Data Dictionary for CBCS Phase 3 Analysis Dataset

Table of Contents

Variable Name	Description	Data Source (Survey: Question Number)	Comments
STUDYID	Study ID		
RACE	Race, used in sampling 1 = Non-African American 2 = African American		
AGESEL	Age at diagnosis		
AGEINT	Age at interview		
AGEGR	5-year age groups 1 = 20-24 2 = 25-29 3 = 30-34 4 = 35-39 5 = 40-44 6 = 45-49 7 = 50-54 8 = 55-59 9 = 60-64 10= 65-69 11= 70-74	Recoded from AGESEL	
FRACT	Sampling probability based on age and race.		Non-African American age <50 = 0.4 Non-African American age 50+ = 0.15 African American age <50 = 1 African American age 50+ = 0.6

WT	Sampling weights	(1/FRACT) – inverse of the sampling probability	Use the WT variable if one is interested in calculating a weighted frequency estimate. If one is interested in calculating statistics such as chi-square on the weighted frequency estimate, use the WT and STRATA in SUDAAN or SAS Proc SurveyFreq to generate the correct weighted estimates and variances. Sampling weights are not needed in regression analysis. Always include age and race in the models to account for the sampling design.
STRATA	Sampling strata	111 = NonAA age <50 112 = NonAA age 50+ 113 = AA age <50 114 = AA age 50+	There are 4 sampling strata based on age and race. Subjects in each stratum have the same sampling probabilities.

Demographics

Variable Name	Description	Data Source (Ph3 = CBCS3 Baseline Survey) (Survey: Question Number)	Comments
MARITAL	Marital status 1 = Never married or lived as married 2 = Married or living as married 3 = Widowed 4 = Separated, divorced, or no longer living as married	Ph3: H1	
SELF_RACE	Self-reported race 1 = White 2 = Black/African American 3 = American Indian, Eskimo 4 = Asian or Pacific Islander 5 = Other	Ph3: H2	
OTHER_RACE	Other race, specify 2 = Multi-racial 3 = Hispanic/Latino 10 = Arab/Arab-Berber	Ph3: H2	Only available for other race (SELF_RACE=5).
ETHNICITY	Are you Hispanic? 1 = Hispanic 2 = Not Hispanic	Ph3: H3	

EDUC	Education 1 = 0 - 8 years 2 = 9-12 years, but not a high school graduate 3 = high school graduate (or GED) 4 = technical or business school 5 = some college 6 = college graduate 7 = post-graduate or professional degree	Ph3: H4	
EDUCAT	Education 1 = HS & Post HS 2 = College+ 3 = < HS	Recoded from EDUC	"<HS" is coded as the reference category.
INCOME	Family income 0 = < $5,000 1 = $5,000 to $10,000 2 = $10,000 to $15,000 3 = $15,000 to $20,000 4 = $20,000 to $30,000 5 = $30,000 to $50,000 6 = $50,000 to $100,000 7 = more than $100,000	Ph3: H12	
MONEY	Family income 1 = 15-30K 2 = 30-50K 3 = >50K 4 = <15K	Recoded from INCOME	"<15K" is coded as the reference category.

FIPS_DX	5-digit FIPS County Code	Obtained from Marc Emerson's Census data file	Updated 01/03/2025. Correct county data for 2 subjects.
COUNTY_DX	County of residence at diagnosis Text	Obtained from Marc Emerson's Census data file.	Updated 01/03/2025. Use the FIPS county code from Marc's data file. Correct county data for 2 subjects.
URBAN_RURAL_DX	Urban/rural status 1 = Urban 2 = Rural	Derived from county of residence at diagnosis	Updated 01/03/2025. County of residence at diagnosis. Based on Rural-urban Continuum Codes, 2013. Urban: RUCC_2013 codes 1-3 Rural: RUCC_2013 codes 4-9
AHEC_DX	Area Health Education Center (AHEC) regions 1 = UNC-Chapel Hill 2 = Area L 3 = Charlotte 4 = Eastern 5 = Greensboro 6 = Northwest 7 = South East 8 = Southern 9 = Wake	Derived from county of residence at diagnosis.	Updated 01/03/2025.

Variable Name	Description	Data Source (Survey: Question Number)	Comments
HH_SIZE	Household size	Ph3: H13	Number of people supported by this income.
Please note there are 2 sets of definition for poverty status (WCHS version)			
POVERTY_2010	Poverty Status 0 = no poverty (Above poverty level) 1 = poverty (Below poverty level)	Derived from INCOME & HH_SIZE	Based on the 2010 federal poverty guideline because the median year of diagnosis in CBCS3 (2008-2013) is 2010. We used the median value of each INCOME category and HH_SIZE to determine the poverty status.
POVERTY_200P	Poverty Status at or below 200% of the 2010 Federal Level 0 = no poverty (Above poverty level) 1 = poverty (Below poverty level)	Derived from INCOME & HH_SIZE	Based on the 200% of the 2010 federal poverty guideline.
Yearly specified poverty status (based on diagnosis year)			
POVERTY_YEARLY	Yearly specified poverty Status 0 = no poverty (Above poverty level) 1 = poverty (Below poverty level)	Derived from INCOME & HH_SIZE	Based on the 2008-2013 HHS yearly poverty guidelines. We used the median value of each INCOME category and HH_SIZE to determine the poverty status.
POVERTY_200P_YEARLY	Yearly specified poverty Status at or below 200% of the HHS Guidelines 0 = no poverty (Above poverty level) 1 = poverty (Below poverty level)	Derived from INCOME & HH_SIZE	Based on the 200% of the 2008-2013 HHS yearly poverty guidelines.

Family History of Breast and Ovarian Cancer

Variable Name	Description	Data Source (Survey: Question Number)	Comments
FFAMHXBC	First-degree family history of breast cancer - parents or sibling(s) 0 = No 1 = Yes	Ph3: B4, B7	
BC_MOM	Breast cancer in mother 0 = No 1 = Yes	Ph3: B4	
MOMAGEBC	Age mother diagnosed with breast cancer	Ph3: B4	Missing for those with no maternal history of breast cancer.
BC_DAD	Breast cancer in father 0 = No 1 = Yes	Ph3: B4	
DADAGEBC	Age father diagnosed with breast cancer	Ph3: B4	Missing for those with no paternal history of breast cancer.
BCSIBYN	Breast cancer in any sibling(s) 0 = No 1 = Yes 98= No siblings	Ph3: B7	Include brothers and sisters.
BC_SIB	Number of siblings with breast cancer 98 = No siblings	Ph3: B7	
MNAGSBC	Minimum age at which sibling diagnosed with breast cancer	Ph3: B7	Missing for those with no sibling history of breast cancer.

BCSISYN	Breast cancer in any sisters 0 = No 1 = Yes 98= No sisters	Ph3: B7	
BC_SIS	Number of sisters with breast cancer 98 = No sisters	Ph3: B7	
MNAGSISBC	Minimum age at which sister diagnosed with breast cancer	Ph3: B7	Missing for those with no sister history of breast cancer.
BCDAUGHYN	Breast cancer in any daughters 0 = No 1 = Yes 98= No daughters	Ph3: B9	
BC_DAUGH	Number of daughters with breast cancer 98 = No daughter	Ph3: B9	
MNAGDAUBC	Minimum age at which daughter diagnosed with breast cancer	Ph3: B9	Missing for those with no daughter history of breast cancer.
FFAMHXOC	First-degree family history of ovarian cancer – mother or sisters 0 = No 1 = Yes	Ph3: B4, B7	
OC_MOM	Ovarian cancer in mother 0 = No 1 = Yes	Ph3: B4	
MOMAGEOC	Age mother diagnosed with ovarian cancer	Ph3: B4	Missing for those with no maternal history of ovarian cancer.

OCSIBYN	Ovarian cancer in any sisters 0 = No 1 = Yes 98= No sisters	Ph3: B7	
OC_SIB	Number of sisters with ovarian cancer 98 = No sisters	Ph3: B7	
MNAGSOC	Minimum age at which sister diagnosed with ovarian cancer	Ph3: B7	Missing for those with no sister history of ovarian cancer.
BCOC_MOM	Breast or ovarian cancer in mother 0 = No 1 = Yes	Ph3: B4	
MOMAGEBCOC	Age mother diagnosed with breast or ovarian cancer	Ph3: B4	Missing for those with no maternal history of breast or ovarian cancer.
BCOCSIBYN	Breast or ovarian cancer in any sisters 0 = No 1 = Yes 98= No sisters	Ph3: B7	
BCOC_SIB	Number of sisters with breast or ovarian cancer 98 = No sisters	Ph3: B7	
MNAGSBCOC	Minimum age at which sister diagnosed with breast or ovarian cancer	Ph3: B7	Missing for those with no sister history of breast or ovarian cancer.

BCMOMLT50	Mother diagnosed with breast cancer before age 50 0 = No 1 = Yes	Recoded from BC_MOM, MOMAGEBC, and B4K (from survey)	Count as No if BC_MOM=No. If MOMAGEBC is unknown and B4K (age at interview) is under 50, count as Yes.
BCOCMOMLT50	Mother diagnosed with breast or ovarian cancer before age 50 0 = No 1 = Yes	Recoded from BCOC_MOM, MOMAGEBCOC, and B4K (from survey)	Count as No if BCOC_MOM=No. If MOMAGEBCOC is unknown and B4K (age at interview) is under 50, count as Yes.
BCSLT50	Number of sisters diagnosed with breast cancer before age 50 98 = No sisters	Ph1: B7 Ph2: B7	Sister is not counted if age of diagnosis is unknown. However, if sister's age at interview is under 50, count the sister
BCOCSLT50	Number of sisters diagnosed with breast or ovarian cancer before age 50 98 = No sisters	Ph1: B7 Ph2: B7	Sister is not counted if age of diagnosis is unknown. However, if sister's age at interview is under 50, count the sister
FBCLT50	Number of 1st degree female relatives (mother, sisters) diagnosed with breast cancer before age 50	Recoded from BCMOMLT50, BCSLT50	If BCMOMLT50=unknown and BCSLT50=0 then count as unknown. If BCMOMLT50=0 and BCSLT50=unknown then count as unknown. Otherwise, sum non-missing data from BCMOMLT50 and BCSLT50.
FBCOCLT50	Number of 1st degree female relatives (mother, sisters) diagnosed with breast or ovarian cancer before age 50	Recoded from BCOCMOMLT50, BCOCSLT50	If BCOCMOMLT50=unknown and BCOCSLT50=0 then count as unknown. If BCOCMOMLT50=0 and BCOCSLT50=unknown then count as unknown. Otherwise, sum non-missing data from BCOCMOMLT50 and BCOCSLT50.

Menstrual History
(Exclude exposure after age of diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
AGEMENA	Age at menarche (range = 7-19)	Ph3: C1	One woman who had never menstruated is coded as missing.
MENARCHE	Age at menarche 1 = < 12 years 2 = 12+	Recoded from AGEMENA	"12+" is coded as the reference category.
MENA13G	Age at menarche 1 = < 13 years 2 = 13+	Recoded from AGEMENA	Cut point obtained from the median of CBCS 1 & 2 controls. "13+" is coded as the reference category.
MENO	Type of menopause experienced 1 = premenopausal 2 = natural menopause 3 = surgical, uterus and 2 ovaries removed 4 = surgical, uterus and 1 ovary removed 5 = surgical, uterus and no ovaries removed 6 = surgical, uterus removed, ovaries unknown 7 = surgical, uterus intact, 2 ovaries removed 8 = surgical, uterus intact, 1 ovary removed 9 = surgical, uterus intact, ovaries intact 10= surgical, uterus intact, ovaries unknown 11= surgical, uterus unk, 2 ovaries removed 12= surgical, uterus unknown, 1 ovary removed 13= surgical, uterus unknown, ovaries intact 14= surgical, uterus unknown, ovaries unknown 15= menopause due to chemo or radiation 16= other menopause 17= Never stopped cycling, but is taking hormone replacement	Ph3: C4-C10	If subject experienced menopause after age of diagnosis, she would be classified as premenopausal for this variable.

MENODATE	Date of menopause	Ph3: C4-C10	This variable goes with the variable MENO. Missing for premenopausal (MENO=1) women.
AGEMENO	Age at menopause	Ph3: C4-C10	This age variable is for the variable MENO. Missing for premenopausal (MENO=1) women.
POSTMENO	Menopausal status 0 = Premenopausal 1 = Postmenopausal	Derived from MENO and AGESEL	For women under age 50, postmenopausal status was assigned to women who had undergone natural menopausal, bilateral oophorectomy, or irradiation to the ovaries; in women aged 50 or older, menopausal status was assigned on the basis of cessation of menstruation.
AGE_POSTMENO	Age at menopause	Derived from AGEMENO and POSTEMNO	This variable goes with POSTMENO. Missing for premenopausal (POSTMENO=0) women.
MENOSURG_DATE	Date of menopausal surgery	Derived from MENO and Ph3: C5	Defined for those with MENO codes 3-14. This is not the same as MENODATE. This is the date of the surgery. Some people experienced menopausal symptoms sometime after surgery; the info would be captured in MENODATE.
AGE_MENOSURG	Age at menopausal surgery	Derived from MENO and Ph3: C5	Defined for those with MENO codes 3-14. This is not the same as AGEMENO.

Pregnancy and Lactation
(Exclude exposure after age of diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
PREGNUM	Number of pregnancies (range: 0-13)	Ph3: C11	Exclude pregnancies after age of diagnosis.
FTPEVER	Ever had full-term pregnancy 0 = No 1 = Yes	Ph3: C12	Full-term pregnancy is defined as 7+ months pregnancy duration or pregnancy resulting in a live birth.
PARITY	Number of full-term pregnancies (range: 0-11)	Ph3: C12	Full-term pregnancy is defined as 7+ months pregnancy duration or pregnancy resulting in a live birth.
AGEFFTP	Age first full-term pregnancy (range: 11-44)	Ph3: C12	Missing for nulliparous.
AGELFTP	Age last full-term pregnancy (range: 13-49)	Ph3: C12	Missing for nulliparous.
DATELFTP	Date of last full-term pregnancy	Ph3: C12	SAS date.
FTPYR	Years since last FTP		Missing for women who never had FTP. If pregnant at time of diagnosis that resulted in a FTP, FTPYR=0.

FTP_INT	Interval between full-term pregnancies 0 = Nulliparous 1 = 1 FTP 2 = 2 FTP, interval <= 1 year 3 = 2 FTP, interval >1 year 4 = 3+ FTP, at least 1 interval <=1 year 5 = 3+ FTP, all intervals >1 year 9 = unable to determine one or more pregnancy intervals	Ph3: C12	
PREG_DX	Pregnant at diagnosis/selection or diagnosed/selected within 2 years after pregnancy 0 = No 1 = Yes	Derived from AGESEL and Ph3: C12	Include all pregnancies regardless of outcome.
PREG_DX5YR	Pregnant at diagnosis/selection or diagnosed/selected within 5 years after pregnancy 0 = No 1 = Yes	Derived from AGESEL and Ph3: C12	Include all pregnancies regardless of outcome.
PREGCUR	Pregnant at time of diagnosis/selection 0 = No 1 = Yes	Derived from AGESEL and Ph3: C12	PREGCUR=1 is different from FTPYR=0 because PREGCUR includes all pregnancies regardless of outcome.
LIVEVER	Ever had live birth 0 = No 1 = Yes	Ph3: C12	Only include pregnancies that resulted in live birth.
NUMLIVEB	Number of live birth pregnancies (range: 0-11)	Ph3: C12	

AGEFLIVE	Age first live birth (range: 11-44)	Ph3: C12	Missing for women who never had live births.
AGELLIVE	Age last live birth (range: 13-49)	Ph3: C12	Missing for women who never had live births.
TERMEVER	Ever had terminated pregnancy 0 = No 1 = Yes	Ph3: C12	Available in CBCS Phase 3 only. Terminated pregnancy = Stillbirths, spontaneous miscarriage, or induced abortion
NUMBTERM	Number of terminated pregnancies (range: 0-11)	Ph3: C12	Available in CBCS Phase 3 only.
AGEFTERM	Age first terminated pregnancy (range: 10-44)	Ph3: C12	Available in CBCS Phase 3 only. Missing for women who never had terminated pregnancy.
ECTEVER	Ever had ectopic/tubal pregnancy 0 = No 1 = Yes	Ph3: C12	
NUMBECT	Number of ectopic/tubal pregnancies (range: 0-4)	Ph3: C12	
AGEFECT	Age first ectopic/tubal pregnancy (range: 16-40)	Ph3: C12	Missing for women who never had induced ectopic/tubal pregnancy.
PRETERM	Ever had preterm birth 0 = No 1 = Yes	Ph3: C12	Definition is slightly different from CBCS1 & 2 because we don't have the stillbirth category in CBCS3. Preterm is defined as <9 months pregnancy duration for single live birth, multiple live birth, and still birth (if data is available). Note: definition of term-pregnancy is different from PARITY and FTPEVER.

LACTEVER	Ever lactated 0 = No 1 = Yes	Ph3: C12	Nulliparous counted as never lactated.
SUMLACT	Lifetime duration lactation (months) (range: 0-104)	Ph3: C12	0 = never lactated or lactated < 2 weeks Some subjects have LACTEVER=1 and SUMLACT=0 (less than 2 weeks of lactation).
LACTMON	Lifetime duration lactation 1 = <= 3 months 2 = 4+ months 3 = Never breast fed	Derived from SUMLACT and LACTEVER	
AGEFLACT	Age at first lactation (range: 11-47)	Ph3: C12	Missing for those who had never lactated.
LACTFAGE	Age at first lactation (years) 1 = <=24 years 2 = 25+ 3 = Never breast fed	Derived from AGEFLACT and LACTEVER	
AGELLACT	Age at last lactation (range: 13-49)	Ph3: C12	Missing for those who had never lactated.
LACTLAGE	Age at last lactation (years) 1 = <=29 years 2 = 30+ 3 = Never breast fed	Derived from AGEFLACT and LACTEVER	
PREGLAC	Number of pregnancies for which lactated (range: 0-8)	Ph3: C12	

LACTKIDS	Number of children breastfed 1 = 1 child 2 = 2+ children 3 = Never breastfed	Derived from PREGLAC and LACTEVER	
LAVGMON	Number of months breastfeeding per child (range: 0-36)	Derived from SUMLACT and NUMLIVEB.	0 = never lactated or lactated < 2 weeks
LACTAVG	Number of months breastfeeding per child 1 = 0-3.9 months 2 = 4+ 3 = Never breast fed	Derived from LAVGMON and LACTEVER	
LONGLACT	The greatest number of months that any individual child was breastfed (range: 0-48)	Ph3: C12	0 = never lactated or lactated < 2 weeks
LACTSUPP	Number of pregnancies for which milk production was suppressed by medication (range: 0-8)	Ph3: C12	
SUPPRESS	Lactation suppressant use 1 = Ever 2 = Never	Derived from LACTSUPP	Note: never is coded as 2.
SUMMILK	Lifetime duration baby drank milk only and have nothing else to eat or drink (months) (range: 0-40)	Ph3: C12	Available in CBCS Phase 3 only.

AFTP	Parity and age at first full term pregnancy 1 = 1 kid, age FTP <26 2 = 1 kid, age FTP 26+ 3 = 2+ kids age FTP <26 4 = 2+ kids age FTP 26+ 5 = Nulliparous	Derived from PARITY and AGEFFTP	
KBFED	Parity and breastfeeding composite 1 = Parity 1-2, never breastfed 2 = Parity 1-2, ever breastfed 3 = Parity 3+, never breastfed 4 = Parity 3+, ever breastfed 5 = Nulliparous	Derived from PARITY and LACTEVER	

Oral Contraceptives Use
(Exclude exposure after age of diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
OCEVER	Ever use oral contraceptives 0 = Never 1 = Ever	Ph3: E1, E3, E4	Ever user is defined as 3+ months of OC use. Exclude OC use after age of diagnosis
OCUSE	Use of oral contraceptives 0 = Never 1 = Current 2 = Former	Ph3: E1, E3, E4	
OCMONTHS	Number of months used OC	Derived from OCUSE, AGESEL and Ph3: E3, E4	Include months of use for never user. Never user: range 0-2 Ever user: range 3-384
OC_REC	Years since last used OC 0 = current user	Derived from OCUSE, AGESEL and Ph3: E3	
OCAGE	Age first used OC	Derived from OCEVER and Ph3: E2	
OCLASTAGE	Age last used OC	Derived from OCEVER and Ph3: E3	

OCYRS	OC use durations 1 = <5 years 2 = 5-10 3 = >10 4 = never	Derived from OCEVER, OCMONTHS	
OCAGEYR	Age at 1st OC use and duration 1 = Age >=20, duration <5 2 = Age >=20, duration 5-10 3 = Age >=20, duration >10 4 = Age <20, duration <5 5 = Age <20, duration 5-10 6 = Age <20, duration >10 7 = Never	Derived from OCAGE, OCYRS	

Endocrine Therapy
(Exclude use before age of diagnosis. That is, if age at last use<age of diagnosis, count as No)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
TAMOXIFEN_S	Initiated Tamoxifen after dx (self-reported) 0 = No 1 = Yes	Ph3: E20a	Women who took Tamoxifen (started and completed) before diagnosis are classified as "No". Women who took Tamoxifen at some point at or following diagnosis are classified as "Yes".
RALOXIFENE_S	Initiated Raloxifene after dx (self-reported) 0 = No 1 = Yes	Ph3: E20b	Women who took Raloxifene (started and completed) before diagnosis are classified as "No". Women who took Raloxifene at some point at or following diagnosis are classified as "Yes".
ARIMIDEX_S	Initiated Arimidex after dx (self-reported) 0 = No 1 = Yes	Ph3: E20c	Women who took Arimidex (started and completed) before diagnosis are classified as "No". Women who took Arimidex at some point at or following diagnosis are classified as "Yes".
AROMASIN_S	Initiated Aromasin after dx (self-reported) 0 = No 1 = Yes	Ph3: E20d	Women who took Aromasin (started and completed) before diagnosis are classified as "No". Women who took Aromasin at some point at or following diagnosis are classified as "Yes".
FEMARA_S	Initiated Femara after dx (self-reported) 0 = No 1 = Yes	Ph3: E20e	Women who took Femara (started and completed) before diagnosis are classified as "No". Women who took Femara at some point at or following diagnosis are classified as "Yes".
ENDOCRINE_S	Initiated Endocrine Therapy after dx (self-reported) 0 = No 1 = Yes	Derived from TAMOXIFEN_S, RALOXIFENE_S, ARIMIDEX_S, AROMADIN_S, and FEMATA_S.	Define as "Yes" if any one of the 5 Endocrine drugs is coded as "Yes".

Hormone Replacement Therapy
(Exclude exposure after age of diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
ESTROGEN	Estrogen replacement therapy (with/without progestin) 0 = Never use 1 = Ever use (3+ months)	Ph3: E22a+E22b+E22d	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of diagnosis. Subjects with unknown months of hormone use are assumed to have 3+ months use and classified as ever user.
ESTROMON	Number of months used estrogen (with/without progestin)	Ph3: E22a+E22b+E22d	Include months of use for never user. Never user: range 0-2 Ever user: range 3-528
AGEFEST	Age first ever used estrogen (with/without progestin) (range: 17-66)	Ph3: E22a+E22b+E22d	
AGELEST	Age last used estrogen (with/without progestin) (range: 19-74)	Ph3: E22a+E22b+E22d	
EST_REC	Years since last use estrogen (with/without progestin) (range: 0-46)	Derived from AGESEL & AGELEST.	For current users, EST_REC=0.
EST_USE	Estrogen replacement therapy (with/without progestin) 0 = Never user 1 = current user 2 = past user	Derived from ESTROGEN & EST_REC	

ESTONLY	Estrogen replacement therapy only (no progestin) 0 = Never use 1 = Ever use (3+ months)	Ph3: E22(b-c) + E22 (d-e)	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of diagnosis. Exclude the times when estrogen were used together with progestin. Subjects with unknown months of hormone use are assumed to have 3+ months use and classified as ever user.
ESTONLY_MON	Number of months used estrogen only (no progestin)	Ph3: E22(b-c) + E22 (d-e)	Include months of use for never user. Never user: range 0-2 Ever user: range 3-528
AGEFEONLY	Age first ever used estrogen only (range: 17-66)	Ph3: E22(b-c) + E22 (d-e)	
AGELEONLY	Age last used estrogen only (range: 18-74)	Ph3: E22(b-c) + E22 (d-e)	
ESTONLY_REC	Years since last use estrogen only (range: 0-48)	Derived from AGESEL & AGELEONLY.	For current users, ESTONLY_REC=0.
ESTONLY_USE	Estrogen replacement therapy only (no progestin) 0 = Never user 1 = current user 2 = past user	Derived from ESTONLY & ESTONLY_REC	

ESTPROG	Estrogen + progestin replacement therapy 0 = Never use 1 = Ever use (3+ months)	Ph3: E22a+E22c+E22e	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of diagnosis. Subjects with unknown months of hormone use are assumed to have 3+ months use and classified as ever user.
EPMON	Number of months used estrogen+progestin	Ph3: E22a+E22c+E22e	Include months of use for never user. Never user: range 0-2 Ever user: range 3-368
AGEFEP	Age first ever used estrogen+progestin (range: 19-61)	Ph3: E22a+E22c+E22e	
AGELEP	Age last used estrogen+progestin (range: 19-74)	Ph3: E22a+E22c+E22e	
EP_REC	Years since last use estrogen+progestin (range: 0-46)	Derived from AGESEL & AGELEP	For current users, EP_REC=0.
EP_USE	Estrogen+progestin replacement therapy 0 = Never user 1 = current user 2 = past user	Derived from ESTPROG & EP_REC	
PROGEST	Progestin replacement therapy only 0 = Never use 1 = Ever use (3+ months)	Ph3: E22f	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of diagnosis. Subjects with unknown months of hormone use are assumed to have 3+ months use and classified as ever user.
PROGMON	Number of months used progestin only	Ph3: E22f	Include months of use for never user. Never user: range 0-2 Ever user: range 3-120

AGEFPROG	Age first ever used progestin only (range: 11-55)	Ph3: E22f	Note: one subject took provera at age 11 for very heavy periods.
AGELPROG	Age last used progestin only (range: 11-60)	Ph3: E22f	
PROG_REC	Years since last use progestin only (range: 0-43)	Derived from AGESEL & AGELPROG	For current users, PROG_REC=0.
PROG_USE	Progestin replacement therapy only 0 = Never user 1 = current user 2 = past user	Derived from PROGEST & PROG_REC	
ANYHRT	Any hormone replacement therapy 0 = Never use 1 = Ever use (3+ months)	Ph3: E22a+E22b+E22d+ E22f	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of diagnosis. Subjects with unknown months of hormone use are assumed to have 3+ months use and classified as ever user.
ANYMON	Number of months used any hormone replacement therapy	Ph3: E22a+E22b+E22d+ E22f	Include months of use for never user. Never user: range 0-2 Ever user: range 3-528
AGEFANY	Age first ever used any hormone replacement therapy (range: 11-66)	Ph3: E22a+E22b+E22d+ E22f	
AGELANY	Age last used any hormone replacement therapy (range: 11-74)	Ph3: E22a+E22b+E22d+ E22f	

HRT_REC	Years since last use any HRT	Derived from AGESEL & AGELANY	For current users, HRT_REC=0.
HRT_USE	Any hormone replacement therapy 0 = Never user 1 = current user 2 = past user	Derived from ANYHRT & HRT_REC	
SUMMEST	Types of estrogen used 1 = Premarin 2 = Estropipate 3 = Estradiol 4 = Esterified Estrogens 5 = DES 6 = Estrovis 7 = Chlortianisene 8 = Estratest 9 = Transdermal Estrogen 10= Multiple Estrogens 99= Unknown	Ph3: E22a+E22b+E22d	Estrogen use with/without progestin.
HORMONE	Type of hormone use 1 = Unopposed estrogen only 2 = Progestin only 3 = Progestin always taken along with estrogen 4 = Progestin sometimes taken along with estrogen 5 = Estrogen and progestin both taken, but never simultaneously 9 = Unknown		Defined for those ever used HRT.

POSTHRT	HRT use among postmenopausal women 0 = Never 1 = Ever	Derived from POSTMENO & ANYHRT IF POSTMENO=1 & ANYHRT=1 THEN POSTHRT=1	SAS label corrected 9/22/2016. Missing for premenopausal women. Lifetime HRT use is counted in this variable. If a postmenopausal woman had only used HRT before menopause, she would be counted as EVER. For example, if a subject's age of menopause is 47 and only had HRT use for 9 months from age 43-44; she would be counted as EVER for this variable.
HRTYRS	Lifetime HRT use duration among postmenopausal women 1 = <5 years 2 = 5-10 3 = >10 4 = Never	Derived from POSTHRT & ANYMON	SAS label corrected 9/22/2016. Missing for premenopausal women. Lifetime HRT use including premenopausal is counted in this variable.

Radiation

Variable Name	Description	Data Source (Survey: Question Number)	Comments
HIGHRAD	History of high dose radiation to chest area 0 = No 1 = Yes	Ph3: D20, D21, D33, D34	Exclude exposure after age of diagnosis. "Yes" is defined as any of the following: 1) Ever had coronary catheterization or angioplasty 2) Had axilla or lung treated or monitored with radiation 3) Had breast treated or monitored with radiation, but did not have breast cancer
ION2	Jobs with ionizing radiation exposure 0 = No 1 = Yes	Ph3: G2	1990 Occupational Classification Codes: 84 (physicians), 95 (RN), 206 (radiologic technicians), 207 (LPN)
JOBSRAD2	Jobs with potential radiation exposure 0 = No 1 = Yes	Ph3: G2	1990 Occupational Classification Codes: 84 (physicians), 95 (RN), 203 (clinical lab technicians), 205 (health record technologists and technicians), 206 (radiologic technicians), 207 (LPN), 447 (nursing aides, orderlies, and attendants)

History of other cancer

OTHER_CANCER	History of other cancer 0 = No 1 = Yes	Ph1: E5 Ph2: D2, D3	Excluding breast cancer. Age of cancer diagnosis <= AGESEL.

Physical Activity

Variable Name	Description	Data Source (Survey: Question Number)	Comments
TOTPHYS_PRE	Total minutes of physical activity per week – 3 months prior to breast cancer diagnosis	Ph3: F6, F7, F8, F9, F10, F11	Total minutes of physical activity per week was calculated as 2*vigorous activity + moderate activity.
TOTPHYS_PRE_CAT	Total minutes of physical activity per week in category (pre diagnosis) 0 = Sedentary 1 = Insufficiently active 2 = Sufficiently active	Derived from TOTPHYS_PRE	Those women reporting $\geq$ 150 minutes of total physical activity per week were classified as sufficiently active, those reporting >0 minutes but <150 minutes per week were classified as insufficiently active, and those reporting no activity were classified as sedentary.
TOTPHYS_POST	Total minutes of physical activity per week – post diagnosis (7 days before interview)	Ph3: F12, F13, F14, F15, F16, F17	Total minutes of physical activity per week was calculated as 2*vigorous activity + moderate activity.
TOTPHYS_POST_CAT	Total minutes of physical activity per week in category (post diagnosis) 0 = Sedentary 1 = Insufficiently active 2 = Sufficiently active	Derived from TOTPHYS_POST	Those women reporting $\geq$ 150 minutes of total physical activity per week were classified as sufficiently active, those reporting >0 minutes but <150 minutes per week were classified as insufficiently active, and those reporting no activity were classified as sedentary.
TOTPHYS_DIFF	Change in total minutes of physical activity per week between pre and post diagnosis	Calculated from (TOTPHYS_POST – TOTPHYS_PRE)	

TOTPHYS_DIFF_CAT	Change in total minutes of physical activity per week between pre and post diagnosis (category) 0 = Increased activity by ≥ 30 minutes 1 = Decreased activity by ≥ 30 minutes 2 = No change, within (<) 30 minutes	Derived from TOTPHYS_DIFF	
TOTMETHRS_PRE	Total weekly metabolic equivalent task (MET) **hours** per week – 3 months prior diagnosis	Ph3: F6, F7, F8, F9, F10, F11	MET values of 4.0 and 8.0 were assigned to moderate-intensity and vigorous-intensity activity.
TOTMETHRS_POST	Total weekly metabolic equivalent task (MET) **hours** per week – post diagnosis (7 days before interview)	Ph3: F12, F13, F14, F15, F16, F17	MET values of 4.0 and 8.0 were assigned to moderate-intensity and vigorous-intensity activity.
TOTMETHRS_DIFF	Change in total weekly MET **hours** per week between pre and post diagnosis	Calculated from (TOTMETHRS_POST – TOTMETHRS_PRE)	
TOTMETMINS_PRE	Total weekly metabolic equivalent task (MET) **minutes** per week – 3 months prior diagnosis	Ph3: F6, F7, F8, F9, F10, F11	MET values of 4.0 and 8.0 were assigned to moderate-intensity and vigorous-intensity activity.
TOTMETMINS_POST	Total weekly metabolic equivalent task (MET) **minutes** per week – post diagnosis (7 days before interview)	Ph3: F12, F13, F14, F15, F16, F17	MET values of 4.0 and 8.0 were assigned to moderate-intensity and vigorous-intensity activity.
TOTMETMINS_DIFF	Change in total weekly MET **minutes** per week between pre and post diagnosis	Calculated from (TOTMETMINS_POST – TOTMETMINS_PRE)	

Alcohol Use

Variable Name	Description	Data Source (Survey: Question Number)	Comments
ALCOHOL	Ever used alcohol 0 = No 1 = Yes	Ph3: F18	Assume first started using alcohol before diagnosis of breast cancer.

Smoking
(Exclude exposure after age of diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
EVERSMOK	Smoking status 0 = Never 1 = Ever	Ph3: F21, F22	
SMOKERS2	Smoking status 0 = Never 1 = Former 2 = Current	Ph3: F21, F22, F23, F24	New definition as of 9/26/2014, this version will be used from now on. If age of smoking cessation ≥ age of diagnosis, the subject would be considered as current smoker. The difference from the old version (SMOKERS) is: If age of smoking cessation is the same as age of diagnosis/selection, the subject would be classified as **current** smoker.
PASSONLY	Never active smokers: exposure to ETS 1 = Exposed to ETS after age 18 2 = Unexposed to active smoke or ETS	Ph3: F21, F22, F30	
SMKGP2	Smoking status 1 = Passive smoking only 2 = Former 3 = Current 4 = No active & no passive	Derived from PASSONLY and SMOKERS2	New definition as of 9/26/2014, this version will be used from now on.

DURATION	Smoking duration 1 = <=10 years 2 = 11 - 20 years 3 = > 20 years 4 = Never	Ph3: F21, F22, F25	
DURGP	Duration of active smoking 1 = Passive smoking only 2 = <=10 years 3 = 11 - 20 years 4 = > 20 years 5 = No active & no passive	Derived from PASSONLY and Ph3: F25	
DOSE	Smoking dose (per day) 1 = < 1/2 pack 2 = 1/2 - 1 pack 3 = >1 pack 4 = Never	Ph3: F21, F22, F26	
DOSEGP	Dose of active smoking (packs per day) 1 = Passive smoking only 2 = < 1/2 pack 3 = 1/2 - 1 pack 4 = >1 pack 5 = No active & no passive	Derived from PASSONLY and Ph3: F26	
AGE_SMK	Age at initiation of smoking (range: 7-60)	Ph3: F22	
AGESTART	Age at initiation of active smoking 1 = Passive smoking only 2 = < 18 3 = 18+ 4 = No active & no passive	Derived from PASSONLY and Ph3: F22	

YRQUITSMK2	Years since quitting smoking – former smokers (range: 1-54)	Derived from AGESEL, SMOKERS2 and Ph3: F24	New definition as of 9/26/2014, this version will be used from now on. Missing for never and current smokers.
YRSTOP2	Years since stopped active smoking 1 = Passive smoking only 2 = < 10 years 3 = 10+ years 4 = No active & no passive	Derived from PASSONLY and YRQUITSMK2	New definition as of 9/26/2014, this version will be used from now on. Missing for current smokers.

Anthropometry

Variable Name	Description	Data Source (Survey: Question Number)	Comments
BMI	BMI based on self-reported usual adult height and weight from 1 year before diagnosis of breast cancer. (range: 14.20-73.96)	Ph3: E23, E24	
BMI18	BMI based on self-report usual adult height and weight at age 18.	Ph3: E23, E30 (18 yo)	
BMI35	BMI based on self-report usual adult height and weight at age 35.	Ph3: E23, E30 (35 yo)	Missing for AGESEL<35.
HEIGHT	Nurse measured height in cm (range: 137.2-188)	Ph3: L1	Anthropometric measurement at interview. Converted from inches to cm using this formula: cm = inch x 2.54
HIGHMED	Height (cm) - median 1 = >=162.5 2 = <162.5	Derived from HEIGHT	Cut points obtained from median of overall controls from CBCS 1 & 2. "<162.5 cm" is coded as the reference category.
HIGHTER	Height (cm) - tertiles 1 = 160-<165 2 = >=165 3 = <160	Derived from HEIGHT	Cut points obtained from tertiles of overall controls from CBCS 1 & 2. "<160 cm" is coded as the reference category.
WEIGHT	Nurse measured weight in kg (range: 38.6-188.9)	Ph3: L2	Anthropometric measurement at interview. Converted from pounds to kg using this formula: kg = pound x 0.4536.

ANTHBMI	BMI based on nurse measured anthropometric data (range: 16.05-69.14)	BMI = (Weight in kilograms / (Height in meters squared))	
BMICAT	BMI based on nurse measured data 1 = 25-<30 2 = 30+ 3 = <25	Derived from ANTHBMI	"<25" is coded as the reference category.
WAISTCM	Waist circumference measurement in cm (range: 55.9-165.1)	Ph3: L3	Anthropometric measurement at interview. In general, 2 measurements were taken. A third measure was taken if the first 2 differed by > 1 inch. If only 2 measurements were available, this variable is the average of the 2. If had third measure, take average of the closest 2. Converted from inches to cm using this formula: cm = inch x 2.54
WAISTMED	Waist circumference (cm) - median 1 = >=87 2 = <87	Derived from WAISTCM	Cut points obtained from median of overall controls from CBCS 1 & 2. "<87 cm" is coded as the reference category.
WAISTTER	Waist circumference (cm)- tertiles 1 = 80-<95 2 = >=95 3 = <80	Derived from WAISTCM	Cut points obtained from tertiles of overall controls from CBCS 1 & 2. "<80 cm" is coded as the reference category.

HIPCM	Hip circumference measurement in cm (range: 73.7-179.1)	Ph3: L3	Anthropometric measurement at interview. In general, 2 measurements were taken. A third measure was taken if the first 2 differed by > 1 inch. If only 2 measurements were available, this variable is the average of the 2. If had third measure, take average of the closest 2. Converted from inches to cm using this formula: cm = inch x 2.54
HIPMED	Hip circumference (cm) - median 1 = >=107 2 = <107	Derived from HIPCM	Cut points obtained from median of overall controls from CBCS 1 & 2. "<107 cm" is coded as the reference category.
HIPTER	Hip circumference (cm) - tertiles 1 = 102-<113 2 = >=113 3 = <102	Derived from HIPCM	Cut points obtained from tertiles of overall controls from CBCS 1 & 2. "<102 cm" is coded as the reference category.
WHRATIO	Waist-hip ratio based on nurse measured anthropometric data (range: 0.59-1.35)	WAISTCM/HIPCM	
WHIPMED	Waist hip ratio – median 1 = >=0.8 2 = <0.8	Derived from WHRATIO	Cut points obtained from median of overall controls from CBCS 1 & 2. "<0.8" is coded as the reference category.
WHIPTER	Waist hip ratio – tertiles 1 = 0.77-<0.84 2 = >=0.84 3 = <0.77	Derived from WHRATIO	Cut points obtained from tertiles of overall controls from CBCS 1 & 2. "<0.77" is coded as the reference category.

WAHEIGHT	Waist height ratio (range: 0.34-1.02)	WAISTCM/HEIGHT	
WHIGHMED	Waist height ratio – median 1 = >=0.54 2 = <0.54	Derived from WAHEIGHT	Cut points obtained from median of overall controls from CBCS 1 & 2. "<0.54" is coded as the reference category.
WHIGHTER	Waist height ratio - tertiles 1 = 0.49-<0.58 2 = >=0.58 3 = <0.49	Derived from WAHEIGHT	Cut points obtained from tertiles of overall controls from CBCS 1 & 2. "<0.49" is coded as the reference category.
WEIGHT_5G	Weight at 5th grade (10 years old) compared to other girls 1 = Thinner 2 = About the same 3 = Heavier	Ph3: E32	
HEIGHT_5G	Height at 5th grade (10 years old) compared to other girls 1 = Shorter 2 = About the same 3 = Taller	Ph3: E33	
WT_5G	Weight at 5th grade (10 years) compared to other girls 1 = Heavier 2 = Thinner/About the same	Recoded from WEIGHT_5G	
HT_5G	Height at 5th grade (10 years) compared to other girls 1 = Taller 2 = Shorter/About the same	Recoded from HEIGHT_5G	

The DASH (Disabilities of the Arm, Shoulder and Hand) Outcome Measure

Variable Name	Description	Data Source (Survey: Question Number)	Comments
DASH	DASH disability/symptom score (range: 0 – 94.17)	Ph3: J53-J82	From DASH scoring instructions: At least 27 of the 30 items must be completed for a score to be calculated. The assigned values for all completed responses are simply summed and averaged, producing a score out of five. This value is then transformed to a score out of 100 by subtracting one and multiplying by 25. This transformation is done to make the score easier to compare to other measures scaled on a 0-100 scale. **A higher score indicates greater disability.** DASH disability/symptom score= $\left[\left(\frac{\text{sum of n responses}}{n}\right) - 1 \right] \times 25$ where n is equal to number of completed responses.
DASH_WORK	DASH work score	Ph3: J83-J86	All four of the questions must be answered in order to calculate the score. Simply add up the assigned values for each response and divide by four (number of items); subtract one and multiply by 25 to get a score out of 100. DASH work score= $\left[\left(\frac{\text{sum of 4 responses}}{4}\right) - 1 \right] \times 25$

QUICKDASH	QuickDASH disability/symptom score	Ph3: J53, J59, J62, J66, J68, J70, J74, J75, J76, J78, and J81	This is a shortened version of DASH based on 11 items. This can be used instead of the DASH score. This variable should be used when comparing baseline to FU7 because QuickDASH is used in the FU7 mail-in survey. A QuickDASH score may not be calculated if there is greater than 1 missing item. QuickDASH disability/symptom score= [(sum of n responses) - 1] x 25 n where n is equal to number of completed responses.

FACIT (Functional Assessment of Chronic Illiness Therapy)/FACT (Functional Assessment of Cancer Therapy) Scales

FACT-B+4 Scoring Guidelines (Version 4)

Instructions:
1. Record answers in the “item response" column. If missing, mark with an X
2. Perform reversals as indicated and sum individual items to obtain a score.
3. Multiply the sum of the item scores by the number of items in the subscale, then divide by the number of items answered. This produces the subscale score.
4. Add subscale scores to derive total scores (TOI, FACT-G & FACT-B).
5. **The higher the score, the better the QOL.**

When there is missing data, prorating subscale is acceptable as long as more than 50% of the items were answered.

Variable Name	**Description**	**Data Source** **(Survey: Question Number)**	**Comments**
FACT_PWB	FACT-B+4 physical well-being subscale (range: 0 – 28)	Ph3: J8-J14	Item Number / Reverse item? / Item response / Item Score J8 4 - = J9 4 - = J10 4 - = J11 4 - = J12 4 - = J13 4 - = J14 4 - = Sum individual item scores: Multiply by 7: Divided by number of items answered: = PWB subscale

FACT_SWB	FACT-B+4 social/family well-being subscale (range: 0 – 28)	Ph3: J15-J21	Item Number / Reverse item? / Item response / Item Score J15 0 + = J16 0 + = J17 0 + = J18 0 + = J19 0 + = J20 0 + = J21 0 + = Sum individual item scores: Multiply by 7: Divided by number of items answered: = SWB subscale
FACT_EWB	FACT-B+4 emotional well-being subscale (range: 0 – 24)	Ph3: J22-J27	Item Number / Reverse item? / Item response / Item Score J22 4 - = J23 0 + = J24 4 - = J25 4 - = J26 4 - = J27 4 - = Sum individual item scores: Multiply by 6: Divided by number of items answered: = EWB subscale
FACT_FWB	FACT-B+4 functional well-being subscale (range: 0 – 28)	Ph3: J28-J34	Item Number / Reverse item? / Item response / Item Score J28 0 + = J29 0 + = J30 0 + = J31 0 + = J32 0 + = J33 0 + = J34 0 + = Sum individual item scores: Multiply by 7: Divided by number of items answered: = FWB subscale

FACT_BCS	FACT-B+4 breast cancer subscale (range: 0 – 36)	Ph3: J35-J43	Item Number / Reverse item? / Item response / Item Score J35 4 - = J36 4 - = J37 4 - = J38 0 + = J39 4 - = J40 4 - = J41 4 - = J42 4 - = J43 0 + = Sum individual item scores: Multiply by 9: Divided by number of items answered: = BC subscale
FACT_ARM	FACT-B+4 arm subscale (range: 0 – 20)	Ph3: J37, J46-J49	Item Number / Reverse item? / Item response / Item Score J37 4 - = J46 4 - = J47 4 - = J48 4 - = J49 4 - = Sum individual item scores: Multiply by 5: Divided by number of items answered: = ARM subscale
FACT_B_TOI	FACT-B Trial Outcome Index (TOI) (range: 0-92)	Sum of FACT_PWB, FACT_FWB, and FACT_BCS	FACT_B_TOI = FACT_PWB + FACT_FWB + FACT_BCS A total score should only be calculated if ALL of the component subscales have valid scores.

FACT_G_TOTAL	FACT-G Total score (range: 0-108)	Sum of FACT_PWB, FACT_SWB, FACT_EWB, and FACT_FWB	The FACT scale is considered to be an acceptable indicator of patient quality of life as long as **overall item response rate** is greater than 80% (e.g., at least 22 of 27 FACT-G items completed). This is not to be confused with individual subscale item response rate, which allows a subscale score to be prorated for missing items if greater than 50% of items are answered. In addition, a total score should only be calculated if ALL of the component subscales have valid scores. FACT_G_TOTAL = FACT_PWB + FACT_SWB + FACT_EWB + FACT_FWB
FACT_B_TOTAL	FACT-B Total score (range: 0-144)	Sum of FACT_PWB, FACT_SWB, FACT_EWB, FACT_FWB, and FACT_BCS	This scale is calculated if the **overall item response rate** is greater than 80% (at least 29 of 36 FACT-B items completed). In addition, a total score should only be calculated if ALL of the component subscales have valid scores. FACT_B_TOTAL = FACT_PWB + FACT_SWB + FACT_EWB + FACT_FWB + FACT_BCS

FACIT-Sp Scoring Guidelines (Version 4)

Instructions:* 1. Record answers in "item response" column. If missing, mark with an X
2. Perform reversals as indicated, and sum individual items to obtain a score.
3. Multiply the sum of the item scores by the number of items in the subscale, then divide by the number of items answered. This produces the subscale score.
4. Add subscale scores to derive total scores (FACIT-Sp12, FACIT-Sp total).
5. **The higher the score, the better the QOL/spiritual well-being.**

When there are missing data, prorating subscale is acceptable as long as more than 50% of the items were answered.

Variable Name	**Description**	**Data Source** **(Survey: Question Number)**	**Comments**
FACIT_MEANING_PEACE	FACIT-Sp Meaning/Peace subscale (range: 0 – 32)	Ph3: J87-J94	Item Number / Reverse item? / Item response / Item Score J87 0 + = J88 0 + = J89 0 + = J90 4 - = J91 0 + = J92 0 + = J93 0 + = J94 4 - = Sum individual item scores: Multiply by 8: Divided by number of items answered: = Meaning/Peace subscale
FACIT_MEANING	FACIT-Sp Meaning subscale (range: 0 – 16)	Ph3: J88, J89, J91, J94	Item Number / Reverse item? / Item response / Item Score J88 0 + = J89 0 + = J91 0 + = J94 4 - = Sum individual item scores: Multiply by 4: Divided by number of items answered: = Meaning subscale

FACIT_PEACE	FACIT-Sp Peace subscale (range: 0 – 16)	Ph3: J87, J90, J92, J93	Item Number / Reverse item? / Item response / Item Score J87 0 + = J90 4 - = J92 0 + = J93 0 + = Sum individual item scores: Multiply by 4: Divided by number of items answered: = Peace subscale
FACIT_FAITH	FACIT-Sp Faith subscale (range: 0 – 16)	Ph3: J95-J98	Item Number / Reverse item? / Item response / Item Score J95 0 + = J96 0 + = J97 0 + = J98 0 + = Sum individual item scores: Multiply by 4: Divided by number of items answered: = Faith subscale
FACIT_SP12	FACIT-Sp12 total subscale (range: 0-48)	Sum of FACIT_FAITH, FACIT_MEANING_PEACE	FACIT_Sp12 total = FACIT_MEANING_PEACE + FACIT_FAITH A total score should only be calculated if ALL of the component subscales have valid scores.
FACIT_SP_TOTAL	FACIT-Sp total score (range: 0-156)	Sum of FACT_PWB, FACT_SWB, FACT_EWB, FACT_FWB, and FACIT_SP12	This scale is calculated if the **overall item response rate** is greater than 80% (at least 32 of 39 FACIT/FACT-B items completed). In addition, a total score should only be calculated if ALL of the component subscales have valid scores. FACIT_SP_TOTAL= FACT_PWB + FACT_SWB + FACT_EWB + FACT_FWB + FACIT_SP12

Tumor Characteristics
(From Medical Record Abstract)

Note: Treatment data is available from the medical record abstract datasets.

Variable	Description	Comment
AJCC_GRP	AJCC Stage 1 = Stage I 1A = Stage IA 1B = Stage IB 2A = Stage IIA 2B = Stage IIB 3A = Stage IIIA 3B = Stage IIIB 4 = Stage IV 88 = Not applicable 99 = Unknown	Obtained from P3MA (ERS) file. 1 = Stage I (diagnosed before 1/1/2010) 1A = Stage IA (diagnosed 2010 and beyond) 1B = Stage IB (diagnosed 2010 and beyond)
STAGE	AJCC Stage 1 = Stage I 2 = Stage II 3 = Stage III 4 = Stage IV	Recoded from AJCC_GRP 1 = 1, 1A, 1B 2 = 2A, 2B 3 = 3A, 3B 4 = 4
SIZE	Tumor size (mm) 998 =Inflammatory carcinoma; diffuse, widespread, ¾ or more of breast 999 = unknown	Can also record size of inflammatory carcinoma if available.
ESTSIZE	Tumor size 1 = ≤ 2 cm 2 = >2 – 5 cm 3 = >5 cm	Recoded from SIZE. SIZE=998 classified as “>5 cm” per Melissa.

NODES_MALIG	Number of nodes positive for malignancy 00 = All nodes examined negative 01-89 = 1-89 nodes are positive 90 = 90 or more nodes are positive 95 = Positive aspiration of lymph nodes performed 97 = Positive nodes documented, number not specified 98 = No nodes examined 99 = Unknown	Recoded from ND_POS and N_STAGE. This variable is almost identical with ND_POS except the value 98 for some cases. If N_STAGE=0 (no lymph node mets) and ND_POS=98 (nodes not examined), assume all nodes negative (NODES_MALIG=0). This is consistent with the definition in CBCS 1 & 2.
NODESTAT	Node status 1 = Positive 2 = Negative	Recoded from ND_POS and N_STAGE. Positive is defined as one of the followings: 1) Number of nodes positive for malignancy >0 2) Staging - Lymph node metastasis If a case has multiple tumors, count as positive if any tumor is node positive. Note: For some clinically staged cases, NODESTAT is positive but ND_POS=0. These cases were considered node positive at time of diagnosis based on imaging studies, or palpation by managing physician. At the time of diagnosis, no nodes were taken. However, after the chemotherapy subsequent surgery in which nodes are taken that may not show any positive ones, which is when 0 is entered under ND_POS.
GRADE	Tumor grade 1 = Well differentiated 2 = Moderately differentiated 3 = Poorly differentiated 4 = Undifferentiated/Anaplastic differentiated 9 = not determined	Not available in CBCS 1 & 2. This is different from the CGRADE (combined grade) variable from the Centralized Pathology Review. The CGRADE variable is the preferred one to use in analysis.

(From Pathology Report Abstract)

Note: Onco_Dx_Assay data is available (for 652 subjects) from the pathology abstract dataset.

Variable	Description	Comment
ERSTAT	ER status 1 = Positive 2 = Negative 3 = Weak Positive / Borderline	Recoded from ER_STS (from pathology abstract file). If percent staining is available, cut point for positivity: 0 = negative 1-10 = weak positive/borderline >10 = positive If percent staining is not available, obtain ER status indicated in record. **Note**: cut point different from the Centralized IHC ER variable.
ER	ER status 1 = Positive 2 = Negative	Recoded from ERSTAT, borderline counted as missing.

PRSTAT	PR status 1 = Positive 2 = Negative 3 = Weak Positive / Borderline	Recoded from PR_STS (from pathology abstract file). If percent staining is available, cut point for positivity: 0 = negative 1-10 = weak positive/borderline >10 = positive If percent staining is not available, obtain PR status indicated in record. **Note**: cut point different from the Centralized IHC PR variable.
PR	PR status 1 = Positive 2 = Negative	Recoded from PRSTAT, borderline counted as missing.
PATH_HER2	HER2 status from IHC/FISH 1 = Positive 2 = Negative 3 = Borderline	Derived from IHC and/or FISH assay from the pathology report. There is another Centralized IHC Her2 variable. User will need to decide which version to use when combining data with CBCS 1 & 2.

Centralized Pathology Review (CPR)

Slides cut from tumor blocks or obtained from hospitals. Histopathologic evaluation done by CBCS study pathologist.

Variable	Description	Comment
HGRADE	Histologic grade 1 = Well diff./good tubule formation 2 = Mod. diff./mod. tubule formation 3 = Poorly diff./scant or no tubule formation 9 = Unknown	From C10 of CPR.
NGRADE	Nuclear grade 1 = Slight pleomorphism 2 = Moderate pleomorphism 3 = Marked pleomorphism 9 = Unknown	From C11 of CPR.
MITOTIC	Mitotic grade 1 = Nottingham MG1 (1) 2 = Nottingham MG2 (2) 3 = Nottingham MG3 (3) 9 = Unknown	From C12 of CPR. The category label is different from Phase 1. However, the 2 versions are equivalent per Dr. Geradts.
CGRADE	Combined grade 1 = Grade I (3-5) 2 = Grade II (6-7) 3 = Grade III (8-9) 9 = Unknown	From C13 of CPR.
TUMOR_NECROSIS	Tumor necrosis 1 = None 2 = Focal 3 = Marked 9 = Unknown	From C14 of CPR.

LYMPH_INFILTRATE	(C15) Lymphocytic infiltrate 1 = None/minimal 2 = Mild 3 = Dense 9 = Unknown	From C15 of CPR.
HISTCAT	Histologic groups 1 = Ductal NOS 2 = Mixed ductal/non-lobular (Except ductal/metaplastic) 3 = Medullary carcinoma 4 = Apocrine carcinoma 5 = Tubular carcinoma 6 = Mucinous carcinoma 7 = Papillary carcinoma 8 = Cribriform carcinoma 9 = Metaplastic carcinoma 10 = Anaplastic carcinoma 11 = Undifferentiated high grade 12 = Lobular carcinomas 13 = Mixed ductal and lobular 14 = Neuroendocrine 15 = DCIS w/ focal invasion 17 = Micropapillary carcinoma 18 = Others 20 = Mixed ductal/metaplastic 21 = Mixed lobular/non-ductal 22 = Pleomorphic lobular carcinoma 23 = Mixed non-ductal/non-lobular 99 = Unknown	This was initially defined by Dr. Bob Millikan so that the same variable can be used in all 3 phases. Categories added and edited by Dr. Geradts in 2016. Categories 17, 18, 20, 21, 22, and 23 are new to CBCS3. Categories 18 & 19 are combined in 1 category (18).

HISGROUP	Histologic groups 1 = Group A 2 = Group B 3 = Group C 4 = Group D 5 = Group E 6 = Group F 99=Unknown	Defined by Dr. Geradts. Group A: HISTCAT= 1, 2, 3, 4, 14, 15, 17 Group B: HISTCAT= 5, 6, 7, 8 Group C: HISTCAT= 9, 10, 11, 20, 22 Group D: HISTCAT= 12 Group E: HISTCAT= 13, 21 Group F: Other/Non-classified For HISTCAT=18 and 23, the histologic group assignment was done on a case by case basis by Dr. Geradts.

Centralized IHC Biomarkers data from UNC Translational Pathology Laboratory (TPL)
Data is available for N=2508 subjects

CENTRAL_ER	IHC-based ER Status 1 = Positive 2 = Negative	1 = weighted percent positive ≥10% 2 = weighted percent positive <10% Note: cut point different from pathology ER_STS and ERSTAT.
WEIGHTED_ PERCENT_POSITIVE_ER	ER – percent positive	
CENTRAL_PR	IHC-based PR Status 1 = Positive 2 = Negative	1 = weighted percent positive ≥10% 2 = weighted percent positive <10% Note: cut point different from pathology PR_STS and PRSTAT.
WEIGHTED_ PERCENT_POSITIVE_PR	PR – percent positive	
CENTRAL_HER2	IHC-based HER2 Status 1 = Positive 2 = Negative	
CENTRAL_P53	IHC-based P53 Status 1 = Positive 2 = Negative	1 = weighted percent positive ≥10% 2 = weighted percent positive <10%
WEIGHTED_ PERCENT_POSITIVE_P53	P53 – percent positive	
CENTRAL_EGFR	IHC-based EGFR Status 1 = Positive 2 = Negative	Positive: Any percent positive >=1%

CENTRAL_CK56	IHC-based CK5/6 Status 1 = Positive 2 = Negative	Positive: Any percent positive >=1%
CENTRAL_Ki67	IHC-based Ki67 Status 1 = Positive 2 = Negative	1 = weighted percent positive ≥7% 2 = weighted percent positive <7%
WEIGHTED_ PERCENT_POSITIVE_Ki67	Ki67 – percent positive	

CBCS3 IHC-based subtyping definitions (from Emma Allott)

Luminal A[a] = (weighted_percent_positive_er ≥10% or weighted_percent_positive_pr ≥10%) and weighted_percent_positive_ki67 <7%

Luminal B[a] = (weighted_percent_positive_er ≥10% or weighted_percent_positive_pr ≥10%) and weighted_percent_positive_ki67 ≥7%

ER-/HER2+ = weighted_percent_positive_er <10% and central_her2_status==3

Basal-like = (weighted_percent_positive_er <10% and weighted_percent_positive_pr <10% and central_her2_status==0) and (anypos_egfr1==1 or anypos_ck561==1)

---

[a]if Ki67 is missing, substitute CGRADE as follows:
Luminal A* = (weighted_percent_positive_er ≥10% or weighted_percent_positive_pr ≥10%) and CGRADE ≤2
Luminal B* = (weighted_percent_positive_er ≥10% or weighted_percent_positive_pr ≥10%) and CGRADE ==3

Variable Name	Description	Comments
IHC_SUBTYPE	IHC-based subtype (text) **LumA** **LumB** **ER-/HER2+** **Basal**	Emma Allott's definition. CBCS3 definition different from CBCS 1 & 2.
CBCS3_SUBTYPE	IHC-based subtype 1 = Basal 2 = LumA 3 = LumB 4 = ER-/HER2+	Emma Allott's definition – recoded from IHC_SUBTYPE. Numeric codes compatible with CBCS1 &2.
CBCS12_SUBTYPE	IHC-based subtype 1 = Basal 2 = LumA 3 = LumB 4 = ER-/HER2+ 5 = Unclassified	Using CBCS 1 & 2 definition **Basal-like:** ER- & PR- & HER2- & (EGFR+ or CK 5/6+) **Luminal A**: HER2- & (ER+ or PR+) **Luminal B**: HER2+ & (ER+ or PR+) **HER2+/ER-**: HER2+ & ER- & PR- **Unclassified**: Negative for all 5 markers Note: some CBCS3 cases undefined using CBCS 1 & 2 definition.

Latent Class Variables

There are 2 versions of latent class variables. One created by Marc Emerson and the other by Matthew Dunn.

Latent class variables from Marc Emerson

E_BESTSES	SES Latent classes 0=High SES/low comorbidity 1= Low SES/high comorbidity
E_BESTACC1	Access to Care 0=Less barriers to care 1=more barriers to care

Latent class variables from Matthew Dunn

Healthcare Access	
D_barriers2class	A latent class variable based on insurance, rural/urban status, self-reported financial and transportation barriers, and job loss
1	Fewer barriers
2	More barriers
SES	
D_Ses3classfinal	A latent class variable based on income, education, US/foreign born status, job type, and marital status
1	High SES
2	Lower SES, highly educated
3	Low SES