Prevention Patient Advisory Board, ad-hoc focus groups, and/or stakeholders from the IHCS will be consulted on development of recruitment language and materials. The IHCS sites will use IRB-approved recruitment language and images with modifications for co-branding to maintain a consistent messaging and look and feel of all study materials.
- **Enrollment Workflow**. The enrollment journey for potential participants in Connect for Cancer Prevention, as currently envisioned, is illustrated in Figure 4. Description of Enrollment Workflow and Data Collection Pre- and Post-consent. Initial outreach to individuals will occur through two strategies: active and passive recruitment.
- **Active recruitment outreach**. The primary recruitment strategy will be active recruitment of individuals who are pre-screened for eligibility by the IHCS (see Section 7.1.1 Screening Activities Performed Prior to Obtaining Informed Consent). Eligible individuals will be identified by each of the IHCSs based on the study inclusion and exclusion criteria (see Section 5.1 Inclusion Criteria and Section 5.2 Exclusion Criteria) using verified algorithms. Eligible individuals will be selected for invitation to participate using different strategies based on criteria appropriate to each IHCS, including, but not limited to:
    - Demographic distribution of target population
    - Upcoming clinical appointments
    - Cancer screening schedule
    - Past diagnosis of cancer precursors
    - Geographic proximity to IHCS clinics
    - Employees of some of the IHCS, following the local requirements (see Section 5.3.1 Participation of Employees)

- Participation in existing research projects at the IHCS or who have indicated an interest in research via the IHCS

The IHCS study staff will send a study invitation that will contain a unique token or Personal Identification Number (PIN) and a link to the secure, participant-facing study website, hosted by NCI. The study website will display information about the study objectives, requirements, eligibility criteria, and list of participating IHCS.

In the pre-consent step, interested individuals will be asked to create a secure account (i.e., log in) on the study website before being directed to the consent process (see Section 9.1.1 Consent and Assent Process and Documentation). IP addresses and account information will not be stored by NCI prior to consent.

- **Active recruitment outreach of underrepresented populations**. Efforts will be made to recruit IHCS members and patients who are in underrepresented populations within the health system's catchment areas (e.g., non-white individuals, living in medically underserved areas). Such efforts can include research coordinators traveling to clinics in underserved areas, tailored recruitment materials and community outreach activities.
- **Active recruitment, pre-consent reminders**: The IHCS staff will send reminders to a potential participant after receiving an invitation to join Connect for Cancer Prevention. Up to 10 reminder contacts will be made defined based on the mode of communication and the IHCS policy. The number of reminders will allow the time, days, and modality to vary and ensure a successful contact attempt. Modes of approaching potential participants for recruitment will be any combination of:
  - Email
  - Phone call
  - Patient portal
  - Postal mail
  - Text (Short Message Service (SMS) and other messaging services)
  - Interactive Voice Response (IVR)
  - In person contact
  - Telemedicine/virtual clinical visits

  During the enrollment process, potential participants will be asked how they found out about the study in the participant-facing MyConnect app and/or when speaking to study staff. We will use these data for internal purposes to review recruitment materials as well as for scientific publications on recruitment strategy outcomes (i.e., enrollment success) overall and by age, race, and geographical region (via deidentified, categorized data provided by the sites).
- **Active recruitment opt out.** Eligible individuals invited as part of the active recruitment strategies can opt out of the reminders in at least three ways (1) during a phone call reminder from site staff, (2) by calling the site or Connect Support Center themselves or (3) by completing a form on the Connect Support Center website. The study staff/interviewers will ask disinterested individuals why they want to opt out of the study recruitment. During the conversion, the study staff/interviewer will attempt to identify, and code opt out reason(s).

The form on Connect Support Center website (linked provided in the recruitment materials) will also query reasons for opting out. Reasons for opting out data will be used to analyze reasons participation might vary by key demographic factors, including age, race, and geographic region. Results from these analyses will be used for internal purposes to revise recruitment materials and for scientific publication purposes.

- **Passive recruitment outreach.** The secondary recruitment strategy will be self-referral of IHCS members and patients. Strategies will be employed to cast a wide net for passive recruitment and direct interested people to the study website to view eligibility criteria, enroll in the study or contact study staff by email or phone for more information. Modes of recruitment may vary by IHCS site and may include:
    - Written advertising within IHCS facilities (e.g., posters, brochures, flyers, banners)
    - Video on IHCS display screens at the IHCS sites and before telemedicine visits
    - Voice recordings before telemedicine visits
    - Social media
    - Radio, TV, print, web advertising
    - Physician, patient and community ambassadors and referrals to the study
    - Tabling at community events, including those hosted by a participating IHCS
    - Announcements in IHCS newsletters/websites
    - "Snowball" participation (formally inviting family and friends of participants)
    - Informing employees via study screensavers, newsletters and employee events
    - Word of mouth
    - Placing information in after-visit summary (e.g., dot.phrase in Epic Systems Corporation (Epic)) given to patients at the end of a clinical visit

  Passive recruits who self-refer into the study can contact study staff or go directly to the public-facing study website to join the study. Similar to active participants, potential participants will be asked how they found out about the study.

- **Participant profile**. After obtaining informed consent (see Section 9.1 Informed Consent Process) and HIPAA authorization, consented individuals will be asked to complete the participant profile on the participant application to verify the identity by the IHCS (i.e., match verification), to maintain contact with participants throughout the study and to trace and link to health registries (such as cancer registries and the National Death Index (NDI)), which are essential to ascertain the primary endpoints (incident cancers, vital status, cause of death) and other health outcomes for this study. The participant profile includes:
    - Name (first, middle, last)
    - Date of birth
    - Sex at birth
    - Contact information
        - Preferred and alternative E-mail
        - Home and/or mobile phone numbers with voicemail and text preferences
        - Primary mailing address
    - Previous cancer diagnosis (yes/no; type and year of diagnosis) – EHR will be the gold standard for determining study eligibility criteria.

Recruitment Workflow			
Who Contacts Participant	*Data Storage and Access*	*Data Elements Shared by Participant*	*Data sent From IHCS to NCI*
Pre Consent			

IHCS send recruitment invitations and reminders	IHCS stores and accesses identifying information	• No data stored from login
		• Site affiliation
		• How did you hear about Connect
		• Token or PIN
			• De-identified data:
			  • Study ID
			  • Indication of active recruitment group (e.g. random selection, upcoming appointment)
			  • Age or age group
			  • Race/ethnicity
			  • Sex

Time of Consent			
IHCS send reminders	• NCI stores name, date, and time of consent • IHCS can access this information	• Name • Date and time of consent	
Post Consent			

NCI sends reminders	• NCI stores participant profile data
	• IHCS can access this information
		• Participant profile
		  • Name
		  • Date of birth
		  • Sex at birth
		  • Cancer history
		  • Email
		  • Phone
		  • Mailing address
		  • Login information (Google account, email, phone)
		• Baseline questionnaires	

Time of Eligibility Verification			
IHCS contacts participants as necessary	Both IHCS and NCI can store and access this information	Corrections to participant profile if needed	• Corrections to participant profile if needed • Eligibility verified (yes/no)
Post Eligibility Verification			
• NCI sends notification about verification status and reminders to complete study components • IHCS contacts participants to schedule biospecimen collections	NCI stores data and IHCS can access this information	• Schedule biospecimen collections • Biospecimen questionnaires after collection • Follow up questionnaires	EMR data and other IHCS data

Version 2.0 *Captured from login information

Figure 4. Description of Enrollment Workflow and Data Collection Pre- and Post-Consent

- **Biospecimens and EHR data.** These will only be collected after IHCS match verification, though consented individuals who were active recruits will be able to complete their baseline questionnaire modules prior to match verification.
- **Data collection for passive recruits**. Consented individuals who enroll through passive recruitment methods will undergo match verification prior completing any study activities including baseline questionnaires. When individuals establish their profiles, they will be reminded that non-verified individuals will be deemed ineligible, such as those who cannot be verified as patients or members of the IHCS.

- **Post-consent reminders**. Reminder contacts, after consent, will be made primarily through digital contact (e.g., SMS text messages, email, patient portals, push notifications through the participant application) by DCEG using the participant application. If these attempts fail, the IHCS staff could also follow-up with telephone calls, depending on IHCS site restrictions and prior successes. Because it is difficult to ascertain whether individuals received, opened, and read the digital contact, up to 10 reminders contacts will be made for each component of the study. The number of reminders will allow the time, days, and modality (e.g., email, mail, text, patient portal, phone) to vary and ensure successful contact attempts.
    - For the user profile, reminders will be sent over the first year until participants complete the user profile.
    - For baseline activities, every attempt will be made to bundle reminders across activities (e.g., one reminder during week 2 would prompt the participant to complete the remaining questionnaire modules and send in the mouthwash specimen) to minimize the number of reminders sent to the participant.
- **Study subject status**. Individuals who do not complete the user profile within one year will be considered ineligible because they will not have sufficient data to go through match verification. Participants will be considered minimally enrolled if they provide consent and HIPAA authorization, complete the user profile and can be verified by their IHCS, but do not complete baseline activities. Fully enrolled participants are minimally-enrolled participants who complete all baseline questionnaire modules (see Section 7.2 Clinical Evaluations) and blood, urine, and mouthwash specimen donation (see Section 7.2 Clinical Evaluations). Linkage with EHR and other IHCS data and acquisition of tissue for prevalent cancer precursors will be obtained for minimally- and fully-enrolled participants.
- **Long-term retention strategies**. Long-term study engagement is vital to the success of this longitudinal study. The Connect for Cancer Prevention Participant Advisory Board (see Appendix 1 Study Governance Structure) will provide guidance on engagement and retention strategies. Each IHCS will employ different strategies based on institution-specific restrictions and experience with their patient population. The activities may include the following:
    - Community engagement at the IHCSs
        - Research lectures and seminars for community members
        - Table at health fairs, block parties, festivals, etc. with printed and live educational content focused on strategies to improve overall health (e.g., magazines/handouts, live cooking demonstrations and tasting using healthy, community-tailored recipes, stretches or exercises, etc.)
        - Highlight self-identified participant stories in different settings/media, including posts or videos on the study website, newsletters, social media or posters
    - Leverage patient-physician relationship supported by
        - Videos of IHCS physicians endorsing study
        - Printed materials (e.g., invitation letter with signature of patient's physician)

        - Physician education about study to support their endorsement of the study during patient encounters
    - Participant feedback survey of enrollment experience sent out to participants after consent and biospecimen collection
        - Survey participants about motivation and value for joining the study
        - Follow-up contact immediately after baseline visit with an email or card to express appreciation to participants and may include messaging on the longitudinal nature of Connect for Cancer Prevention
    - General participant communication
        - Branded study materials, referral cards for friends and family, information regarding next steps, promotion of study website, application and other ways to stay connected
        - Newsletters highlighting study updates, achievements, lay summaries of research findings, descriptive overview of the cohort and other relevant information
        - Enrollment anniversary, birthday and holiday cards
        - Study website and participant application updated with general health, screening and other cancer-related information
        - Snap (short, quick-response) questions to attract participants to the study application to view and complete study-related activities (e.g., questionnaire modules, digital health technology activities). These might include multiple-choice poll questions on a health- or wellness-related topic posted at regular intervals or a brief fact summary sourced from trusted, public domain sources (such as National Library of Medicine, NIH institutional websites, etc.)
    - Participant return of results
        - Relevant information obtained from participant's response to questionnaires (e.g., dietary/nutrient intake, physical activity) compared to the cohort distribution and expert consensus recommendations, and accompanied with referral information
        - Biological assay results, including but not limited to genetic ancestry (see Section 7.3.2 Samples for Genetic/Genomic Analysis)
- **Updating contact information**. Long-term efforts to maintain updated contact information on all participants will be done through the secure participant app via app notifications, emails, or SMS text messages. If an individual does not respond to internet-based contact, attempts could be made to contact them by regular mail or telephone. Contact information could also be updated from data pulls from the IHCS EHR searches of public listings or commercial databases such as LexisNexis Accurint using name and social security number.

5.6.1 Costs

This section is not applicable to our study.

5.6.2 Compensation

Monetary and non-monetary compensation will be provided for completing study activities. At baseline, monetary compensation will include parking reimbursement (where applicable) plus

compensation of $25 after completion of the four questionnaire modules and donation of the blood sample. During each follow-up contact for biospecimen collection, compensation of no more than $25 will be provided. In addition, participants might be given the option to receive non-monetary compensation (e.g., data summaries from questionnaires such as diet composition) or branded study materials (e.g., pens, pins, tote bags).

6 PARTICIPANT DISCONTINUATION/WITHDRAWAL

6.1 PARTICIPANT DISCONTINUATION/WITHDRAWAL FROM THE STUDY

With the minimal risks of this observational study, withdrawals due to reportable events are not expected. If one should occur due to an adverse event, the participant discontinuation or withdrawal from the study will be recorded on the Case Report Form (CRF). The majority of withdrawals are expected to be those described below and will be submitted to the IRB as part of the Continuing Review process.

Participant discontinuation from the study can be due to voluntary participant withdrawal and/or revocation. It should be noted that existing data and biospecimens from deceased participants will continue to be available for future research as death is an endpoint in this study.

- **Participants request for withdrawal**: Participants will be able to request to refuse to continue with the active study activities and withdraw from the study at any time by mail, email or verbal communication to study staff (e.g., IHCS staff, help desk). At the time of withdrawal, the study staff will inquire about the reason for withdrawal and explain to the participant that already collected data and biospecimens will be retained and analyzed even after withdrawal. No further data will be collected, and no additional contact will be made with the withdrawn participant. Withdrawal status will be documented in the study's record system. If a participant wishes to re-join the study, they will be able to do so after providing a new consent.
- **Revocation**: Participants who choose to withdraw from the study research and also revoke authorization in writing for continued use or disclosure of their personal health information that was already obtained in the research, analysis of that personal health information will only continue to the extent necessary to protect the integrity of the study. Previously collected data and biospecimens stored at the NCI central repository will be destroyed (see Section 7.3 Biospecimen Evaluations) and excluded from future use; however, data or biospecimens that were already distributed to other institutions for approved research use will not be able to be retrieved.

6.2 LOST TO FOLLOW-UP

Every attempt will be made to obtain updated longitudinal data and study endpoints with long-term follow-up of study participants via the secure participant application. Even participants who leave the IHCS will continue to be followed through serial contacts, questionnaires, and collection of biospecimens (e.g., serial blood specimens) whenever possible (see Section 7.3 Biospecimen Evaluations). Participants who no longer respond to the secure participant application, email, text messages or letters and are still members of the IHCS may be telephoned

by IHCS study staff using updated information from the EHR (Section 5.6 Strategies of Recruitment and Retention). Existing data and biospecimens from lost to follow-up participants will continue to be available for future research.

Despite the study staff's best efforts, there will be participants lost to follow-up given the long duration of the study. Depending on the statistical analysis, different approaches can be implemented to handle missing data or loss to follow-up (i.e., individuals for whom we cannot determine cancer or vital status via active or passive follow-up). An example is multiple imputation, which is a general-purpose methodology that can be applied. In some situations, loss-to-follow-up could be related to the underlying biological processes. In these cases, the probability of leaving the IHCS or having missing serial information could be related to an individual's risk of developing study endpoints. Shared random parameter models or pattern mixture models can be used to explicitly account for this type of missing data mechanism. For any planned analysis, the analysis plan should include a detailed description of how missing data or lost to follow up will be addressed.

7 STUDY ASSESSMENTS AND PROCEDURES

7.1 SCREENING PROCEDURES

7.1.1 Screening activities performed prior to obtaining informed consent

A HIPAA Waiver of Authorization will be sought by the IHCS to screen for potential eligible study participants and conduct recruitment outreach prior to their informed consent. Participant screening and recruitment activities might include:

- Email, written, in person or telephone communications (e.g., phone call or text message) with prospective subjects (to cease when an individual has been offered participation and refused)
- Review of existing administrative information (e.g., appointment history, upcoming visits) and/or medical records (e.g., medical history, physical examination, laboratory studies)
- Review of local tumor registry and related accession lists
- Review of existing MRI, X-ray, or CT images
- Review of existing pathology specimens/reports from diagnostic specimens

These activities will be performed exclusively by the IHCS who have access to their patient/member information, in accordance with the HIPAA Waiver of Authorization and their privacy policies, thus minimizing any risks to the individuals. To accomplish this task, the IHCS will develop and deploy an algorithm to identify members or patients meeting the eligibility criteria. The algorithm may include existing EHR data, tumor registry data, pathology data and other existing data sources (e.g., claims data, mortality data) to identify potentially eligible participants and exclude those with a personal history of invasive cancer.

7.1.2 Screening activities performed after a consent for screening has been signed

This section is not applicable to our study.

7.2 CLINICAL EVALUATIONS

- **Physical examination**. In follow-up activities, participants might be asked to have anthropometric measurements taken in the clinic or at home.
- **Radiographic or other imaging assessments**. Archived images performed in the past or future as part of the participant's medical care will be collected as part of the study. Currently, the study protocol does not include *de novo* imaging. Any future plans will be submitted to the IRB prior to the initiation of study activities.
- **Questionnaires**. Questionnaires will be completed on the participant application. They will be designed to be completed on any web-enabled device (e.g., mobile phone, tablet, computer) by an individual or with assistance from study staff (e.g., at an IHCS visit or through the help desk). All questionnaires will be submitted to the IRB for approval prior to implementation. Each questionnaire module is expected to take, on average, 20-30 minutes but not to exceed an hour to complete. Participants will receive reminders to complete their questionnaires (see "Reminders" in Section 5.6 Strategies for Recruitment and Retention). Participants will receive relevant information obtained from their responses to questionnaires (e.g., dietary/nutrient intake, physical activity) compared to the cohort or national distribution and national expert consensus recommendations.
    - **Baseline Questionnaires**: Participants will be asked to complete an electronic baseline questionnaire administered in four modules (see Section 12 Attachments) to collect information on factors, such as demographics, behavior, anthropometry, environment and medical history, via the secure participant application. Participants will be strongly encouraged to complete their baseline questionnaire within three months of enrollment but will be permitted to complete them at any time. For each biospecimen collection, participants will be asked to complete a questionnaire about recent exposures (see Section 7.3 Biospecimen Evaluations).
    - **Follow-up Questionnaires**. In addition to baseline questionnaires, participants will be asked to complete follow-up questionnaires, on average, every 6 months for the duration of the study. Questionnaires will include questions related to health as well as updates to basic contact information. It may be necessary to send more frequent questionnaires to accurately capture certain exposures such as dietary intake and physical activity. At six months after enrollment, participants will be asked to log their dietary intake using the NCI's Diet History Questionnaire III (DHQIII). Subsequently, participants will be asked to report 24-hour dietary intake and physical activity using validated instruments (e.g., NCI's Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24) and Activities Collected over Time over 24 hours (ACT24)).
- **Linkage to EHR data or insurance claims data**. EHR data, broadly defined to include medical history, medical imaging and any other health related data such as pathology reports, and insurance claims data will be requested from the participating IHCS for each of the consented participants with HIPAA authorization. The EHR requests will be made at baseline to obtain participant data prior to study enrollment and at each year after enrollment until the participant leaves the IHCS. The EHR data will be transferred from the IHCS to

NCI using a secure file transfer protocol server with data encrypted in transit and during rest (i.e., Box). In addition to the participants' IHCS EHRs, the study will explore the possibility of acquiring health records from other sources using existing and newly emerging tools. These alternate methods could include data linkages (e.g. imaging facilities, pharmacies) or obtaining records directly from participants, particularly for those who leave the IHCS or who receive health care outside of their IHCS. EHR from other healthcare providers can also be requested.

- **Assessment of adverse events.** Unexpected adverse events will be submitted to the IRB as soon as they occur.

7.3 BIOSPECIMEN EVALUATIONS

Multiple biological specimens, including dedicated research samples and discard clinical specimens, will be collected at various time points during the study (see Table 2). The baseline biospecimen collection will include dedicated research samples of blood, saliva and urine. In addition, tissue samples from cancer precursors, other benign conditions and non-cancerous tissue collected as part of the participant's medical care will be requested from the IHCS at baseline. During follow-up, we plan to repeat collection of the baseline specimens and obtain other biospecimen types and tissue from incident cancers and metastases or precursors. Detailed SOPs and MOOPs will be developed with details on these procedures. To ensure high uniformity of processing across the sites, detailed quality assurance and quality control programs will be established.

[redacted]Types of biospecimens collected from study participants

Biospecimen type	Research or Clinical Discard Specimens	Collection tube/kit and volume	Collection timing
Blood	Research or Clinical Discard	Serum and plasma tubes, <48 ml	Baseline and every 2-3 years
Urine	Research	Spot urine, 10 ml from a larger collection cup	Baseline and follow up
Saliva	Research	Scope mouthwash or other collection kit, 10 ml	Baseline and follow up
Feces	Research or Clinical Discard	FOBT or FIT, or self-collection kit	Follow up
Hair	Research	Self-collection kit	Follow up
Toenails	Research	Self-collection kit	Follow up

Cervical specimens	Clinical Discard	Liquid-based cytology containers	Follow up
Tissue	Clinical Discard	Formalin-fixed paraffin embedded (FFPE) or fresh frozen blocks, tissue/cytology slides	Follow up

The collection methods for dedicated research samples will depend on various factors such as the specimen type and infrastructure at the IHCS sites. Participants will be able to either donate biospecimens by (a) scheduling a biospecimen donation appointment at a site-specific clinic, (b) walking in as they would for any clinic blood draw via a standing research blood/urine draw order or (c) submitting a self-collected sample if appropriate for the specimen type. Urine and blood samples will be shipped to a site-specific regional laboratory or to a central facility for processing (e.g., blood fractionated and aliquoted, urine aliquoted) according to standard operating procedures. Efforts will be made to minimize the "needle to freezer" time to the extent possible.

- **Blood**. Blood will be collected at baseline and during follow up (we anticipate on average every 2-3 years). Whenever possible, the research collection will be drawn with a blood draw for clinical purposes. At each research collection, participants will provide up to 48 ml of blood for research purposes, including serum, ethylenediaminetetraacetic acid (EDTA) plasma, heparin plasma, and whole blood. Numbers of tubes will depend on the sizes of blood tubes available for use at each IHCS. A re-draw at a future date may be requested in the event of an issue with the sample (e.g. hemolyzed/ clotted samples).
- **Urine**. The baseline urine collection will be collected with the blood collection or at home depending on the IHCS. The baseline collection will be a 10ml spot urine sample collected in a standard size collection cup (~90-120 ml). Future collections might involve other protocols, such as collection of first morning void.
- **Saliva**: Saliva will be collected at baseline using a mouthwash protocol. The collection will occur at the same time as a blood/urine collection or at home using a kit that will be mailed to the participant's home for self-collection and return. The anticipated procedure will be a 30-second oral rinse and gargle with 10 ml of Scope mouthwash (or saline mouthwash for those avoiding alcohol). Self-collection kits would be returned via the USPS at no cost to the participant. Future repeat collection of saliva using this or modified protocols might occur.
- **Tissue**. IHCS sites will use EHR or other medical data and tumor registries to identify relevant tissue samples collected any time before or after baseline enrollment. Tissue samples of interest can include primary cancers, metastases, cancer precursors, other benign conditions and non-cancerous tissue. Cancer precursors, benign conditions and non-cancerous tissue can be identified adjacent to a cancer or collected independently as part of diagnostic workup without a cancer diagnosis. All clinical tissue specimens (e.g., formalin-fixed paraffin-embedded tissue blocks or slides) selected for study will be sent to a central

pathology laboratory for review. If possible, research sampling and processing will occur centrally and returned to the IHCS sites as soon as possible. If the blocks have clinical value, they will not be depleted. If required for patient care, tissue will be returned immediately upon request.

- **Other collections**: Other sample types and environmental samples will be considered for collection as the scientific need arises. The details of these protocols will be provided with a future study modification. Examples of such samples may include:
  - Hair and/or toenail samples (self-collection kits)
  - Fecal samples (self-collection kits)
  - Discard clinical samples (e.g., cervical swabs, fecal tests)
  - Environmental samples (e.g. water, house dust, clothes dryer lint)
- **Biological specimen questionnaires.** At the time of or shortly after providing a biological specimen, participants will be asked to complete a short electronic questionnaire (see Section 12 Attachments) to assess pre-analytical factors that might affect biomarker measurements. Questionnaire items may include fasting status and medication use at the time of collection, date of last menstrual period (for premenopausal women), current health status and other questions. Future biospecimen collection questionnaires will be submitted to the IRB for approval prior to use.
- **Sample storage, tracking and disposition**. Biological specimens will be stored temporarily at appropriate temperatures at the collection facilities (e.g., IHCS) or participant's home (e.g. refrigerated, ambient) until transferred to the Connect processing facility (Frederick National Laboratory operated currently by Leidos Biomed). Shipment procedures will be outlined in the SOPs to maintain the cold chain needed to prevent specimen degradation. Specimens collected by participants at home will be shipped directly to the Connect processing facility. Tracking of the biological specimens will be done by the IHCS and the Connect processing facility through the study data system.

The samples obtained under this protocol will be stored indefinitely at appropriate temperatures at the Connect central repository (Frederick National Laboratory operated currently by Leidos Biomed). De-identified information of these samples (e.g., sample ID, date of collection) will be tracked in DCEG's instance of the Biospecimen Inventory and Resource Management system. Biospecimens will be destroyed for participants at the time of revocation or if they are no longer of scientific value. The Connect central repository will dispose of biological specimens in accordance with all regulations. Because biospecimens could potentially contain biohazardous materials, the repository will not be able to return unused specimens back to participants or their families. Culturally sensitive practices may be developed as a response to requests from participants or their relatives, for instance cultural ceremonies performed by and accredited tribal healer or medicine man to administer blessing or last rites to samples prior to their destruction. The DCEG Coordinating Center will record any loss or unanticipated destruction of samples as a deviation. Reporting will be per the requirements of Section 6 Participant Discontinuation and Withdrawal.

7.3.1 Correlative Studies for Research

The biospecimens obtained under this protocol will be used in established, experimental and future assays. For example, they will be used to measure blood levels of hormones, inflammation markers, metabolomics; urine metabolomics; saliva and fecal microbiome; heavy metals in hair and toenails; cervical swabs for HPV testing; and somatic mutations in tissue samples. In subsequent biospecimen collections, samples could be collected for other assays or products (e.g., cell lines). The study will remain open so long as biospecimens are viable and data analyses continue.

7.3.2 Samples for Genetic/Genomic Analysis

7.3.2.1 Description of the Scope of Genetic/Genomic Analysis

Germline and somatic DNA and/or Ribonucleic Acid (RNA) will be extracted from different sample types, including blood, saliva or tissue samples (non-cancerous, precursor, and tumor). The extracted DNA/RNA will be used for genetic/genomic analyses to address a wide range of research and scientific questions including, but not limited to, genetic susceptibility to cancer and other traits, germline-somatic relationships, cancer transformation and progression through characterization of early somatic changes in non-cancerous or precursor tissues and the role of infectious agents and the microbiome in relation to cancer risk. Examples of genetic/genomic analysis that could be performed include DNA genome-wide genotyping arrays; targeted, whole exome or whole genome DNA sequencing; messenger RNA (mRNA)/microRNA (miRNA) sequencing; genome-wide chromatin accessibility analyses; and epigenetic changes such as DNA methylation arrays or pyrosequencing. As knowledge and technology evolves, it is possible that the biospecimens may be used in genetic tests that have yet to be developed.

7.3.2.2 Description of How Privacy and Confidentiality of Medical Information/Biological Specimens Will be Maximized

Steps to protect the privacy and confidentiality of all study participant data is described in detail under Publication and Data Sharing Policy (see Section 9.11). Briefly, all data and samples from the study will be coded and the key to personal identifiers will be securely stored with highly restricted access at NCI. Although participant identifiers will not be attached to the data or samples, genomic information or the combination of extensive clinical/demographic information collected from study participants can be unique to an individual. However, in order for anyone to use these sources of data to identify specific individuals, they would need to have access to a database with the same combination of information. Genetic and genomic data from this study will be available in accord with the NIH Genomic Data Sharing (GDS) Policy (see Section 7.3.2 Samples for Genetic/Genomic Analysis). Researchers involved in this study and collaborators who request access to study data will be prohibited from attempting to use study information to identify individuals as defined in their Data Use Agreements. This study is covered under the NCI Certificate of Confidentiality.

7.3.2.3 Management of Results

Any genetic information generated during the study will be used for research purposes and may be returned to the participant based on future amendments to the protocol. Specific protocols for the return of results (e.g., ancestry, clinically actionable genetic mutations, non-clinically important phenotypes such as cilantro taste, clinical blood counts) will be reviewed and approved by the IRB before efforts proceed. The plan will outline how to educate participants about the risk and benefits of medically actionable results, the procedure to confirm results in a Clinical Laboratory Improvement Amendments (CLIA) environment and appropriate quality assurance procedures, and other procedures.

8 STATISTICAL CONSIDERATIONS

Connect for Cancer Prevention is a research resource that will be used to address a wide range of scientific questions related to cancer etiology, precursor to tumor transformation, risk assessment and early detection, as described in the Introduction (see Section 2). Here we describe the primary statistical endpoints, an overview of analysis strategies and an evaluation of the design characteristics for this prospective cohort study. An important aspect of this cohort study is the prospective (i.e., prior to cancer diagnosis) and serial collection of comprehensive information on risk factors for cancer initiation and progression in a population setting. Therefore, a primary aim of statistical analyses will be to integrate complex high-dimensional and serial data to address scientific questions of interest. Statistical analysis plans and power calculations to address specific research aims will be written at the time of application for data access through the Connect for Cancer Prevention Data/Specimen Access Committee (see Appendix 1 Study Governance Structure).

A formal Statistical Analysis Plan (SAP) is not appropriate for this study. The comprehensive nature of the Connect for Cancer Prevention cohort will be an invaluable resource to test hundreds to thousands of hypotheses for cancer prevention and other general research uses.

8.1 STATISTICAL HYPOTHESIS

- Primary Endpoint(s): Cancer transformation, incidence, progression, and mortality
- Secondary Endpoint(s): General research use

8.2 SAMPLE SIZE DETERMINATION

Many of the study objectives require complex longitudinal statistical analyses involving integration between multiple data sources that do not lend themselves to simple sample size and power determination. For power calculations, we focus on estimating power for (i) testing an association between a baseline binary variable (such as a behavioral, environmental, or genetic factor) and cancer incidence, (ii) prospective follow-up on all participants with baseline and longitudinal biomarkers, behavioral and environmental factors, and (iii) nested case-control design in which biomarkers are assessed longitudinally on all cases and on a control series that is matched with respect to age and length of follow-up (i.e., incidence density sampling).

A sample size of 200,000 participants was chosen to provide sufficient numbers of expected cancer endpoints during the follow up period to evaluate key research questions with adequate statistical power. As for any epidemiological studies, evaluation of some hypotheses will likely require combining data with other cohorts to further increase the sample size. The data/sample collection in Connect for Cancer Prevention is designed to facilitate these pooling efforts with newer cohorts.

Table 3 shows estimates of the number of incident cases expected to occur in the first ten or fifteen years after enrollment based on cancer incidence rates from the SEER 18 registries for 2010-2012 (all race/ethnic groups combined). Calculations assume a uniform age distribution at study entry between 40 and 65 years, an equal number of men and women in all age groups, and enrollment of an equal number of participants/years over five years. In addition, calculations assume an annual loss to follow-up of 4% for participants aged 40-44 years old at study entry, 3% for 45-49 years, 3% for 50-54 years, 2% for 55-59 years and 2% for 60-64 years. The incident number of cancers was based on the first primary cancer in a specific organ site.

[redacted]Estimated Total Number of Incident Invasive Cancers for a Cohort Size of 200,000 Participants Aged 40-65 Years Old at Enrollment, After 10 or 15 Years of Follow-up

Cancer site	Total Number of Expected Incident Cancers	
	10 years of follow up	15 years of follow up
Prostate (males only)	1,951	3,595
Breast (females only)	1,566	2,553
Lung & Bronchus	1,151	2,165
Colorectum	877	1,566
Melanoma of the skin	410	681
Non-Hodgkin Lymphoma	369	649
Uterine corpus	345	579
Urinary bladder	327	622
Thyroid	232	356
Pancreas	221	410
Leukemia	203	365
Ovary	146	246

Table 3 illustrates that the expected number of events varies widely across cancer types, indicating that we will have substantially more power to address questions related to common cancers such as prostate, breast, and lung relative to rarer cancers (e.g., pancreas, leukemia, and ovary). These data will be used for all power calculations in this section.

Power to study baseline factors in relation to cancer risk

Table 4 shows the results of power calculations for testing an association between a binary baseline factor (e.g., dichotomous genetic, environmental, or behavioral factor) and cancer incidence prostate, colorectal, and pancreatic cancer as three examples, using the expected number of events from Table 3 These calculations will serve as a conservative estimate of power when studying the association of a continuous or ordinal exposure variable and cancer incidence. These three cancers were chosen to demonstrate cancer types with different event rates (i.e., incidence) so that the results are generalizable to a large spectrum of cancer types. The results show high power for detecting a relative risk of 1.5 for all but the rare cancer types when the baseline variable has a higher frequency 0.10. When the dichotomous variable has a frequency higher than 0.30, rare cancers such as pancreatic cancer still show moderate to high power to detect this association.

Estimated Power for the Association (Relative Risk Of 1.5) Between a Dichotomous Baseline Variable and Cancer Incidence. For a Cohort Size Of 200,000 Participants Aged 40-65 Years Old at Enrollment, After 10 And 15 Years of Follow-Up*. Power is Computed for a Probability of the Dichotomous Baseline Variable Being Positive Of 0.5, 0.3., 0.1, And 0.02.

Cancer Site	10 Years of Follow Up				15 Year Follow Up			
	Probability of dichotomous variable				Probability of dichotomous variable			
	0.50	0.30	0.10	0.02	0.50	0.30	0.10	0.02
Prostate	1.00	1.00	1.00	0.56	1.00	1.00	1.00	0.86
Colorectal	1.00	1.00	0.91	0.25	1.00	1.00	1.00	0.46
Pancreatic	0.84	0.73	0.31	0.06	0.98	0.95	0.56	0.11

*Calculations use proportions estimated from Table 5 and are based on a two-group test of proportions conducted at the 0.05 significance level with a two-sided test.

Power to study longitudinal biomarker changes in relation to cancer risk

The serial collection of biological specimens will allow study of the relationship between changes in biomarkers over time and cancer incidence. We propose to collect serial biospecimens for biomarker analyses, on average every 2-3 years over an extended follow-up. As an example, Table 5 shows power for examining the association between the slope of a longitudinal biomarker and prostate cancer incidence. Power is shown for different ratios of the residual error (measurement-specific biological or technical error) relative to individual level and slope (units per year) standard deviation (SD). The prospective design corresponds to analyzing longitudinal data on all 100,000 men enrolled in the study, while the nested design corresponds to a nested case-control study where cases are matched to controls (incidence density sampling

with 1:2 matching). The effect size considers is a relative risk of 1.5 corresponding to a 1 SD change in individual slope.

Power to Detect an Association Between the Slope of a Biomarker and Prostate Cancer Incidence. Calculations are Based on Detecting a Relative Risk of 1.5 For a 1 Standard Deviation Change in the Individual Slope.*

Ratio of true baseline SD to slope SD	Ratio of residual SD to slope SD	Prospective Design		Nested Case-Control Design	
		10-year follow up	15-year follow up	10-year follow up	15-year follow up
2.5	**7.5**	1.00	1.00	0.94	0.97
2.5	**22.5**	0.61	0.99	0.36	0.72
7.5	**7.5**	1.00	1.00	0.94	0.97
7.5	**22.5**	0.60	0.99	0.36	0.74

*Based on the joint model, where the random slope in a linear mixed model is included as a covariate in a Cox proportional hazards model. We assume a standard deviation for individual slope of 0.4. Power was evaluated for a 0.05 significance level, with a simulation study with 10,000 realizations. For these calculations, we assumed that participants provide a blood samples on average every 2-3 years across all ages.

For a prospective design in which biomarkers are analyzed on all serial measurements, the study will have high power to detect a relative risk of 1.5 for a 1 SD change in individual slope over a 15-year follow-up for all four scenarios. For the shorter 10-year follow-up, the study will have high power unless the residual variation is much larger than the individual slope variation (e.g., a ratio of 22.5 results in a power of 0.60). For a nested case-control study (1:2 matching), the study will have high power to detect a relative risk of 1.5 for both the 10- and 15-year follow-up period unless the residual error is very large relative to the individual slope variation.
The most important reason for collecting serial biomarker measurements is to characterize the effects of the dynamics of the longitudinal process on cancer incidence. However, due to measurement error in biomarker measurements, even if this process is not changing across time, there may be sizable power gains by using a longitudinal design as compared with simply using the observed baseline biomarker measurement. Table 6 shows the power to detect a relative risk of 1.2 for a 1 SD change in the unknown true baseline value.

Power to Detect an Association Between a Baseline Biomarker and Prostate Cancer Incidence With a Prospective Design. Calculations are Based on Detecting a Relative Risk Of 1.2 for a 1 Standard Deviation Change in Individual Slope*

	Modeled Baseline	Observed Baseline

Ratio of residual SD to true baseline SD	Ratio of base SD to slope SD	10-year follow up	15-year follow up	10-year follow up	15-year follow up
1.0	**7.5**	1.00	1.00	1.00	1.00
3.0	**7.5**	0.56	0.85	0.69	1.00
1.0	**6.0**	1.00	1.00	1.00	1.00
3.0	**6.0**	0.53	0.84	0.65	0.98

*Modeled baseline based on the joint model, where the random intercept (true baseline measurement) in a linear mixed model is included as a covariate in a Cox proportional hazards model. We assume a standard deviation for individual slope of 0.4. Observed baseline was based on a Cox proportional hazards model using the observed baseline measurement as a covariate. Power was evaluated for a 0.05 significance level, with a simulation study with 10,000 realizations.

For example, for a large residual error (e.g., the ratio of residual SD to true baseline SD of 3), the power increases from 0.85 to 0.98 when using the modeled baseline as compared with the observed baseline biomarker measurement. These calculations demonstrate the power advantages of using the longitudinal model to properly account for the measurement error in the baseline values (i.e., power advantage of modeled baseline over observed baseline).

Expected sample size for biomarkers of early detection of cancer

Identifying markers for early detection is a primary objective of this cohort. Based on collecting serial measurements every 2-3 years on average (for instance, four years from ages 40 to 60 and every two years from 60 to 80), we performed statistical calculations to examine the expected proportion of diagnosed cancers that will have a blood specimen within 1 and 2 years of a cancer diagnosis for early detection studies. For a 10-year follow-up, 42% and 78% of breast cancer cases will have provided a blood sample within 1 and 2 years of diagnosis. These proportions are 46% and 88% for prostate cancer and 46% and 86% for lung/bronchial cancer. These design calculations suggest that we will have a reasonable number of cases with blood draws close to diagnosis to study early detection biomarkers.

Expected sample size to study risk factors for cancer precursors

In the study, cancer precursors will occur more commonly and rapidly than cancer endpoints and thus will be investigated in earlier stages of the study. Cancer precursors could include benign breast disease, colon polyps, cervical cancer precursors, lung nodules, esophageal metaplasia/dysplasia, and endometrial hyperplasia. We evaluated the power for a cancer precursor risk of 5 and 15%. Table 7 shows the power to detect a relative risk of 1.5 for the association between a dichotomous variable (e.g., exposure, behavioral factor, or genetic trait) and the risk for a cancer precursor.

[redacted]Power to Detect an Association Between a Baseline Variable and Cancer Precursor Risk. Calculations are Based on Detecting a 1.5 Relative Risk with a Chi-Squared

Test at the 0.05 Significance Level. Power is Shown For Different Sample Sizes and Probability for a Positive Binary Factor.

Sample size (hypothetical numbers of endpoints)	Cancer precursor prevalence: 15%			Cancer precursor prevalence: 5%		
	Frequency of a positive baseline factor			Frequency of a positive baseline factor		
	0.50	0.30	0.10	0.50	0.30	0.10
1,000	0.75	0.62	0.25	0.31	0.22	0.08
5,000	1.00	1.00	0.90	0.90	0.80	0.37
10,000	1.00	1.00	1.00	1.00	0.98	0.70

The results illustrate that we will have high power to detect a relative risk of 1.5 with a sample size of 10,000 even for relatively rare dichotomous baseline factors.

8.3 STATISTICAL ANALYSES

8.3.1 General Approach

The longitudinal assessment of questionnaires and specimen collections along with repeated assessment of EHR provide a unique resource for longitudinal data integration with the goal of improving our understanding of the progression to cancer and its precursors in healthy populations.

This section describes standard analyses that can be used to analyze the primary endpoints of the cohort. It is important to note that as Connect for Cancer Prevention will be a research resource for the scientific community, data will be used to conduct a wide range of analyses by many investigators. Therefore, here we are just describing what we consider as standard methods to address the primary objectives, rather than an extensive description of all possible analyses. Depending on the scientific question, adjustments for multiple comparisons should be considered. Although this cohort will be a unique data source, novel risk factors and biological mechanisms should be confirmed in other cohorts whenever possible.

8.3.2 Analysis of the Primary Endpoints

Statistical models for prospective data such as Cox proportional hazards models will be used to estimate the hazard ratios for the relationship between of baseline variables and changes in exposure/biomarkers over time on cancer incidence, as well as on the incidence of important cancer precursors. Standard longitudinal data analytic techniques will be used to examine mean changes in time-dependent variables across time. For instance, linear mixed models will be used to estimate changes in continuous biomarkers across time and to characterize the heterogeneity in these changes across participants and generalized estimating equations and generalized linear mixed models will be used to estimate and test for changes in discrete variables across time.

In addition to applying the standard statistical methodologies described above, integrating complex high-dimensional data and studying the longitudinal dynamics of exposures and biomarkers on the incidence of cancer and its precursors will involve the development of novel analytic strategies. Examples include:

- Joint modeling strategies between longitudinal and time-to-event data can be used to study the relationship between changes in biomarker and life-style variables and cancer incidence (6). Specifically, these models can be used to examine the association between longitudinal variables and cancer endpoints (i.e., time-to-event) and to develop dynamic risk predictors that provide updated assessments of cancer risk with increasing amounts of longitudinal data. Longitudinal patterns will be characterized using mixed models with polynomial or cubic splines to represent biomarker dynamics over time. Joint models are an appealing approach since they can measure the association between the trajectory of a time-dependent variable and the risk of an event at any given time.
- Methodological approaches for joint models that allow for multiple variables of different types (gaussian, binary, and ordinal variables) will be developed.
- Transition model approaches will be used to study the effect of biomarkers on the cancer disease process (e.g. dynamics from a precursor to cancer).
- Machine learning techniques will be applied and developed for using multi-dimensional irregularly observed longitudinal measurements for risk prediction and early detection.
- Latent variable modeling approaches will be developed for integration longitudinal data from various sources (e.g., biomarkers, questionnaires, and EHR).
- Intensively collected longitudinal data collected from wearable devices will be analyzed with a two-stage approach where time-series models will be fit on a subject-specific basis, followed by relating summarizes of these models to disease endpoints.
- The development and validation of risk prediction models to estimate the absolute risk of developing cancer over a specified time period will integrate information on known and novel exposures and biomarkers using methodologies, such as the Individualized Coherent Absolute Risk Estimators (iCARE).
- The development of novel biomarkers for early detection will be examined through algorithms that incorporate change-point methodology in the longitudinal process, such as methods similar to the Risk of Ovarian Cancer Algorithm (ROCA).

8.3.3 Analysis of the Secondary Endpoint(s)

A variety of statistical approaches such as those outlined above will be employed to utilize the data resources for general research purposes.

8.3.4 Baseline Descriptive Statistics

Univariate and bivariate distributions will provide important insight into recruitment activities and later into the behavior of the data.

8.3.5 Subgroup Analyses

The distributions of baseline data will be summarized estimating the distribution of age, race/ethnicity, and gender by site. We will also evaluate associations between baseline variables using correlations and relative risk estimates. Longitudinal data will be explored using spaghetti plots where subject-specific trajectories are overlaid on a single plot.

8.3.6 Tabulation of Individual Participant Data

Tabulation of individual participant data will not be used for hypothesis testing.

8.3.7 Exploratory Analyses

Not applicable to this repository of data and biospecimens

9 REGULATORY AND OPERATIONAL CONSIDERATIONS

9.1 INFORMED CONSENT PROCESS

9.1.1 Consent/Assent Procedures and Documentation

- **Summary of process**. Individuals wishing to participate in Connect for Cancer Prevention will be asked to provide a single universal consent electronically for all study activities. Consent procedures will occur in the secure participant application. Individuals can complete the consent process in a location and time frame of their choosing. They will be able to exit the consent and return at a later time to consent or decline participation. Individuals can consult with family, friends, medical providers or others as needed prior to providing consent. Information about the study will be presented in an electronic educational module first (to be provided in a subsequent amendment) as well as the consent document (see Section 12 Attachments).
- **Request waiver of consent documentation**. The main contact with study participants, including invitation to enroll, is indirect, mainly through the electronic participant application. The reliance on the participant application is to facilitate long-term communication with this large number of study participants. Obtaining an ink signature prior to the initiation of study activities (such as the electronic questionnaires) would be impractical.
    - Research described in this study is no greater than minimal risk and involves procedures for which written consent is normally not required outside the research context. Research from this study is also expected to only offer indirect benefit to participants and participation is voluntary. The waiver will not affect the rights and welfare of the subjects.
- **Addressing participant questions**. Beyond the educational module (to be detailed in a subsequent IRB amendment) and consent document, additional information about the study will be found on the study's websites. If the potential participant has trouble understanding the consent information, trained study support staff at the IHCS or central study support staff via the Connect for Cancer Prevention Help Desk may answer questions to achieve informed consent, per 45 Code of Federal Regulations (CFR)

§46.116 (a). Contact information for study support will be visible throughout the process in case the participant has any questions. This approach helps to ensure consistency between enrolling centers as well as ensuring that an individual's decisional autonomy is respected.

- **Consent procedure**. The materials presented will be consistent across the IHCS sites but may be customized based on an individual's geographic location, enrollment method, or IHCS affiliation.
  1. Prior to consent, interested individuals will create a secure account (i.e., log in) on the study website using an email address or cellular phone number. After authenticating the email address or cellular phone number, the individuals will be directed to the consent process.
  2. To facilitate participant understanding, we plan to provide a web-based educational module that highlights the key aspects of the study activities and procedures, consent, and HIPAA authorization (details to be provided in a subsequent amendment).
  3. Individuals will be presented with the full consent document (see Section 12 Attachments). The individual will have to scroll through the entire document before advancing to the next step.
  4. If an individual decides to consent to enroll into the study, the individual will electronically consent by checking a box indicating the individual's agreement to the following statement:

     "By clicking "Agree" and typing your first and last name, you agree to the following:

     1.I have read this form.

     2.If I have questions, I can contact the Connect Support Center at [website] or [phone].

     3.I agree to voluntarily provide my survey answers, medical information, and samples (blood, urine, saliva, tissue, etc.) as part of the Connect for Cancer Prevention study. I understand that my information and samples will be securely stored and shared with researchers as described in this consent form."

     The individual will also enter their first and last name. A record of the consent including the participant's name, date and time of consent, and consent version will be stored securely by the DCEG Coordinating Center and maintain an audit trail. The participant will be able to view, download and print their signed consent form. It will also be made available to the recruiting IHCS study staff and a copy may be stored in their medical record at the institution's discretion.

  5. The HIPAA authorization document will then be presented. The "signature" process will be the same as described above for consent.

- **Language**. Informed consent and other study materials will be available only in English at the time of the initial study launch. Other languages (e.g., Spanish) spoken among the IHCS patient population will be considered in subsequent years of recruitment. Any translated materials will be generated through an IRB- approved translation procedure and provided to the IRB for their records. Verbal translation into languages without official translation will not be allowed.
- **Assent**. Assent process will not be established since we will not be enrolling minors under this protocol.

9.1.2 Consent for minors when they reach the age of majority

This section is not applicable to our study.

9.1.3 Telephone consent

This section is not applicable to our study.

9.1.4 Telephone assent

This section is not applicable to our study.

9.1.5 Considerations for Consent of NIH employees

This section is not applicable to our study.

9.1.6 Consent of Subjects who are/become Decisionally Impaired

Decisionally impaired individuals are not expected to consent to this study. There are no plans to proactively evaluate participants for impaired status given the minimal risk nature of this study.

9.2 STUDY DISCONTINUATION AND CLOSURE

This study may be temporarily suspended or prematurely terminated if there is sufficient reasonable cause. Written notification, documenting the reason for study suspension or termination, will be provided by the suspending or terminating party to study participants, investigator, funding agency and regulatory authorities. If the study is prematurely terminated or suspended, the Principal Investigator (PI) will promptly inform study participants, the Institutional Review Board (IRB), and sponsor and will provide the reason(s) for the termination or suspension. Study participants will be contacted, as applicable, and be informed of changes to study visit schedule.

Circumstances that may warrant termination or suspension include, but are not limited to:

- Loss or insufficient funding
- Determination of unexpected, significant, or unacceptable risk to participants
- Insufficient compliance to protocol requirements
- Data that are not sufficiently complete and/or evaluable

Study may resume once concerns about safety, protocol compliance, and data quality are addressed, and satisfy the IRB.

9.3 CONFIDENTIALITY AND PRIVACY

Participant confidentiality and privacy is strictly held in trust by the participating investigators and their staff. This confidentiality is extended to cover testing of biological samples and genetic tests in addition to the self-reported questionnaire, electronic health record, and other medical information relating to participants. Therefore, the study protocol, documentation, data, and all other information generated will be maintained in confidentiality to the full extent permitted by law (https://grants.nih.gov/policy/humansubjects/coc/what-is.htm).

All data (including PII, PHI and research data from the participant application and other sources) obtained during the conduct of the study will be stored in the Connect Data Platform, a FedRamp cloud environment hosted by NCI/DCEG. Research data and samples collected from participants will be coded (i.e., all personal identifiers will be removed from data/samples), and the "key" for the code will be securely stored at NCI with highly restricted access by trained members of the Connect Coordinating Center. Security measures to restrict access to PII, PHI and the code key will include storing this information in a password-protected database separately from other research data that will be more broadly accessible. Any written data will be kept in a locked filing cabinet. Data transmissions, such as the transfer of the electronic health record and other medical information, to the Platform will be conducted through a secure file transfer protocol with the data encrypted in transit and at rest. The Platform will ensure implementation of privacy and security policies by all components and enforce those policies upon users.

PII will be accessed by the Connect Coordinating Center and IHCS study staff for administrative purposes, maintaining communication with the study participants and coordinating future study activities. Administrative activities include data linkages of personal identifying information with state, national or other registries of demographic or health outcomes, environmental exposures, Census data, or other geocoded data. In limited cases, such as geocoded data, investigators from other institutions, including universities, private companies, or non-profit organizations, might also request location information. Representatives of the IRB and/or regulatory agencies may also inspect all files required to be maintained by the investigator, including but not limited to, research records for the participants in this study.

Study participant research data for statistical analysis and scientific reporting will be accessible through the Connect Data Platform. Individual participants and their research data will be identified by a unique study identification number. Access to the Connect Data Platform will be secured through two-factor authentication.

Other procedures that will be adhered to ensure confidentiality of the data include:

- The consent form and HIPAA authorization (see Section 9.1.1 Consent and Assent Procedures and Documentation) detail information about confidentiality and the risk of potential data breaches.
- Most study activities will be conducted through the study participant application, which will allow interested individuals and participants to perform study activities in as private a setting as possible.

- Only specimens and data that have been deidentified and coded will be made available to researchers for analyses. The results of these tests will not be linked to personal identifier information.
- No individual results will be presented in publications or other reports.
- No information from the study will be placed in the participants' medical records by NCI or its collaborative research team.
- All directly identifiable information will be protected by systems meeting or exceeding the Federal Information Systems Management Act (FISMA) Moderate standards and authorized to operate by NIH and the NIH Office of the Chief Information Officer (OCIO). Devices such as tablets and laptops that may be used by the IHCS to register and collect participant information will be shut down automatically after a few minutes of non-use. Transmission of information between the IHCS and NCI also complies with high standards of security and is included under the review and approval purview of the NIH OCIO.
- To minimize the risk of inappropriate disclosure of PII and PHI, all personnel involved in this research (including IHCS and NCI employees and contracting institutions) will undergo mandatory annual training in human subject's research, including knowledge assessment regarding privacy protection and HIPAA rules.
- Should we become aware that a major breach in our plan to protect subject confidentiality and study data has occurred, this will be reported expeditiously per requirements in Section 9.9.2 Unanticipated Problem Reporting.
- To further protect the privacy of study participants, a Certificate of Confidentiality has been issued by the National Institutes of Health (NIH). This certificate protects identifiable research information from forced disclosure. It allows the investigator and others who have access to research records to refuse to disclose identifying information on research participation in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level. By protecting researchers and institutions from being compelled to disclose information that would identify research participants, Certificates of Confidentiality help achieve the research objectives and promote participation in studies by helping assure confidentiality and privacy to participants.

9.4 FUTURE USE OF STORED SPECIMENS AND DATA

De-identified study data, such as survey or EHR data, and biospecimens (collectively referred to as the "Connect Resource") will be available for the research community through a continuum of open to controlled access , managed by the Connect Resource Access Committee (see Appendix 1 Study Governance Structure) and available on the Connect Data Platform (see Section 9.11.1 Human Data Sharing Plan). Sharing research data supports the mission of the NIH and is essential to facilitate the translation of research results into knowledge, products, and procedures that improve human health. Therefore, the Connect scientific leadership will encourage extensive and appropriate use of the Connect Resource to address a wide range of scientific questions. Research findings that emerge from the Connect Resource will be broadly communicated to the American public through the study's public facing website, press releases of scientific

publications, and other communication opportunities. The scientific value of the Connect Resource is expected to extend several decades.

9.5 SAFETY OVERSIGHT

For this minimal risk, observational study, the PI with support from the Connect Senior Scientist and study management staff will provide safety oversight of the study activities.

9.5.1 Principal Investigator/Research Team

The study management team will meet on a regular basis (approximately weekly) when subjects are being actively enrolled/evaluated on the study.

All data will be collected in a timely manner and reviewed by the PI or the Connect Senior Scientist. Events meeting requirements for expedited reporting as described in HRPP Policy 801 will be submitted within the required timelines.

9.6 CLINICAL MONITORING

Not applicable to this study.

9.7 QUALITY ASSURANCE AND QUALITY CONTROL

The study will be conducted in accordance with procedures identified in the protocol and SOPs for data/sample collection and quality management. Study staff, including but not limited to those at the IHCS sites, the NCI central repository, and support services contractor, will be trained and follow the SOPs. Regular monitoring and audits for the various study components will be performed to assure protocol compliance, data quality and proper storage and handling of samples.

Electronic monitoring of key metrics for recruitment, biospecimen collection and shipment and help desk issues will be monitored weekly by the Connect Coordinating Center. On-site review by the Connect Coordinating Center will be conducted early, for initial assessment and training and throughout the study to ensure monitoring practices are performed consistently across all participating sites.

9.8 DATA HANDLING AND RECORD KEEPING

9.8.1 Data Collection and Management Responsibilities

Data and biospecimen collection and accurate documentation are the responsibility of the study research team under the supervision of the Connect for Cancer Prevention Executive Committee (see Appendix 1 Study Governance Structure). DCEG will serve as the Connect Coordinating Center for this multi-center study, ensuring data accuracy, consistency and timeliness. The Connect Coordinating Center will use strategies to ensure consistency in data quality and on-going adherence to standards for data collection and formatting:

- Define standards for shared data items, formats, and metadata elements will be used by all participating sites to collect and structure data files from recruitment, enrollment and all data collection activities, including data collected through the participant application (e.g., user

profile, questionnaires). These standards will be updated periodically to reflect new data collection activities as well as reviewed for improvements in instrumentation or changes in data collection protocols. Data completeness and data quality assessments including measures of unusable or unlinked data fields will be part of the data standards document. Included in the data quality assessments will be estimates of the number of participant records deemed unusable because of quality issues and documentation of data collection methods that can be improved to prevent further loss of participant records.

- A review of the adherence to the standards for data collection items and for data quality will be conducted. This review will consist of regular audits of submitted data for metadata completeness and consistency and a review of reports on data completeness and quality assessments. Of primary importance, the data quality assessments will aid the IHCSs to improve data collection protocols and documentation as the study progresses.

9.8.2 Study Records Retention

Study documents should be retained as per the HHS and NIH Intramural Records Retention Schedule, as applicable.

9.9 UNANTICIPATED PROBLEMS

9.9.1 Definition of Unanticipated Problems (UP)

Any incident, experience, or outcome that meets **all** the following criteria:

- Unexpected in terms of nature, severity, or frequency given (a) the research procedures that are described in the protocol-related documents, such as the Institutional Review Board (IRB)-approved research protocol and informed consent document; and (b) the characteristics of the participant population being studied; and
- Related or possibly related to participation in the research ("possibly related" means there is a reasonable possibility that the incident, experience, or outcome may have been caused by the procedures involved in the research); and
- Instances in which the research places participants or others (which many include research staff, family members or other individuals not directly participating in the research) at a greater risk of harm (including physical, psychological, economic, or social harm) than was previously known or expected.

9.9.2 Unanticipated Problem Reporting

The investigator will report unanticipated problems (UPs) to the NIH IRB as per Policy 801.

9.10 PROTOCOL DEVIATIONS AND NON-COMPLIANCE

We will follow the reporting requirements in Policy 801 for Reporting Research Events and Policy 802 for Non-Compliance Human Subjects Research. Because this is an observational cohort study with no intervention, experimental treatments or procedures performed as a part of this protocol, and with minimal risks for the participants, we request a waiver of reporting requirements to the NCI Clinical Director for research-related anticipated adverse events.

The Principal Investigator/Executive Committee will meet on a regular basis (e.g., monthly by phone) when participants are being actively enrolled. All data will be collected in a timely manner and reviewed by the principal investigator or a lead associate investigator. Any safety concerns, new information that might affect either the ethical and/ or scientific conduct of the study will be immediately reported to the IRB.

The DCEG Coordinating Center will ensure that all IHCS sites or participating laboratories physically working with Connect for Cancer Prevention participants have procedures in place for responding to incident such as physical injury that occurred while on site, for instance, known risk associated with blood draws (see Section 2.3.3 Assessment of Potential Risks and Benefits There is no foreseeable risk associated with saliva or urine collections

Unexpected adverse events and unanticipated problems which are not consistent with the known or foreseeable risk of adverse events associated with the research procedures will be documented by the IHCS sites. These reportable events will be filed the NIH and the IRB in accordance with the HHS requirement for disclosing reportable evens and unanticipated problems.

9.10.1 NIH Definition of Protocol Deviation

A protocol deviation is any changed, divergence, or departure from the IRB-approved research protocol.

- **Major deviations**. Deviations from the IRB approved protocol that have, or may have the potential to, negatively impact the rights, welfare or safety of the subject, or to substantially negatively impact the scientific integrity or validity of the study.
- **Minor deviations**. Deviations that do not have the potential to negatively impact the rights, safety or welfare of subjects or others, or the scientific integrity or validity of the study.

9.11 PUBLICATION AND DATA SHARING POLICY

9.11.1 Human Data Sharing Plan

The principles of the Connect Resource (see Section 9.4 Future use of Stored Specimens and Data) access and use policy will be consistent with the NIH policy for data management and sharing (draft) and the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles (https://commonfund.nih.gov/commons). Specifically, the procedures and processes that have been applied to access to the Connect Resource derive from the following key principles:

- The Connect Resource will be available to scientific researchers without preferential or exclusive access for any person, to the extent possible. All researchers will be subject to the same application process and approval criteria.
- Aggregated data that cannot be used to identify individuals will be available on the Connect Data Platform without review or registration of users. Limited aggregated data will be widely available in real-time (e.g., verified participant counts with demographic information). Other aggregated data will be updated on a regular basis (e.g., annually).

- Individual-level data, biospecimens, or other Connect Resources will be under controlled-access. Scientific researchers will request Resource access through a multistep process coordinated by the Connect Coordinating Center and directed by the RAC (see Appendix 1 Study Governance Structure). A scientific proposal, submitted by scientific researchers, will be reviewed to ensure that the proposed aims are consistent with the access procedures, the study protocol, and the consent that was provided by the participants, and that they have any relevant ethics approval, if required. The RAC may seek review of proposals by review committees of scientific experts. If there are any intellectual property issues (e.g., patent filings) or contractual obligations that would preclude sharing and secondary research with the data, data will not be made available.
- Access to the biological samples will be carefully controlled and coordinated since they are limited and depletable resources. Cancer-related hypotheses will be prioritized. The quantity of sample that is required will be judged against the potential benefits of the research project, with advice from appropriate experts when required. Requirements to obtain access will be determined at a later time but can include preliminary data on influence of pre-analytical factors, assay reliability among others. When appropriate, internal controls will be incorporated into the study sample sets and interim data reporting will be required to evaluate data quality. Priority will be given to hypotheses with cancer endpoints. Researchers will be expected to return to the Resource any derived data.
- Researchers with approved projects by the RAC will be able to access individual-level data and specimens after receiving IRB approval, if required, signing a Data Transfer Agreement (DTA) and/or Materials Transfer Agreement (MTA) as appropriate with NCI (see Section 9.12 Collaborative Agreements).
- Documentation of resource access concepts and the investigative team will be publicly available, when project is approved, to facilitate collaborative research and satisfy human subject and Privacy Act compliance.
- Safeguards will be maintained to ensure the anonymity and confidentiality of participants' data and samples. Researchers will agree to a data use policy that details their responsibility to avoid attempts to identify participants, and the individual-level data and/or samples provided by Connect to researchers will be de-identified.
- To improve data traceability and reproducibility of analyses/results according to FAIR data principles, the policy for the Project is to provide data access for analyses through a cloud-hosted Connect Data Platform (see Section 5.2.1 Human Data Sharing Plan). Downloading local versions of data will be strongly discouraged. Special requests for downloading the data will be considered by the RAC, if required analysis tools are not available on the Connect Data Platform and/or if data cannot be read remotely by the available analysis tools. If analyses are run in other environments, derived data and code must be submitted through the data platform to ensure FAIR principles, prior to any publication.
- The Connect Data Platform will also facilitate the return of and access to derived data (e.g., geo-linked data, biospecimen assay results), metadata, data dictionary, and

annotated code to all researchers, not only the ones performing specific analyses. To ensure compliance, limited covariate data will be provided until investigators return derived data and affiliated data and document. Data science infrastructure is emerging to support contained analytical environments with suitable user-facing analytical environment. The Connect team is an active participant in ongoing assessment of researcher-facing analytical solutions being considered by NCI/CBIIT.

9.11.2 Genomic Data Sharing Plan

This study will comply with the NIH Genomic Data Sharing (GDS) Policy, which applies to all NIH-funded research that generates large-scale human or non-human genomic data, as well as the use of these data for subsequent research. Large-scale data include genome-wide association studies (GWAS), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, epigenomic, and gene expression data. De-identified genomic data will be deposited on Connect Data Platform no later than at the time of publication. Access to the genomic resources will be through controlled access.

9.12 COLLABORATIVE AGREEMENTS

9.12.1 Agreement Type

- MTAs will be established, if required, between the NCI and IHCS (e.g., for tissue retrieval).
- DTA and MDTAs will be established between NCI and researchers requesting access to individual-level data or biospecimens (see Section 5.2.1 Human Data Sharing Plan)

Once the agreements are executed, we will provide a copy and the number identifying the agreements to the IRB.

9.13 CONFLICT OF INTEREST POLICY

The independence of this study from any actual or perceived influence is critical. Therefore, any actual conflict of interest of persons who have a role in the design, conduct, analysis, publication, or any aspect of this observational study will be disclosed and managed. Furthermore, persons who have a perceived conflict of interest will be required to have such conflicts managed in a way that is appropriate to their participation in the design and conduct of this observational study. The study leadership in conjunction with the National Cancer Institute has established policies and procedures for all study group members to disclose all conflicts of interest and will establish a mechanism for the management of all reported dualities of interest.

10 ABBREVIATIONS

ASA24	Automated Self-Administered 24-Hour Dietary Assessment Tool
CDCC	Connect Data Coordinating Center
CFR	Code of Federal Regulations
CLIA	Clinical Laboratory Improvement Amendments
CRF	Case Report Form
CT	Computed Tomography

DCEG	Division of Cancer Epidemiology and Genetics
DHQIII	Diet History Questionnaire III
DNA	Deoxyribonucleic Acid
DTA	Data Transfer Agreement
EDTA	Ethylenediaminetetraacetic Acid
EHR	Electronic Health Records
Epic	Epic Systems Corporation
FAIR	Findability, Accessibility, Interoperability, and Reusability
FedRAMP	Federal Risk and Authorization Management Program
FFPE	Formalin-fixed Paraffin Embedded
FIPS	Federal Information Processing Standard
FISMA	Federal Information Systems Management Act
GCP	Good Clinical Practice
GDS	Genomic Data Sharing
GWAS	Genome-Wide Association Studies
HIPAA	Health Insurance Portability and Accountability Act
HIV	Human Immunodeficiency Virus
HHS	United States Department of Health & Human Services
iCARE	Individualized Coherent Absolute Risk Estimators
ICH	International Conference on Harmonisation
ID	Identification
IHCS	Integrated Health Care Systems
IHC	Immunohistochemistry
IRB	Institutional Review Board
IT	Information Technology
IVR	Interactive Voice Response
MGUS	Monoclonal Gammopathy of Undetermined Significance
MOOP	Manual of Operating Procedures
MRI	Magnetic Resonance Imaging
MTA	Materials Transfer Agreement
mRNA	Messenger RNA
miRNA	microRNA
NAACCR	North American Association of Central Cancer Registries
NCI	National Cancer Institute
NDI	National Death Index
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NORC	National Opinion Research Center
OCIO	Office of the Chief Information Officer
OMB	Office of Management and Budget

PI	Principal Investigator
PII	Personal Identifying Information
PIN	Personal Identification Number
PHI	Personal Health Information
RACC	Resource Access Coordinating Committee
ROCA	Risk of Ovarian Cancer Algorithm
RNA	Ribonucleic Acid
SAP	Statistical Analysis Plan
SMS	Short Message Service
SNP	Single Nucleotide Polymorphisms
SOA	Schedule of Activities
SOP	Standard Operating Procedure
SD	Standard Deviation
UP	Unanticipated Problem
US	United States
USGCB	US Government Configuration Baseline
VPR-CLS	Virtual Pooled Registry Cancer Linkage System

11 REFERENCES

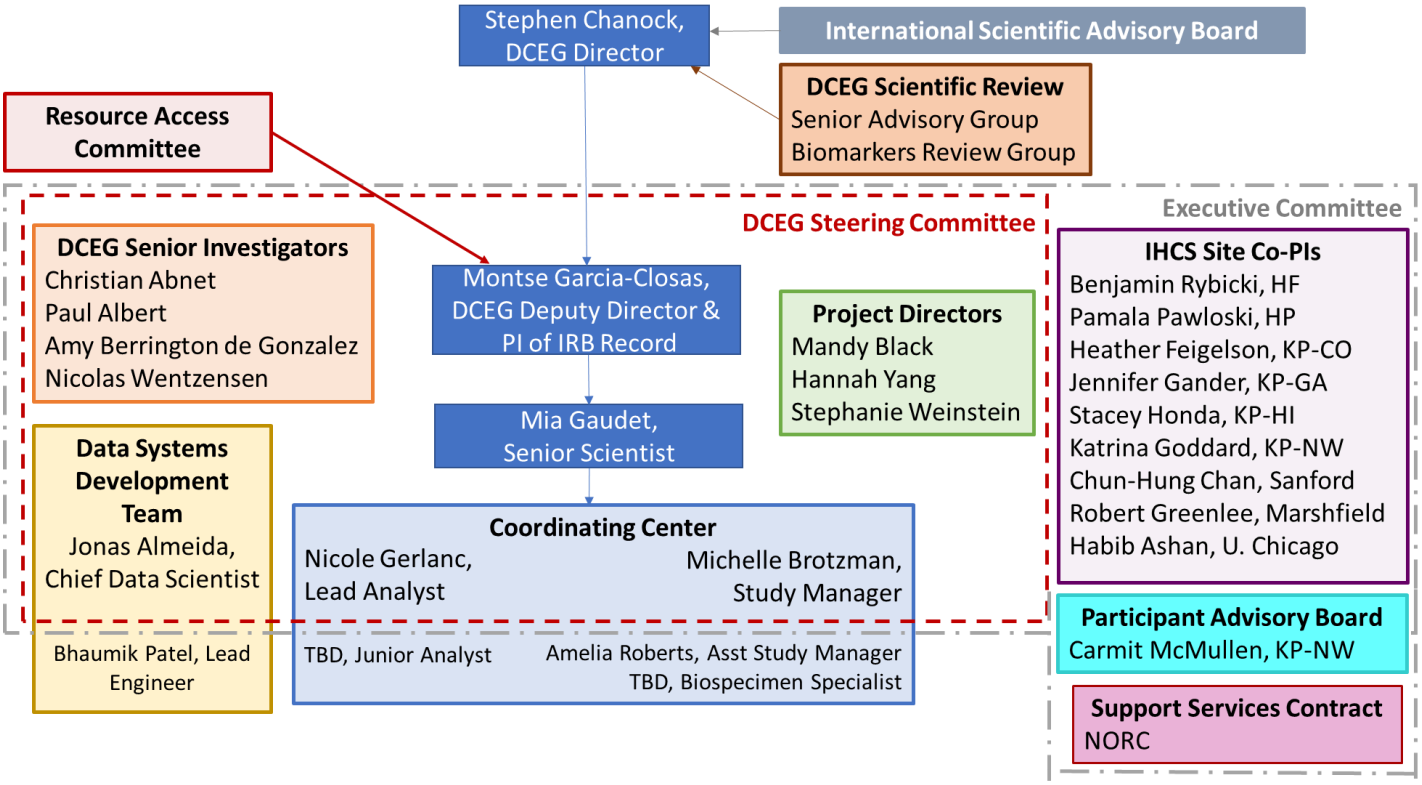
1. Rebbeck TR, Burns-White K, Chan AT, Emmons K, Freedman M, Hunter DJ, et al. Precision prevention and early detection of cancer: fundamental principles. Cancer Discov. [Internet]. 2018[cited 2020 June 26];8(7):803–811. Available from: doi:10.1158/2159-8290.CD-17-1415.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. [Internet]. 2020[cited 2020 June 26];70(1):7–30. Available from: doi:10.3322/caac.21590.
3. National Cancer Institute. Cancer Statistics [Internet]. Bethesda (MD): United States Government, U.S. Department of Health and Human Services; 2018[cited 2020 June 26]. Available from: https://www.cancer.gov/about-cancer/understanding/statistics.
4. Schottenfeld D, Thun MJ, Linet MS, Cerhan JR, Haiman CA. (2017). Schottenfeld and Fraumeni Cancer Epidemiology and Prevention. 4th edition. New York, NY: Oxford University Press. 2017.
5. Wiltbank TB, Giordano GF, Kamel H, Tomasulo P, Custer B. Faint and prefaint reactions in whole-blood donors: an Analysis of predonation measurements and their predictive value. Transfusion. [Internet]. 2008[cited 2020 June 30];48(9). Avaliable from: doi: 10.1111/j.1537-2995.2008.01745.x.
6. Rizopouos, D. Joint models for longitudinal and time-to-event data: with applications in R. Boca Raton, FL: Chapman & Hall. 2012.

12 ATTACHMENTS

Attachment Item	Filename	Version #	Version Date	Category	Description
1. Cover Memo	Connect_CoverMemo_V1_08072020	V1	08/07/2020	Cover Memo	Cover memo for initial IRB submission for Connect
2. Protocol	Connect_Protocol_V1_08062020	V1	08/06/2020	Protocol	Study protocol for Connect
3. Informed Consent (Long Form)	Connect_Consent_V1_08072020	V1	08/13/2020	Consent - Redacted	Long-form version of participant consent form
4. HIPAA Authorization	Connect_HIPAA_V1_08072020	V1	08/07/2020	Other	HIPAA Authorization to use and disclose protected health information form
5. Study Personnel Page	Connect_StudyPersonnel_V1_08052020	V1	08/05/2020	Study Personnel List	List of study personnel for Connect
6. Participant Profile Electronic Questionnaire	Connect_QParticipantProfile_V1_05052020	V1	5/5/2020	Study Instrument	Paper version of electronic participant profile questionnaire

7. Module 1 Electronic Questionnaire: Background and Overall Health	Connect_QModule1_V1_06162020	V1	6/16/2020	Questionnaire	Paper version of electronic baseline questionnaire module 1
8. Module 2 Electronic Questionnaire: Medications, Reproductive Health, Exercise, and Sleep	Connect_QModule2_V1_04072020	V1	4/7/2020	Questionnaire	Paper version of electronic baseline questionnaire module 2
9. Module 3 Electronic Questionnaire: Smoking, Alcohol, and Sun Exposure	Connect_QModule3_V1_04202020	V1	4/20/2020	Questionnaire	Paper version of electronic baseline questionnaire module 3
10. Module 4 Electronic Questionnaire: Where You Live and Work	Connect_QModule4_V1_04202020	V1	4/20/2020	Questionnaire	Paper version of electronic baseline questionnaire module 4
11. Blood & Urine Specimen Electronic Questionnaire	Connect_QBlood&Urine_V1_07132020	V1	07/13/2020	Questionnaire	Paper version of electronic specimen (blood/urine) questionnaire
12. Saliva Specimen Electronic Questionnaire	Connect_QSalivaSpecimen_V1_07132020	V1	07/13/202	Questionnaire	Paper version of electronic saliva specimen (mouthwash) questionnaire

APPENDIX 1: CONNECT FOR CANCER PREVENTION GOVERNANCE STRUCTURE



The Connect for Cancer Prevention organizational structure includes leadership committees of the **DCEG Steering Committee** and the **Executive Committee**, which are supported by the functions of the **Coordinating Center** and the **Resource Access Committee** and external guidance from the **International Scientific Advisory Board** and the **Connect Patient Advisory Board.** Scientific review of the Connect for Cancer Prevention study will be conducted by the DCEG **Senior Advisory Group** and the **Biomarkers Research Group**.

- The **DCEG Steering Committee** includes a core group of DCEG Senior Investigators and key staff members, including the Project Directors and leadership in the Data Systems Development Team and Coordinating Center. The committee's responsibilities include the overall operational and scientific coordination of the project; monitoring study progress and budget; managing the contracts to the IHCS sites, support services, and the biorepository; obtaining study approvals from NIH and NCI and producing reports; and, establishing collaborative relationships with other institutions within NIH as well as external academic or industry groups. The DCEG Steering Committee is chaired by the Connect Senior Scientist and coordinated by the DCEG Study Manager. Together, they work closely with the site Co-PIs and other personnel from the IHCS, biorepository, support services contractor, and the

Data Systems Development personnel on the day-to-day study management and operational activities.

- The **Executive Committee** is comprised of DCEG Steering Committee members, IHCS Site Co-Principal Investigators, the Principal Investigator of the Participant Advisory Board, and the project lead from the support services provider. This committee provides scientific and operational direction of the study. The Executive Committee reviews and provides input on the study protocol as well as any proposed sub-studies for ancillary data and specimen collection. The Executive Committee will review and address unanticipated problems and adverse events during study enrollment and follow-up. The Executive Committee is led by the Connect Senior Scientist from DCEG and a rotating co-chair from among the Co-Principal Investigators of the IHCS sites on one-year terms. The committee meets at least monthly. Final governance or scientific decisions that are central to the integrity of the study are decided by the DCEG Director. In this regard, the Executive Committee is advisory to NCI leadership.
- The **Resource Access Committee** (RAC) will be chaired by the DCEG Senior Scientist, coordinated by a DCEG Staff Scientist, and include representation from the DCEG Connect team, IHCS sites, external investigators using Connect resources, patient representatives, and funding agencies. An Expert Review Panel composed of national and international experts in specific areas, will be convened to evaluate scientific proposals for data and specimen usage. Final approval of proposals will be made by the Connect PI, Dr. Montse Garcia-Closas. If there are any questions about whether a proposed analysis is authorized within the signed informed consent, the NIH IRB will be consulted.
- The **International Scientific Advisory Board** is comprised of experts from the extramural and international scientific community. These individuals have expertise in a range of domains including exposure assessment, biobanking, ethics, participant engagement, patient advocates, bioinformatics, and others. Scientific advisory board members serve 5-year terms and advise the NCI and DCEG Directors.
- The **Connect Patient Advisory Board** led by Carmit McMullen, PhD includes one to two representatives from each of the participating IHCS. The Board had its inaugural meeting in July 2020. They will meet at least monthly to advise cohort leadership from the perspective of patient and participant stakeholders. They will review study recruitment procedures, value proposition, communication materials among other study aspects. To ensure integration with the larger Connect leadership Dr. McMullen is a member of the Executive Committee and at least one member of the Connect Coordinating Center will attend the meetings, as appropriate.