Data Dictionary for CBCS Phase 1 & Phase 2 Analysis Dataset

Table of Contents

Variable Name	Description	Data Source (Ph1 = CBCS 1 Survey; Ph2 = CBCS 2 Survey)	Comments
STUDYID	Study ID		
STUDY	Phase of study 0 = Phase 1 (invasive) 1 = Phase 2 (invasive) 2 = Phase 2 (CIS)		
CASE	Case/control status 0 = control 1 = case		
RACE	Race, used in sampling 1 = Non-African American 2 = African American		
AGESEL	Age at selection/diagnosis		Age at diagnosis for cases and selection into study for controls.
AGEINT	Age at interview		
AGEGR	5-year age groups, used in sampling 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	These are the age groups used for frequency matching of controls in Phase 1 & Phase 2 invasive. However, for CIS, the age groups are: 20-34, 35-44, 45-54, 55-64, and 65-74.

FRACT	Sampling probability based on age group, case/control status, and race.		Different probabilities in Phase 1, Phase 2 invasive and CIS.
WT	Sampling weights	(1/FRACT) – inverse of the sampling probability	Use the WT variable if one is interested in calculating a prevalence estimate among women in North Carolina, based upon the CBCS controls. If one is interested in calculating statistics such as chi-square on the weighted prevalence estimate, use the WT and STRATA in SUDAAN or SAS Proc SurveyFreq to generate the correct weighted estimates and variances.
STRATA	Sampling strata		This is based on the study phase, case/control status, race, and age group. Subjects in each stratum have the same sampling probabilities. For invasive cases, age is categorized as <50 and 50+; for invasive controls, age is categorized by the 5-year age intervals. Examples: 1 = Ph1, <50 NonAA cases 7 = Ph1, 50+ NonAA cases 23 = Ph1, 20-24 NonAA controls 44 = Ph1, 70-74 AA controls 51 = Ph2, 50+ AA cases 108 = CIS, 65-74 AA controls
CBCSOFF	CBCS offset term		Define as the natural log of the ratio of the sampling probability for a case in a specific age-race-stratum to the sampling probability for a control from the same age-race-stratum. For case/control analysis, this needs to be included it in the OFFSET option of SAS Proc Logistics to account for the sampling design. However, it is not needed for case only analysis.

Demographics

Variable Name	Description	Data Source (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	Ph1: H1 Ph2: I1	
SELF_RACE	Self-reported race 1 = White 2 = Black/African American 3 = American Indian, Eskimo 4 = Asian or Pacific Islander 5 = Other	Ph1: H2 Ph2: I2	
OTHER_RACE	Other race, specify 2 = Multi-racial 3 = Hispanic/Latino 99= Unspecified	Ph1: H2 Ph2: I2	Only available for other race (SELF_RACE=5).
ETHNICITY	Are you Hispanic? 1 = Hispanic 2 = Not Hispanic	Ph1: H3 Ph2: I3	

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	Ph1: H4 Ph2: I4	
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	Ph1: H11 Ph2: I14	
MONEY	Family income 1 = 15-30K 2 = 30-50K 3 = >50K 4 = <15K	Recoded from INCOME	"<15K" is coded as the reference category.

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 0 = No 1 = Yes	Ph1: B4, B7 Ph2: B4, B7	
BC_MOM	Breast cancer in mother 0 = No 1 = Yes	Ph1: B4 Ph2: B4	
MOMAGEBC	Age mother diagnosed with breast cancer	Ph1: B4 Ph2: B4	Missing for those with no maternal history of breast cancer.
BC_DAD	Breast cancer in father 0 = No 1 = Yes	Ph1: B4 Ph2: B4	
DADAGEBC	Age father diagnosed with breast cancer	Ph1: B4 Ph2: B4	Missing for those with no paternal history of breast cancer.
BCSIBYN	Breast cancer in any sibling 0 = No 1 = Yes 98= No siblings	Ph1: B7 Ph2: B7	
BC_SIB	Number of siblings with breast cancer 98 = No siblings	Ph1: B7 Ph2: B7	
MNAGSBC	Minimum age at which sibling diagnosed with breast cancer		Missing for those with no sibling history of breast cancer.

BCSISYN	Breast cancer in any sisters 0 = No 1 = Yes 98= No sisters	Ph1: B7 Ph2: B7	
BC_SIS	Number of sisters with breast cancer 98 = No sisters	Ph1: B7 Ph2: B7	
MNAGSISBC	Minimum age at which sister diagnosed with breast cancer	Ph1: B7 Ph2: B7	Missing for those with no sister history of breast cancer.
BCDAUGHYN	Breast cancer in any daughters 0 = No 1 = Yes 98= No daughters	Ph1: B9 Ph2: B9	
BC_DAUGH	Number of daughter with breast cancer 98 = No daughter	Ph1: B9 Ph2: B9	
MNAGDAUBC	Minimum age at which daughter diagnosed with breast cancer	Ph1: B9 Ph2: B9	Missing for those with no daughter history of breast cancer.

FFAMHXOC	First-degree family history of ovarian cancer – mother or sister 0 = No 1 = Yes	Ph1: B4, B7 Ph2: B4, B7	
OC_MOM	Ovarian cancer in mother 0 = No 1 = Yes	Ph1: B4 Ph2: B4	
MOMAGEOC	Age mother diagnosed with ovarian cancer	Ph1: B4 Ph2: B4	Missing for those with no maternal history of ovarian cancer.
OCSIBYN	Ovarian cancer in any sister 0 = No 1 = Yes 98= No sisters	Ph1: B7 Ph2: B7	
OC_SIB	Number of sisters with ovarian cancer 98 = No sisters	Ph1: B7 Ph2: B7	
MNAGSOC	Minimum age at which sister diagnosed with ovarian cancer		Missing for those with no sister history of ovarian cancer.

Menstrual History
(Exclude exposure after age of selection/diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
AGEMENA	Age at menarche (range: 8-21)	Ph1: C1 Ph2: C1	
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 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	Ph1: C5-C11 Ph2: C4-C10	If subject experienced menopause after date of selection/diagnosis, she would be classified as premenopausal for this variable.

MENODATE	Date of menopause	Ph1: C5-C11 Ph2: C4-C10	This variable goes with the variable MENO. Missing for premenopausal (MENO=1) women.
AGEMENO	Age at menopause	Ph1: C5-C11 Ph2: 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 menopausal	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 Ph1: C6 Ph2: 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 Ph1: C6 Ph2: C5	Defined for those with MENO codes 3-14. This is not the same as AGEMENO.

Pregnancy and Lactation
(Exclude exposure after age of selection/diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
PREGNUM	Number of pregnancies	Ph1: C14 Ph2: C11	Exclude pregnancies after age of selection/diagnosis.
FTPEVER	Ever had full-term pregnancy 0 = No 1 = Yes	Ph1: C15 Ph2: 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: 0-16)	Ph1: C15 Ph2: 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)	Ph1: C15 Ph2: C12	Missing for nulliparous.
AGELFTP	Age last full-term pregnancy (range: 14-46)	Ph1: C15 Ph2: C12	Missing for nulliparous.
DATELFTP	Date of last full-term pregnancy	Ph1: C15 Ph2: 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	Ph1: C15 Ph2: C12	
PREG_DX	Pregnant at diagnosis/selection or diagnosed/selected within 2 years after pregnancy 0 = No 1 = Yes	Derived from AGESEL and Ph1: C15 Ph2: C12	
PREGCUR	Pregnant at time of diagnosis/selection 0 = No 1 = Yes	Derived from AGESEL and Ph1: C15 Ph2: 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	Ph1: C15 Ph2: C12	Only include pregnancies that resulted in live birth.
NUMLIVEB	Number of live birth pregnancies (range: 0-16)	Ph1: C15 Ph2: C12	
AGEFLIVE	Age first live birth (range: 11-44)	Ph1: C15 Ph2: C12	Missing for women who never had live births.
AGELLIVE	Age last live birth (range: 14-46)	Ph1: C15 Ph2: C12	Missing for women who never had live births.

AGELLIVE	Age last live birth (range: 14-46)	Ph1: C15 Ph2: C12	Missing for women who never had live births.
IABEVER	Ever had induced abortion 0 = No 1 = Yes	Ph1: C15 Ph2: C12	
NUMBIAB	Number of induced abortions (range:0-7)	Ph1: C15 Ph2: C12	
AGEFIAB	Age first induced abortion (range: 9-45)	Ph1: C15 Ph2: C12	Missing for women who never had induced abortions.
SABEVER	Ever had spontaneous abortion 0 = No 1 = Yes	Ph1: C15 Ph2: C12	
NUMBSAB	Number of spontaneous abortions (range: 0-13)	Ph1: C15 Ph2: C12	
AGEFSAB	Age first spontaneous abortion (range: 10-51)	Ph1: C15 Ph2: C12	Missing for women who never had induced abortions.
ECTEVER	Ever had ectopic/tubal pregnancy 0 = No 1 = Yes	Ph1: C15 Ph2: C12	
NUMBECT	Number of ectopic/tubal pregnancies (range: 0-3)	Ph1: C15 Ph2: C12	
AGEFECT	Age first ectopic/tubal pregnancy (range: 17-42)	Ph1: C15 Ph2: C12	Missing for women who never had induced ectopic/tubal pregnancy.

PRETERM	Ever had preterm birth 0 = No 1 = Yes	Ph1: C15 Ph2: C12	Preterm is defined as <9 month pregnancy duration for single, multiple live births, and stillbirth. Note: definition of term-pregnancy is different from PARITY and FTPEVER.
LACTEVER	Ever lactated 0 = No 1 = Yes	Ph1: C15 Ph2: C12	
SUMLACT	Lifetime duration lactation (months) (range: 0-111)	Ph1: C15 Ph2: C12	0 = never lactated or lactated < 2 weeks
LACTMON	Lifetime duration lactation 1 = >0-3 months 2 = 4+ months 3 = Never breast fed	Derived from SUMLACT and LACTEVER	
AGEFLACT	Age at first lactation (range: 11-44)	Ph1: C15 Ph2: C12	Missing for those who 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: 11-44)	Ph1: C15 Ph2: C12	Missing for those who 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: 0-12)	Ph1: C15 Ph2: C12	
LACTKIDS	Number of children breastfed 1 = 1 child 2 = 2+ children 3 = Never breast fed	Derived from PREGLAC and LACTEVER	
LAVGMON	Number of months breastfeeding per child (range: 0-48)	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)	Ph1: C15 Ph2: C12	0 = never lactated or lactated < 2 weeks
LACTSUPP	Number of pregnancies for which milk production was suppressed by medication (range: 0-9)	Ph1: C15 Ph2: C12	
SUPPRESS	Lactation suppressant use 1 = Ever 2 = Never	Derived from LACTSUPP	Note: never is coded as 2.
LACTUNAB	Number of pregnancies for which unable to lactate (range: 0-7)	Ph1: C15 Ph2: C12	

LACTCNOT	Number of pregnancies for which chose not to lactate (range: 0-15)	Ph1: C15 Ph2: C12	
LACTPROB	Number of pregnancies for which not lactated due to prior problems (range: 0-4)	Ph2: C12	Available in Phase 2 data 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 selection/diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
OCEVER	Ever use oral contraceptives 0 = Never 1 = Ever	Ph1: D1, D3, D4 Ph2: E1, E3, E4	Ever user is defined as 3+ months of OC use. Exclude OC use after age of selection/diagnosis
OCUSE	Use of oral contraceptives 0 = Never 1 = Current 2 = Former	Ph1: D1, D3, D4 Ph2: E1, E3, E4	
OCMONTHS	Number of months used OC	Derived from OCUSE, AGESEL and Ph1: D4 Ph2: E4	Include months of use for never user. Never user: range 0-2 Ever user: range 3-336
OC_REC	Years since last used OC 0 = current user	Derived from OCUSE, AGESEL and Ph1: D3 Ph2: E3	
OCAGE	Age first used OC	Derived from OCEVER and Ph1: D2 Ph2: E2	
OCLASTAGE	Age last used OC	Derived from OCEVER and Ph1: D3 Ph2: 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	

Hormone Replacement Therapy
(Exclude exposure after age of selection/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)	Ph1: D21a+D21c+D21e Ph2: E11a+E11c+E11e	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of selection/diagnosis.
ESTROMON	Number of months used estrogen (with/without progestin)	Ph1: D21a+D21c+D21e Ph2: E11a+E11c+E11e	Include months of use for never user. Never user: range 0-2 Ever user: range 3-444
AGEFEST	Age first ever used estrogen (with/without progestin) (range: 17-73)	Ph1: D21a+D21c+D21e Ph2: E11a+E11c+E11e	
AGELEST	Age last used estrogen (with/without progestin) (range: 17-74)	Ph1: D21a+D21c+D21e Ph2: E11a+E11c+E11e	
EST_REC	Years since last use estrogen (with/without progestin) (range: 0-44)	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)	Ph1: D21(a-b) + D21 (c-d) + D21 (e-f) Ph2: E11(a-b) + E11 (c-d) + E11 (e-f)	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of selection/diagnosis. Exclude the times when estrogen were used together with progestin.
ESTONLY_MON	Number of months used estrogen only (no progestin)	Ph1: D21(a-b) + D21 (c-d) + D21 (e-f) Ph2: E11(a-b) + E11 (c-d) + E11 (e-f)	Include months of use for never user. Never user: range 0-2 Ever user: range 3-444
AGEFEONLY	Age first ever used estrogen only (range: 17-73)	Ph1: D21(a-b) + D21 (c-d) + D21 (e-f) Ph2: E11(a-b) + E11 (c-d) + E11 (e-f)	
AGELEONLY	Age last used estrogen only (range: 17-74)	Ph1: D21(a-b) + D21 (c-d) + D21 (e-f) Ph2: E11(a-b) + E11 (c-d) + E11 (e-f)	
ESTONLY_REC	Years since last use estrogen only (range: 0-44)	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)	Ph1: D21b+D21d+D21f Ph2: E11b+E11d+E11f	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of selection/diagnosis.
EPMON	Number of months used estrogen+progestin	Ph1: D21b+D21d+D21f Ph2: E11b+E11d+E11f	Include months of use for never user. Never user: range 0-2 Ever user: range 3-300
AGEFEP	Age first ever used estrogen+progestin (range: 20-72)	Ph1: D21b+D21d+D21f Ph2: E11b+E11d+E11f	
AGELEP	Age last used estrogen+progestin (range: 25-74)	Ph1: D21b+D21d+D21f Ph2: E11b+E11d+E11f	
EP_REC	Years since last use estrogen+progestin (range: 0-44)	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)	Ph1: D21g Ph2: E11g	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of selection/diagnosis.
PROGMON	Number of months used progestin only	Ph1: D21g Ph2: E11g	Include months of use for never user. Never user: range 0-2 Ever user: range 3-192
AGEFPROG	Age first ever used progestin only (range: 16-70)	Ph1: D21g Ph2: E11g	
AGELPROG	Age last used progestin only (range: 16-71)	Ph1: D21g Ph2: E11g	

PROG_REC	Years since last use progestin only (range: 0-42)	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)	Ph1: D21a+D21c+D21e+ D21g Ph2: E11a+E11c+E11e+ E11g	Ever user is defined as 3+ months of hormone use. Exclude hormone use after age of selection/diagnosis.
ANYMON	Number of months used any hormone replacement therapy	Ph1: D21a+D21c+D21e+ D21g Ph2: E11a+E11c+E11e+ E11g	Include months of use for never user. Never user: range 0-2 Ever user: range 3-444
AGEFANY	Age first ever used any hormone replacement therapy (range: 16-73)	Ph1: D21a+D21c+D21e+ D21g Ph2: E11a+E11c+E11e+ E11g	
AGELANY	Age last used any hormone replacement therapy (range: 16-74)	Ph1: D21a+D21c+D21e+ D21g Ph2: E11a+E11c+E11e+ E11g	
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	Ph1: D21a+D21c+D21e Ph2: E11a+E11c+E11e	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		Defined for those ever used HRT.
POSTHRT	Postmenopausal HRT use 0 = No 1 = Yes	Derived from POSTMENO & ANYHRT	
HRTYRS	Postmenopausal HRT use durations 1 = <5 years 2 = 5-10 3 = >10 4 = never	Derived from POSTHRT & ANYMON	

NSAIDS Use
(Available for Phase 2 data only)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
NSDEVER	Ever/never use of NSAIDS 0 = never user 1 = ever user	Ph2: E14, E15, E16, E17	Never user includes subjects who took acetaminophen only, and subject who never used prescription or over the counter NSAIDS. Based on pre-diagnosis use.
NSAIDUSR	Type of NSAID user - based on length of pre-breast cancer diagnosis use 0 = nonuser 1 = occasional user (< 3 months or <= 7 days/month) 2 = short-term regular user (>=8 days/month for < 3 years) 3 = long-term regular user (>=8 days/month for >=3 years)	Ph2: E14, E15, E16, E17	Non-user includes subjects who took acetaminophen only, and subject who never used prescription or over the counter NSAIDS.
BEFDXTYPE	Type of NSAID use - based on pre-dx prescription vs. non-prescription use 0 = nonuser 1 = non-prescription use only 2 = prescription use only 3 = both non-prescription and prescription use	Ph2: E14, E15, E16, E17	Non-user includes subjects who took acetaminophen only, and subject who never used prescription or over the counter NSAIDS.
ASPEVER	Ever/never use of ASPIRIN 0 = never user 1 = ever user	Ph2: E14, E15, E16, E17	Based on pre-diagnosis use.

Radiation

Variable Name	Description	Data Source (Survey: Question Number)	Comments
HIGHRAD	History of high dose radiation to chest area 0 = No 1 = Yes	Ph1: E12, E13, E15, E16 Ph2: D22, D23, D30, D31	Exclude exposure after age of selection/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	Ph1: G2, G8 Ph2: G2	1990 Occupational Classification Codes: 84 (physicians), 95 (RN), 206 (radiologic technicians), 207 (LPN)
JOBSRAD2	Jobs with potential radiation exposure 0 = No 1 = Yes	Ph1: G2, G8 Ph2: 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)

Physical Activity

Variable Name	Description	Data Source (Survey: Question Number)	Comments
PHYSICAL	Physical activity 0 = No 1 = Yes	Ph1: F7 Ph2: F6	

Fruits & Vegetables Consumption

Variable Name	Description	Data Source (Survey: Question Number)	Comments
WINVEG_M	Weekly vegetables in winter (1/2 cup-sized servings) 0 = < 14 1 = 14+	Ph1: F10 Ph2: F7	Cut points obtained from median of overall controls.
SUMVEG_M	Weekly vegetables in summer (1/2 cup-sized servings) 0 = < 14 1 = 14+	Ph1: F11 Ph2: F8	Cut points obtained from median of overall controls.
WINFRU_M	Weekly fruits or fruit juices in winter (1/2 cup-sized servings) 0 = < 8 1 = 8+	Ph1: F12 Ph2: F9	Cut points obtained from median of overall controls.
SUMFRU_M	Weekly fruits or fruits juices in summer (1/2 cup-sized servings) 0 = < 12 1 = 12+	Ph1: F13 Ph2: F10	Cut points obtained from median of overall controls.

Alcohol use
(Exclude exposure after age of selection/diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
ALCOHOL	Ever used alcohol 0 = No 1 = Yes	Ph1: F16, F17 Ph2: F17, F18	Coded as "No" if first started drinking alcohol after age of selection/diagnosis.
ALCSTART	Age first started drinking alcohol (range: 4-70)	Ph1: F17 Ph2: F18	Missing for non-drinker.
ALCLT25	Ever alcohol use during ages less than 25 0 = No 1 = Yes	Ph1: F18 (Before age 25) Ph2: F19 (Before age 25)	
ALC2549	Ever alcohol use during ages 25-49 0 = No 1 = Yes	Ph1: F18 (Ages 25-49) Ph2: F19 (Ages 25-49)	Missing for those AGESEL less than 25.
ALC50PL	Ever alcohol use after age 50 0 = No 1 = Yes	Ph1: F18 (Since age 50) Ph2: F19 (Since age 50)	Missing for those AGESEL less than 50.

DRINKSWK25	Drinks per week before age 25	Ph1: F18 (Before age 25) Ph2: F19 (Before age 25)	Number of drinks per month divided by 4. Number of drinks per year divided by 52. Missing for unknown alcohol use.
DRINKSWK49	Drinks per week from ages 25 to 49	Ph1: F18 (Ages 25-49) Ph2: F19 (Ages 25-49)	Number of drinks per month divided by 4. Number of drinks per year divided by 52. Missing for unknown alcohol use and for AGESEL less than 25.
DRINKSWK50	Drinks per week since age 50	Ph1: F18 (Since age 50) Ph2: F19 (Since age 50)	Number of drinks per month divided by 4. Number of drinks per year divided by 52. Missing for unknown alcohol use and for AGESEL less than 50.

Smoking
(Exclude exposure after age of selection/diagnosis)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
EVERSMOK	Smoking status 0 = Never 1 = Ever	Ph1: F19 Ph2: F20	
SMOKERS	Smoking status 0 = Never 1 = Former 2 = Current	Ph1: F19, F21, F22 Ph2: F20, F22, F23	This version of the variable was used in analyses before 9/26/2014. If age of smoking cessation is after age of selection/diagnosis, the subject would be considered as current smoker. If age of smoking cessation is the same as age of diagnosis/selection, the subject would be classified as former smoker.
SMOKERS2	Smoking status 0 = Never 1 = Former 2 = Current	Ph1: F19, F21, F22 Ph2: F20, F22, F23	New definition as of 9/26/2014, this version will be used from now on. If age of smoking cessation $\geq$ age of selection/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.

DURATION	Smoking duration 1 = <=10 years 2 = 11 – 20 years 3 = > 20 years 4 = Never	Ph1: F19, F23 Ph2: F20, F24	
DOSE	Smoking dose (per day) 1 = < 1/2 pack 2 = 1/2 – 1 pack 3 = >1 pack 4 = Never	Ph1: F19, F24 Ph2: F20, F25	
PASSONLY	Never active smokers: exposure to ETS 1 = Exposed to ETS after age 18 2 = Unexposed to active smoke or ETS	Ph1: F19, F30 Ph2: F20, F29	
SMKGP	Smoking status 1 = Passive smoking only 2 = Former 3 = Current 4 = No active & no passive	Derived from PASSONLY and SMOKERS	
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 Ph1: F23 Ph2: F24	
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 Ph1: F24 Ph2: F25	

AGE_SMK	Age at initiation of smoking (range: 5-60)	Ph1: F20 Ph2: F21	
AGESTART	Age at initiation of active smoking 1 = Passive smoking only 2 = < 18 3 = 18+ 4 = No active & no passive	Derived from PASSONLY and Ph1: F20 Ph2: F21	
YRQUITSMK	Years since quitting smoking – former smokers (range: 0-50)	Derived from AGESEL, SMOKERS and Ph1: F22 Ph2: F23	This version of the variable was used in analyses before 9/26/2014. 0 = less than 1 year Missing for never and current smokers.
YRQUITSMK2	Years since quitting smoking – former smokers (range: 1-50)	Derived from AGESEL, SMOKERS2 and Ph1: F22 Ph2: F23	New definition as of 9/26/2014, this version will be used from now on. Missing for never and current smokers.
YRSTOP	Years since stopped active smoking 1 = Passive smoking only 2 = < 10 years 3 = 10+ years 4 = No active & no passive	Derived from PASSONLY and YRQUITSMK	This version of the variable was used in analyses before 9/26/2014. Missing for 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 ago (range: 13.10-62.23)	Ph1: E17, E18 Ph2: E20, E21	
BMI18	BMI based on self-report usual adult height and weight at age 18.	Ph1: E17, E21 (18 yo) Ph2: E20, E24 (18 yo)	
BMI35	BMI based on self-report usual adult height and weight at age 35.	Ph1: E17, E21 (35 yo) Ph2: E20, E24 (35 yo)	Missing for AGESEL<35.
HEIGHT	Nurse measured height in cm (range: 137-188)	Anthropometric measurement at interview	
HIGHMED	Height (cm) - median 1 = >=162.5 2 = <162.5	Derived from HEIGHT	Cut points obtained from median of overall controls. "<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. "<160 cm" is coded as the reference category.
WEIGHT	Nurse measured weight in kg (range: 35-165)	Anthropometric measurement at interview	
ANTHBMI	BMI based on nurse measured anthropometric data (range: 13.17-68.37)	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: 55.25-165)	Anthropometric measurement at interview	In general, 2 measurements were taken. A third measure was taken if the first 2 differed by > 1cm. If only 2 measurements were available, this variable is the average of the 2. If had third measure, take average of the closest 2.
WAISTMED	Waist circumference (cm) - median 1 = >=87 2 = <87	Derived from WAISTCM	Cut points obtained from median of overall controls. "<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. "<80 cm" is coded as the reference category.
HIPCM	Hip circumference measurement in cm (range: 68.05-170.05)	Anthropometric measurement at interview	In general, 2 measurements were taken. A third measure was taken if the first 2 differed by > 1cm. If only 2 measurements were available, this variable is the average of the 2. If had third measure, take average of the closest 2.
HIPMED	Hip circumference (cm) - median 1 = >=107 2 = <107	Derived from HIPCM	Cut points obtained from median of overall controls. "<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. "<102 cm" is coded as the reference category.

WHRATIO	Waist-hip ratio based on nurse measured anthropometric data (range: 0.59-1.34)	WAISTCM/HIPCM	
WHIPMED	Waist hip ratio – median 1 = >=0.8 2 = <0.8	Derived from WHRATIO	Cut points obtained from median of overall controls. "<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. "<0.77" is coded as the reference category.
WAHEIGHT	Waist height ratio (range: 0.33-1.02)	WAISTCM/HEIGHT	
WHIGHMED	Waist height ratio – median 1 = >=0.54 2 = <0.54	Derived from WAHEIGHT	Cut points obtained from median of overall controls. "<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. "<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	Ph1: E22 Ph2: E25	
HEIGHT_5G	Height at 5th grade (10 years old) compared to other girls 1 = Shorter 2 = About the same 3 = Taller	Ph1: E23 Ph2: E26	
WT_5G	Weight at 5th grade (10 years old) compared to other girls 1 = Heavier 2 = Thinner/About the same	Recoded from WEIGHT_5G	
HT_5G	Height at 5th grade (10 years old) compared to other girls 1 = Taller 2 = Shorter/About the same	Recoded from HEIGHT_5G	

Tumor Characteristics
(Available for cases only)

Variable Name	Description	Data Source (Survey: Question Number)	Comments
STAGE	AJCC/UICC Stage Grouping 1 = Stage I 2 = Stage II 3 = Stage III 4 = Stage IV	Medical record abstract	Available for Phase 1 & Phase 2 invasive cases only. For some analysis, CIS cases can be considered Stage 0.
ERSTAT	ER Status 1 = Positive 2 = Negative 3 = Borderline	Medical record abstract and IHC staining	For Phase 1 & Phase 2 invasive cases, 88% of data were from medical records, 12 % were from IHC staining. For CIS cases, data was from IHC staining.
PRSTAT	PR Status 1 = Positive 2 = Negative 3 = Borderline	Medical record abstract and IHC staining	For Phase 1 & Phase 2 invasive cases, 87% of data were from medical records, 13 % were from IHC staining. Data not available for CIS cases.
ER	ER Status 1 = Positive 2 = Negative	Medical record abstract and IHC staining	Borderline ER status is set to missing. This version of the ER status was used in most of the previous CBCS analyses.
PR	PR Status 1 = Positive 2 = Negative	Medical record abstract and IHC staining	Borderline PR status is set to missing. This version of the PR status was used in most of the previous CBCS analysis.

NODESTAT	Node status 1 = Positive 2 = Negative	Medical record abstract	Available for Phase 1 & Phase 2 invasive cases only. Positive is defined as one of the followings: 1) Number of nodes positive for malignancy >0 2) Lymph node metastasis
NODES_MALIG	Number of nodes positive for malignancy (range: 0-40)	Medical record abstract	Available for Phase 1 & Phase 2 invasive cases only.
ESTSIZE	Estimated tumor size 1 = ≤2 cm 2 = >2-5 cm 3 = >5 cm	Medical record abstract	Available for Phase 1 & Phase 2 invasive cases only.
TUMSIZE	Estimated tumor size 1 = >2 cm 2 = ≤2 cm	Recoded from ESTSIZE	Available for Phase 1 & Phase 2 invasive cases only. "≤ 2 cm" is coded as the reference category.
HGRADE	Histologic grade of invasive cancer 1=Well diff./good tubule formation 2=Mod. diff./mod. tubule formation 3=Poor diff./scant or no tubule 9=Unknown	Slides cut from tumor blocks or obtained from hospitals. Histopathologic evaluation done by CBCS study pathologist.	Available for Phase 1 invasive cases only.
HISGRADE	Histologic grade of invasive cancer 1=Poor diff./scant or no tubule 2=Mod/Well diff.	Recoded from HGRADE	Available for Phase 1 invasive cases only. "Mod/Well diff" is coded as the reference category.

NGRADE	NGRADE Nuclear grade 1=Slight pleomorphism 2=Moderate pleomorphism 3=Marked pleomorphism 9=Unknown	Slides cut from tumor blocks or obtained from hospitals. Histopathologic evaluation done by CBCS study pathologist.	Available for Phase 1 invasive cases only.
NUGRADE	Nuclear grade 1=Marked pleomorphism 2=Moderate/Slight	Recoded from NGRADE	Available for Phase 1 invasive cases only. "Moderate/Slight" is coded as the reference category.
MITOTIC	Mitotic rate 1= <=5 mitotses/10 HPF 2=6-10 3=>10 9=Unknown	Slides cut from tumor blocks or obtained from hospitals. Histopathologic evaluation done by CBCS study pathologist.	Available for Phase 1 invasive cases only.
MINDEX	Mitotic rate 1= >10 mitotses/10 HPF 2= <= 10	Recoded from MITOTIC	Available for Phase 1 invasive cases only. "<=10" is coded as the reference category.
CGRADE	Combined grade 1=Grade I 2=Grade II 3=Grade III 9=Unknown	Slides cut from tumor blocks or obtained from hospitals. Histopathologic evaluation done by CBCS study pathologist.	Available for Phase 1 invasive cases only.

HISTCAT	Histologic types 1 = Ductal NOS 2 = Mixed ductal/non-lobular 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 lobular and ductal 14 = Neuroendocrine 15 = DCIS w/ focal invasion 99 = Unknown	Slides cut from tumor blocks or obtained from hospitals. Histopathologic evaluation done by CBCS study pathologist.	Available for Phase 1 & Phase 2 invasive cases only.
HISGROUP	Histologic groups 1 = Group A (Ductal – less favorable outcome) 2 = Group B (Ductal – more favorable outcome) 3 = Group C (Less differentiated-less favorable outcome) 4 = Group D (Lobular) 5 = Group E (Mixed lobular and ductal)	Classify histologic types into 5 groups.	Available for Phase 1 & Phase 2 invasive cases only. Group A: HISTCAT= 1, 2, 3, 4, 14, 15 Group B: HISTCAT= 5, 6, 7, 8 Group C: HISTCAT= 9, 10, 11 Group D: HISTCAT= 12 Group E: HISTCAT= 13

The following variables are available for CIS cases only. The data are from the CIS Centralized Pathology Review. (Part B: Histopathologic Evaluation of H&E slides sent from referring pathologist)

CIS_GP	CIS subgroup 1 = DCIS 2 = LCIS 3 = DCIS w/ microinvasion	Determined by study pathologist or PI.	Available for CIS cases only.
COMEDO1	Comedo type DCIS – definition 1 (most strict) 1 = Comedo type DCIS 0 = NonComedo	Part B: questions B17, B21, B22, B24. B17=1 (Comedo necrosis) B21=3 (Marked nuclear pleomorphism) B22=3, 4 (Large or very Large nuclei) B24=3, 4 (Prominent nucleoi)	Available for CIS cases only. Comedo type DCIS is defined as ALL of the followings: Comedo necrosis Large or very large nuclei Marked nuclear pleomorphism Prominent nucleoli NonComedo = All else (including missing data)
COMEDO2	Comedo type DCIS – definition 2 (middle) 1 = Comedo type DCIS 0 = NonComedo	Part B: questions B17, B21, B22. B17=1 (Comedo necrosis) B21=3 (Marked nuclear pleomorphism) B22=3, 4 (Large or very Large nuclei)	Available for CIS cases only. Comedo type DCIS is defined as ALL of the followings: Comedo necrosis Large or very large nuclei Marked nuclear pleomorphism NonComedo = All else (including missing data)

COMEDO3	Comedo type DCIS – definition 3 (least strict) 1 = Comedo type DCIS 0 = NonComedo	Part B: questions B17, B21, B22, B24. B17=1 (Comedo necrosis) B21=3 (Marked nuclear pleomorphism) B22=3, 4 (Large or very Large nuclei) B24=3, 4 (Prominent nucleoi)	Available for CIS cases only. Comedo type DCIS is defined as: Comedo necrosis Any two of the following: Large or very large nuclei Marked nuclear pleomorphism Prominent nucleoli NonComedo = All else (including missing data)
CISTYPE1	CIS type – comedo definition 1 1 = DCIS comedo type 2 = DCIS noncomedo type 3 = DCIS w/ microinvasion 4 = LCIS	Derived from CIS_GP & COMEDO1	Available for CIS cases only.
CISTYPE2	CIS type – comedo definition 2 1 = DCIS comedo type 2 = DCIS noncomedo type 3 = DCIS w/ microinvasion 4 = LCIS	Derived from CIS_GP & COMEDO2	Available for CIS cases only.
CISTYPE3	CIS type – comedo definition 3 1 = DCIS comedo type 2 = DCIS noncomedo type 3 = DCIS w/ mcroinvasion 4 = LCIS	Derived from CIS_GP & COMEDO3	Available for CIS cases only.

Lab Results (from slides)
(Available for cases only)

Variable Name	Description	Data Source	Comments
IHC_HER2	IHC Her2 status 1 = Positive 2 = Negative	Lab work	For Phase 1 & Phase 2 invasive: HER2 positive is defined as: localization of stain = *membrane or membrane+cytoplasm* intensity of stain = *weak, moderate, or strong* Percent positive = *>=10* For CIS cases: Positive: scores 2-3 Negative: score 0-1
IHC_P53	IHC P53 status 1 = Positive 2 = Negative	Lab work	P53 positive is defined as: localization of stain = *nucleus or nuclear+cytoplasm* intensity of stain = *weak, moderate, or strong* Percent positive = *>=10*
CK5_6_2	CK 5/6 (Cytokeratin) 1 = Positive 2 = Negative	Lab work	Positive: scores 1, 2 Negative: score 0
HER1DEF1	HER1 aka EGFR 1 = Positive 2 = Negative	Lab work	Positive: scores 1, 2, 3 Negative: score 0

FAK	FAK expression 1 = Positive 2 = Negative	Lab work	Available for Phase 1 cases only.
ECAD	Ecadherin 1 = Positive 0 = Negative	Lab work	Available for Phase 1 cases only.
AMP_HER2	HER2 amplified 1 = Positive 0 = Negative	Lab work	Available for Phase 1 cases only.
PRAD_1	PRAD amplified score 1 = Positive 0 = Negative	Lab work	Available for Phase 1 cases only.
KI_67A	KI-67 1 = Low 2 = High	Lab work	Available for CIS cases only. Low: scores 0, 1 High: scores 2, 3, 4
KI_67B	KI-67 1 = Low 2 = Intermediate 3 = High	Lab work	Available for CIS cases only. Low: scores 0, 1 Intermediate: score 2 High: scores 3, 4

Tumor Subtype
(Available for cases with complete data in these markers: ER, PR, IHC HER2, CK5/6, and HER1.)

Variable Name	Description	Data Source	Comments
SUBTYPE	Tumor subtype 1 = Basal-like 2 = Luminal A 3 = Luminal B 4 = HER2+/ER- 5 = Unclassified	Derived from: 1) ER 2) PR 3) IHC_HER2 4) CK5_6_2 5) HER1_DEF1	Data for all 5 markers must be present. **Basal-like:** ER negative and PR negative and HER2 negative and (HER1 positive or CK 5/6 positive) **Luminal A**: HER2 negative and (ER positive or PR positive) **Luminal B**: HER2 positive and (ER positive or PR positive) **HER2+/ER-**: HER2 positive and ER negative and PR negative **Unclassified**: Negative for all 5 markers PR data is not available in CIS, and it is not used in the definition of CIS tumor subtype.

Variable Name	Description	Data Source	Comments
SUBTYPE14	Tumor subtype 1 = Basal-like 2 = HER2+/ER- 3 = LumA 2014 4 = LumB 2014 5 = Lum NOS 6 = Unclassified	Derived from: 1) ER 2) PR 3) IHC_HER2 4) CK5_6_2 5) HER1_DEF1	**Defined for Phase 1 & 2 Invasive only in 2014.** Data for all 5 markers must be present (includes borderline PR status as data instead of treating it as missing like previously; borderline ER status is treated as missing) **Basal-like (previous definition):** HER2- and ER- and PR- and (HER1+ or CK5/6 +) **HER2+/ER- (previous definition):** HER2+ and ER- and PR- **Luminal A 2014**: HER2- and (ER+ or ER-) and PR+ with >20% cell positive **Luminal B 2014:** Exclude cases with negative in all 5 markers. 1. HER2- and (ER+ or ER-) and (PR+ with 0-20 % cell positive or PR- or PR borderline) OR 2. HER2+ and (ER+ or PR+) (the previous luminal B definition) **Luminal NOS**: HER2- and (ER+ or ER-) and PR+ with unknown % positivity **Unclassified (previous definition)** Negative for all 5 markers: ER- and PR- and HER2- and HER1- and CK5/6-

AIMS (Ancestry) variables
(Available for subjects with genotyping data using Illumina assay) (N=3748)

Variable Name	Description	Data Source	Comments
MLE1_EURO	Proportion of European ancestry (range: 0.045 – 1.00)	Obtained from 144 AIMS SNPs	
MLE2_AFRO	Proportion of Africa ancestry (range: 0 – 0.955)	Obtained from 144 AIMS SNPs	= (1 – MLE1_EURO)