Data Currently Available
(And How to Access It)

Chance Hohensee

Data Training

September 9, 2016

Introduction

- The WHI dataset is large and complex
- There are different cohorts within the 161,808 WHI participants, and different study periods during which data were collected
- Data collection is not uniform across cohorts and time periods

Purpose of this talk

- This talk will focus on the resources available to help you find what data are available for which cohorts, during which study periods. Specifically:
  - WHI Investigator datasets
  - Documentation
  - Genetic Data
  - VDE
  - BioLINCC

WHI Investigator Datasets

- Data downloadable from the WHI website:
  - Requires that have a P&P approved paper proposal and that you submit a signed Data Use Agreement
  - If you are conducting an ancillary study you must submit any new data that your study generated to the CCC
  - Once the requirements are submitted, you will receive account credentials that will let you sign in to the WHI website
  - Once you receive your credentials, you will have 90 days to download the files you need

WHI Investigator datasets (continued)

- To download data from the WHI website:
  - Sign In (upper left corner)
  - From the main website, click on Researchers
  - On the left side, under "Data", click on "Download"
  - Data can be downloaded by category
  - Data are in formatted text (*.dat) and accompanied by PDF data dictionaries and SAS programs that will read them into SAS

Data Documentation

- The dataset documentation page provides a wealth of useful information
- Access does not require sign in

Data Documentation (cont.)

- Here are some highlights:
  - Variable Search
  - WHI Data Preparation
  - Frequency for WHI 1993-2005 (table)
  - Frequency For Extension Studies (table)
  - Participant characteristics over time
  - Data Collection forms

WHI Data Preparation Document (Read this first)

- A short (~20 pgs.) overview of the WHI data, including:
  - An overview of the different data collection periods and forms, and which groups were included in the collection
  - General data file information
  - Information for specific datasets (e.g. diet, bone density, outcomes)
  - Tables detailing which outcomes were collected for a given cohort and study period

Frequency for WHI 1993-2005

- Data forms collected at each participant contact
- Includes:
  - Collection time points
  - Cohorts that received them
  - Variables in each form

Data forms collected (continued)

Frequency of Data Collection

Form #	Form Name	Screening CT and OS				CT																				OS			
		SV 0	SV 1	SV 2	SV 3	4-6 wk	6 m	1 Yr	4 wk	6 m	2 Yr	6 m	3 Yr	6 m	4 Yr	6 m	5 Yr	6 m	6 Yr	6 m	7 Yr	6 M	8 Yr	6 m	9 Yr	Close Out	An-nual	3 Yr	6 Yr	9 Yr
2	Eligibility Screen	X																											
4	HRT Washout		H																										
10	HT Manage/Safety Interview					H	H	H		H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H			
17	CaD Manage/Safety Interview								C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C			
20	Personal Information		X	X	X																								
25	Participant Treatment Assignment – HT[1]																												
28	Participant Treatment Assignment – CaD																									X			
30	Medical History		X																										
31	Reproductive History			X																									
32	Family History				X																								
33	Medical History Update						X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			

Key: X = All Participants
D = DM
H = HRT
C = CaD
O = OS
% = Percentage (subsample) of participants
BD = Bone Density sites

Form and variables	Timing and Subsample Notes (See table above for frequency of collection)
Form 30 - Medical History Questionnaire -- hospitalization history; history of medical conditions; history of heart, circulatory, or coagulation problems; history of arthritis, gallbladder disease, thyroid disease, hypertension, angina, peripheral arterial disease and related procedures, colonoscopy or sigmoidoscopy, stool guaiac; history of cancers (site, age at diagnosis); recent history of falls or syncopal episodes; history of fractures (site, age, number).	
Form 31 - Reproductive History -- age at menarche; history of menstrual irregularity and amenorrhea; history of menopausal symptoms; history of pregnancy, pregnancy outcomes, infertility; history of breast feeding; history of gynecologic and breast surgeries.	
Form 32 - Family History Questionnaire -- number of full-blooded sisters and brothers, daughters, and sons; parental age or date of death; relatives' history of' diabetes, myocardial infarction, stroke, cancers; fractures in parents (site, age).	
Form 39 - Cognitive Assessment -- expanded mini mental status examination.	HT cohort aged 65 and over.

Frequency For Extension Studies (table)

- Shows the forms that were used in the WHI extension, and the cohorts that received them.

Data Collection Schedule for Extension Studies

Form #	Data collection Study years: Oct 1 thru Sept 30	Extension Study 2005-2010					Extension Study 2010-2015					Extension Study 2015-2020				
		Yr 1 2005-06	Yr 2 2006-07	Yr 3 2007-08	Yr 4 2008-09	Yr 5 2009-10	Yr 1 2010-11	Yr 2 2011-12	Yr 3 2012-13	Yr 4 2013-14	Yr 5 2014-15	Yr 1 2015-16	Yr 2 2016-17	Yr 3 2017-18	Yr 4 2018-19	Yr 5 2019-20
--	Re-consent and Personal Information Update	X					X									
33	Medical History Update	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
150	Hormone Use Update	HT	HT	HT	HT	HT										
85	Mammogram	HT	HT													
134	Addendum to Medical History Update	X														
151	Activities of Daily Life (ADL)	X	X	X	X	X	X		X	X	X	X	X	X	X	X

X = All Extenstion participants

HT = Previous Hormone Trial participants

MRC = Previous Hormone Trial participants and all African American and Hispanic participants consented to Ext. 2010-2015

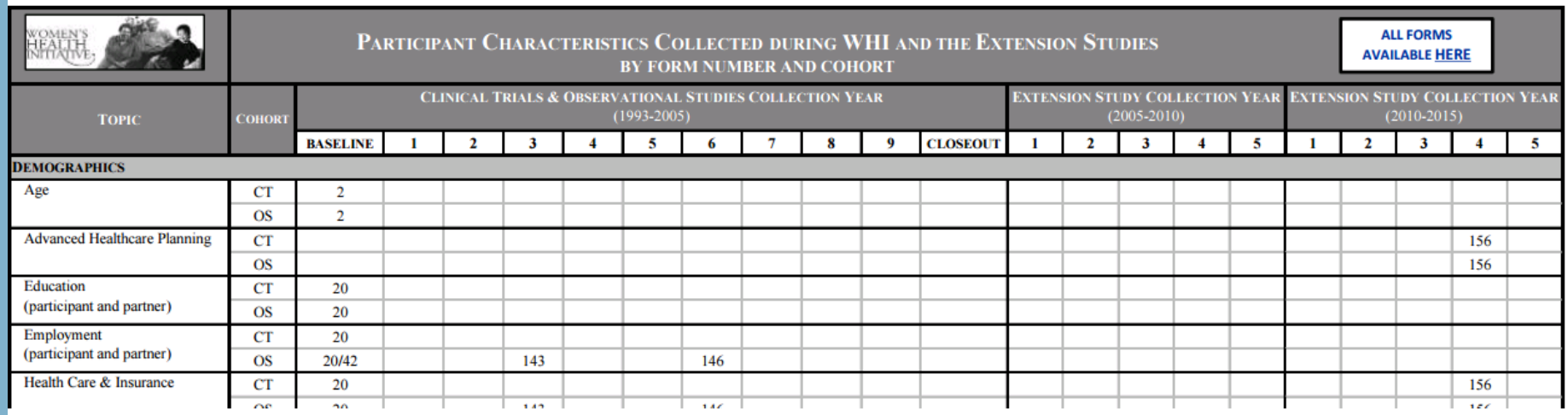
B = All participants who previously reported diagnosis of breast cancer

%DM = 4.6% of Dietary Modification Trial participants, with half in first part of Extension and half near end of Extension

LLS = All participants who consented to the Long Life Study

Participant Characteristics Over Time

- A table of various participant characteristics. Includes the following:
  - Cohorts in which the characteristics were tracked
  - Time point(s) in which they were collected
  - Form used for collection



PARTICIPANT CHARACTERISTICS COLLECTED DURING WHI AND THE EXTENSION STUDIES
BY FORM NUMBER AND COHORT

ALL FORMS AVAILABLE HERE

TOPIC	COHORT	CLINICAL TRIALS & OBSERVATIONAL STUDIES COLLECTION YEAR (1993-2005)											EXTENSION STUDY COLLECTION YEAR (2005-2010)					EXTENSION STUDY COLLECTION YEAR (2010-2015)				
		BASELINE	1	2	3	4	5	6	7	8	9	CLOSEOUT	1	2	3	4	5	1	2	3	4	5
DEMOGRAPHICS																						
Age	CT	2																				
	OS	2																				
Advanced Healthcare Planning	CT																				156	
	OS																				156	
Education (participant and partner)	CT	20																				
	OS	20																				
Employment (participant and partner)	CT	20																				
	OS	20/42			143			146														
Health Care & Insurance	CT	20																			156	

Data Dictionaries

- …have a header with general information about the dataset

Form 2 - Eligibility Screening

File Name
f2_ctos_inv.dat

Data as of	Population	Data Collected	1 row per	Rows	File Created
Sep 12, 2005	CT+OS	Baseline	Participant	161,808	Aug 29, 2012

Extension Study 2, Self Reported, Non-Adjudicated Outcomes for MRC Participants

File Name
outc_self_x2_mrc_inv.dat

Data as of	Population	Data Collected	1 row per	Rows	File Created
Sep 30, 2015	MRC	Ext2	Participant	22,316	Dec 1, 2015

Data Dictionaries (continued)

- Variable information with standard formats for categorical and continuous variables

HORM — Female hormones ever — Col#18

Did you ever use any female hormones like estrogen (Premarin) or progesterone (Provera)?These might be pills, skin patches, implants, creams, suppositories, shots, or birth control pills. (This does not include birth control pills you used before you were 50.)

Value	Description	N	%
0	No	52,644	32.5
1	Yes	104,983	64.9
	Missing	4,181	2.6

Usage Notes:
Not collected on all versions of Form 2.

F33BKKNEEDY — Days enrollment to F33 Broken Knee — Col#13

N	Missing	Min	Max	Mean	Std Dev
224	22,092	4738	7737	5,903.098	595.641

Data We Cannot Release

- Participant personal identifiers: name, address, zip code, date of birth
- Real WHI Participant ID
- Clinical Center ID
- Any date (including specific date, month or year)
- Geocodes (latitude/longitude coordinates), solar irradiance (i.e, latitude, Langleys or Watts) determined from participant zip code or clinical center ID
- CMS raw or computed data

WHI Virtual Data Enclave (VDE)

- Allows investigators to access data not otherwise available outside of the CCC
- Has a base cost of $3,000
- Requires proof of local IRB approval and signed Data Use Agreement
- Excludes participant name, street address, phone number, and Social Security Number
- Documentation and more information available here.

Genetic Data

- Two sources:
  - dbGaP
    - Provides molecular and phenotype data linked by dbGaP ID
    - Genotype data available in Plink format, imputed data available in MACH format
    - Best when using genotyped (i.e. non-imputed) data or when you need an entire dataset (i.e. all SNPs genotyped or imputed in a specific study)
  - WHI CCC
    - Best when you need specific imputed SNPs

dbGaP

- The main study page gives an overview of the WHI and links to sub-studies that utilize participants from the WHI cohort:
  - phs000386 WHI SHARe
  - phs000281 GO-ESP WHISP
  - phs000315 WHI GARNET
  - phs000503 WHISE
  - phs000227 PAGE WHI
  - phs000675 WHIMS+
  - phs000746 WHI Harmonized and Imputed GWAS

dbGaP Sub-study pages

- In general, sub-study pages provide:
  - Study overview
  - The number of samples and subjects (click on the Molecular Data tab)
  - Race/Ethnicity of the participants
- Example: SHARe

BioLINCC

- Biologic Specimen and Data Repository Information Coordinating Center
- Provides public access to NHLBI funded population-based biospecimen and data resources

WHI data on BioLINCC

- Does *not* require WHI investigator status as a condition of access
- BioLINCC currently includes WHI data as of September 2005
- An update is currently in the planning stages

Conclusion

- The WHI is complex - start with the WHI Data Preparation Document.
- Make use of the documentation available on the WHI SharePoint site – particularly the Data section under "researchers"
- If you have any questions, ask the helpdesk (helpdesk@whi.org)