WHI LILAC Cohort
Data Preparation and Use Guide
Investigator Datasets
October 2023
LILAC data as of February 19, 2023

Table of Contents

Page

1. Introduction

The Women's Health Initiative (WHI) Life and Longevity after Cancer (LILAC) study offers an important opportunity to advance cancer research by extending the original WHI studies to examine survivorship in women diagnosed with cancer during their participation in WHI. WHI participants with a diagnosis of one of the following eight cancers and consented to ongoing follow-up were approached with the baseline/consent mailing: breast, colorectal, lung, endometrial, ovarian (including fallopian tube and primary peritoneum cancers), melanoma, lymphoma, and leukemia. Deceased women were enrolled with a partial waiver of consent. The LILAC datasets include enrollment information, cancer diagnosis and treatment details, and long-term cancer outcomes for WHI participants who were included in the LILAC survivorship cohort during the first funding period (2/15/13 to 8/31/19).

Details of cancer treatment and outcomes are available from two sources: direct medical record abstraction and self-reported data from cancer survivors. Data from these sources vary considerably in terms of the level of detail, timeliness and specificity. Medical records abstraction and self-reported data are based on LILAC datasets as of February 19, 2023.

Data that was collected during the WHI clinical trial, the observational study, and the extension studies are available for use with approved paper proposals. Data dictionaries are at https://www.whi.org/datasets. These data can be linked with the LILAC study data by WHI participant ID.

To understand the diagnosis of cancer and its treatment on trajectories of aging, the accelerated aging phenotype, and age-related comorbidities, we established a cohort of age-matched WHI participants who were cancer-free as of February 28th, 2020. To facilitate longitudinal data analysis of the LILAC cancer survivors and the cancer-free cohort, the WHI CCC has prepared datasets populated with much of the extensive demographic, physical, mental, social health, and clinical event data collected under the WHI protocol. Cancer diagnosis date defines a key time point for each LILAC participant and her matched cancer-free controls, referred to as the index date and coded as time zero. The timing of other data elements is relative to the index date. Some participants selected as cancer-free controls subsequently developed invasive cancer during the additional follow-up time (after February 28th, 2020). The details of their cancer diagnosis can be determined by linking to the WHI Investigator dataset (https://www.whi.org/datasets/outcomes). For details about LILAC participant datasets, please see the companion document in this submission titled **WHI LILAC Participants and Age-Matched Cancer-Free Cohort – Dataset Guide**

2. Data File Setup

Each data set is provided as a tab-delimited ASCII file (.DAT) with a header row containing the variable names. The code needed to create SAS datasets from the ASCII files is provided in the files with the .SAS extension. To read the ASCII files into any other statistical program, refer to the INFILE statement in the SAS code file for the order of the variables and to the PROC FORMAT section for the values of all categorical variables. A PDF data dictionary is included for each dataset.

Not all data files have the same number of records since not every form was completed by each participant. The first variable in each file (ID) is the unique WHI participant ID. All files are linked by this identifier, which must be used to merge the data files. Order of the variables after ID generally matches the order of the questions on the most recent version of the form. In general, computed variables have been added at the end of the appropriate form. The form questions used in the computation of the computed variables have been noted in the variable descriptions.

3. Data Conventions

Dates

No actual dates are included in the data files. All dates have been converted to the number of days since index date (cancer diagnosis for LILAC participants and corresponding reference date for controls. When only the month and year were recorded, the first day of the month was used to convert the date. For abstraction form dates, missing months were assigned to July and missing days to the first day of the month. Calculated dates

that were inconsistent with diagnosis or surgery dates were reassigned to the day after diagnosis/surgery. Both start and stop dates were recorded for some treatments. Stop dates that were inconsistent with start dates were reassigned to one day after the start date.

Data Edits

The data entry system validates data at the time of entry to prevent invalid values for all categorical variables. Broad range checks applied to continuous variables have set out-of-range responses to missing. There still may be values that appear questionable to the user; **it is up to the user to examine all data before proceeding with data analysis.**

Consistency checks between data items on different forms were not done. Therefore, discrepancies do exist. Again, it is up to the user to carefully examine the data and determine which values are most appropriate for the specific analyses.

LILAC Forms

See https://www.whi.org/md/370/home for images of the baseline questionnaire, annual survey forms, and https://www.whi.org/md/370/data for images of abstraction forms.

Missing Data

Missing data can result from a form not being required, a form not being completed, a particular question on a form not being answered or not required because it was part of a skip pattern, or a question not being asked on all versions of a form. If an entire form is missing for a participant, that participant does NOT have a record in the data file.

Mark-All-That-Apply Questions

For LILAC baseline and survey forms, questions involving "mark all that apply" responses have been recoded. Each possible response has been turned into a yes/no variable with a "yes" coded if the response was marked and "no" otherwise. If all possible responses for the question were missing, all possible responses are set to missing. For abstraction forms, please see below for how each form's data was individually handled.

4. Specific Data Set Information

4.1 Participant Selection and Cancer Characteristics

Included in this resource is a dataset, *lilac_cancer_characteristics_inv.dat*, which contains enrollment and cancer detail information for LILAC cancer survivors. WHI participants were eligible to be included in the LILAC cohort if they had a confirmed invasive cancer diagnosis of breast, colorectal, lung, endometrial, ovarian, melanoma, lymphoma, or leukemia during WHI follow-up. Women reporting cancer history (other than non-melanoma skin cancer) prior to WHI enrollment were not eligible. Women still in active WHI follow-up were invited by mail to participate and were asked for their consent to medical records and tissue release. Women consenting to one or more components of LILAC were enrolled into the cohort. Deceased women were enrolled into the cohort under a partial waiver of consent. A small number of women were diagnosed with more than one LILAC-designated cancer at the same time (double primary). Variables that define the cancer(s) for which women were enrolled in LILAC are provided, as well as summary cancer treatment information derived from medical record abstraction or self-report. Details about the SEER coding of tumor characteristics can be found in the WHI data preparation guide.

4.2 LILAC Baseline Questionnaire (Form 340)

Participants with no cancer (other than non-melanoma skin cancer) prior to enrollment in WHI, still in active WHI follow-up, and a confirmed invasive cancer diagnosis during WHI follow-up of one of the selected cancers were sent *Form 340 - Baseline Questionnaire*. Non-responders to *Form 340* were called by WHI staff and asked to provide responses to questions 1-4 and 6 only. For the participants who responded by phone, the remaining *Form 340* responses are missing. The baseline dataset, *lilac_form_340_inv.dat*, includes self-reports of chemotherapy, hormonal/endocrine therapy, radiotherapy, and biological therapies. It also includes data on conditions common to many cancers or their treatments (e.g., lymphedema, cardiotoxicity, nephrotoxicity,

neurotoxicity), symptoms after treatment completion, pain, depression, anxiety, fatigue, distress, social support, weight, marital status, and insurance coverage.

4.3 Annual Survey (Form 370)

The first LILAC annual follow-up questionnaire was sent to the participants that responded to the baseline consent mailing by completing the LILAC baseline questionnaire. This dataset, *lilac_form_370_inv.dat*, includes current weight; weight at first cancer diagnosis; weight two years after cancer diagnosis; intentional weight loss after cancer diagnosis; current or ever use of selected medications; financial toxicity; cancer worry; social networks and social support; peripheral neuropathy; participation in cancer support groups or online peer support groups; cognitive functioning; physical functioning; exercise; body image; symptom within the past 4 weeks; selected nutrition/diet; and lymphedema.

4.4 Follow-up Survey (Form 371)

Additional measures collected in the second annual follow-up questionnaire, *lilac_form_371_inv.dat*, include: depression, anxiety, fatigue and distress; and unmet needs of cancer survivors (e.g., pain, physical functioning, memory/concentration, weight changes, end of life planning, etc.).

4.5 Computed Variables

One computed variable has been added to the end of *Form 340*, the baseline survey. Twelve recreational physical activity constructs have been added to the end of *Form 370*. The variable descriptions provide details about the computations.

4.6 Medical Record Abstraction

For a subset of LILAC-enrolled participants, medical records were collected in order to learn about cancer treatments and cancer recurrences. Medical records were retrieved and abstracted for LILAC women whose cancer diagnoses occurred in the year 2000 or later, and who consented to medical records release or were deceased and enrolled in LILAC under partial waiver of consent. Medical records for participants who were enrolled in Medicare fee for service (FFS) A+B at diagnosis were not abstracted.

For women with double primaries (two cancers diagnosed on the same day), two abstraction forms were usually completed, one for each cancer site.

Because of the relationship between ovarian, fallopian tube, and primary peritoneal cancers, all three of these cancer sites are defined as ovarian cancer for LILAC purposes, and one abstraction form was completed.

First course of treatment was defined as “all planned treatments administered after the original diagnosis of cancer in an attempt to destroy or modify the cancer tissue”. These guidelines use the documented first course of therapy (treatment plan) from the medical record, which ends when the treatment plan is completed (regardless of the time frame). Additionally, the first course of therapy ends when there is documentation of disease progression, recurrence, or treatment failure. We did not limit how far after diagnosis a treatment start date could be to be coded as first course of treatment.

The following medical record abstraction datasets are provided for each cancer site.

Cancer Site	Dataset name
Breast	lilac_form_342_inv.dat
Lung	lilac_form_343_inv.dat
Colorectal	lilac_form_344_inv.dat
Ovarian/fallopian tube/primary peritoneum	lilac_form_345_inv.dat
Endometrium	lilac_form_346_inv.dat
Melanoma	lilac_form_347_inv.dat
Leukemia	lilac_form_348_inv.dat
Lymphoma	lilac_form_349_inv.dat

Each dataset contains detailed information obtained from the medical records about cancer-directed surgery, radiation, molecular testing, endocrine and bone therapies (if applicable), cancer recurrences, and a summary variable for receipt of chemotherapy or other targeted therapy. Details about the agents that each participant received as part of first course chemotherapy/targeted therapy can be obtained by using ID to merge with the LILAC Chemo Meds file (*lilac_chemo_meds_inv.dat*). (See *Section 4.6.1*)

4.6.1 Drug Therapies

The dataset *lilac_chemo_meds_inv.dat* contains information about documented chemotherapy or targeted therapy agents that a participant received across all LILAC-designated cancers.

Drug therapies may have been written in the medical record as separate agents or as regimens. Regimens were separated into their agent components and details about each agent received, start of the therapy, and whether the treatment was neoadjuvant are provided. Only the first administration of a specific agent is included in the dataset. Other therapies that were given together with chemo/targeted therapies (e.g., steroids, chemo-protectants, or other ancillary agents) have been included if documented in the medical record as part of the regimen or treatment plan.

4.7 Medicare Data

Detailed cancer treatment information for participants who were enrolled in Medicare fee for service (FFS) A+B at diagnosis can be obtained by applying to use CMS data within the WHI Virtual Data Enclave.