CBCS3 data for Elizabeth Hass
Examining the Impact of Diagnostic Delay on Care Quality, Tumor Biology, and Breast Cancer Survival: Aim 2 Preliminary Data
Dataset: Hass_CBCS3_032425 (N = 2998)

Variable name	Description	Comments
STUDYID	CBCS Study ID	
AGESEL	Age at BC diagnosis	
RACE	Race, used in sampling 1 = Non-African American 2 = African American	
SELF_RACE	Self-reported race 1 = White 2 = Black/African American 3 = American Indian, Eskimo 4 = Asian or Pacific Islander 5 = Other	
OTHER_RACE	Other race, specify 2 = Multi-racial 3 = Hispanic/Latino 10 = Arab/Arab-Berber	Only available for other race (SELF_RACE=5).
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	
ETHNICITY	Are you Hispanic? 1 = Hispanic 2 = Not Hispanic	
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	

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	
MONEY	Family income 1 = 15-30K 2 = 30-50K 3 = >50K 4 = <15K	Recoded from INCOME. "<15K" is coded as the reference category.
URBAN_RURAL_DX	Urban/rural status 1 = Urban 2 = Rural	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	Updated 01/03/2025.
FFAMHXBC	First-degree family history of breast cancer - parents or sibling(s) 0 = No 1 = Yes	

BCDAUGHYN	Breast cancer in any daughters 0 = No 1 = Yes 98= No daughters	
FFAMHXOC	First-degree family history of ovarian cancer – mother or sisters 0 = No 1 = Yes	
BCMOMLT50	Mother diagnosed with breast cancer before age 50 0 = No 1 = Yes	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	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	Sister is not counted if the 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	Sister is not counted if the 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	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	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.

AGEMENA	Age at menarche (range = 7-19)	One woman who had never menstruated is coded as missing.
MENA13G	Age at menarche 1 = < 13 year 2 = 13+	Cut point obtained from the median of CBCS 1 & 2 controls. "13+" is coded as the reference category.
MENO	Type of menopause experience 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	If the subject experienced menopause after age of diagnosis, she would be classified as premenopausal for this variable.
MENODATE	Date of menopause	This variable goes with the variable MENO. Missing for premenopausal (MENO=1) women.
AGEMENO	Age at menopause	This age variable is for the variable MENO. Missing for premenopausal (MENO=1) women.
POSTMENO	Menopausal status 0 = Premenopausal 1 = Postmenopausal	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	This variable goes with POSTMENO. Missing for premenopausal (POSTMENO=0) women.
PREGNUM	Number of pregnancies (range: 0-13)	Exclude pregnancies after age of diagnosis.
LIVEVER	Ever had live birth 0 = No 1 = Yes	Only include pregnancies that resulted in live birth.
NUMLIVEB	Number of live birth pregnancies (range: 0-11)	
OCEVER	Ever use oral contraceptives 0 = Never 1 = Ever	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	
TAMOXIFEN_S	Initiated Tamoxifen after dx (self-reported) 0 = No 1 = Yes	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	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	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	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	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	Define as “Yes” if any one of the 5 Endocrine drugs is coded as “Yes”.
ANYHRT	Any hormone replacement therapy 0 = Never use 1 = Ever use (3+ months)	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.
HRT_USE	Any hormone replacement therapy 0 = Never user 1 = current user 2 = past user	
OTHER_CANCER	History of other cancer 0 = No 1 = Yes	Excluding breast cancer. Age of cancer diagnosis <= AGESEL.
ALCOHOL	Ever used alcohol 0 = No 1 = Yes	Assume first started using alcohol before diagnosis of breast cancer.
EVERSMOK	Smoking status 0 = Never 1 = Ever	Account for age of diagnosis when all smoking variables were derived.

SMOKERS2	Smoking status 0 = Never 1 = Former 2 = Current	If age of smoking cessation > age of diagnosis, the subject would be considered as current smoker
BMICAT	BMI based on nurse measured data 1 = 25-<30 2 = 30+ 3 = <25	"<25" is coded as the reference category.
WAISTCM	Waist circumference measurement in cm (range: 55.9-165.1)	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.
FACT_G_TOTAL	FACT-G Total score (range: 0-108)	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)	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_TOTAL	FACIT-Sp total score (range: 0-156)	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
I included STRATA and WT just in case you need them – Jessica.		
STRATA	Sampling strata 111 = NonAA age <50 112 = NonAA age 50+ 113 = AA age <50 114 = AA age 50+	This is based on the race and age group. Subjects in each stratum have the same sampling probabilities.
WT	Sampling weights – inverse of the sampling probabilities	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**.

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.
NODESTAT	Node status 1 = Positive 2 = Negative	Recoded from ND_POS and N_STAGE. Positive is defined as one of the following: 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.
GRADE	Tumor grade 1 = Well differentiated 2 = Moderately differentiated 3 = Poorly differentiated 4 = Undifferentiated/Anaplastic differentiated 9 = not determined	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.

ERSTAT	ER status 1 = Positive 2 = Negative 3 = Weak Positive / Borderline	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	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

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.

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

CBCS3 Baseline Survey Variables from Section D and question I6

Please refer to this file for SAS variables (handwritten words) and codes for “Other, specify” fields:
Hass CBCS3 Baseline D & I questions.pdf

Nanostring data available for 1969 CBCS3 subjects

Variable Name	Variable Description	Details
SAMPLEID	Sample (Block) ID that RNA came from	
Basal	Correlation to Basal centroid in PAM50 algorithm	Continuous Numeric
Her2	Correlation to Her2 centroid in PAM50 algorithm	Continuous Numeric
LumA	Correlation to LumA centroid in PAM50 algorithm	Continuous Numeric
LumB	Correlation to LumB centroid in PAM50 algorithm	Continuous Numeric
Normal	Correlation to Normal centroid in PAM50 algorithm	Continuous Numeric
PAM50_Subtype	PAM50 Subtype	Basal = Basal-like, Her2=Her2, LumA = LuminalA, LumB=LuminalB, Normal=Normal-like
PAM50_Confidence	Confidence in final PAM50_Subtype call	Continuous Numeric
ROR_S	Risk of recurrence subtype only score (ROR-S)	Continuous Numeric
ROR_S_Group	ROR-S risk category	low = Low, med =Intermediate, high =High
Proliferation_Score	Proliferation Score from PAM50 assay	Continuous Numeric
ROR_P	Risk of recurrence + proliferation score (ROR-P)	Continuous Numeric
ROR_P_Group	ROR-P risk category	low = Low, med =Intermediate, high =High
ROR_T	Risk of recurrence + size score (ROR-T)	Continuous Numeric
ROR_T_Group	ROR-T risk category	low = Low, med =Intermediate, high =High
ROR_PT	Risk of recurrence + proliferation + size score (ROR-PT)	Continuous Numeric
ROR_PT_Group	ROR-PT risk category	low = Low, med =Intermediate, high =High

ER_Score	Expression of ESR1 in PAM50 assay	Continuous Numeric
Her2_Score	Expression of ERBB2 in PAM50 assay	Continuous Numeric
P53_Score	Correlation to P53-Mut-Like	Continuous Numeric
P53_Subtype	P53 RNA-based subtype	Mut-like = mutant-like, WT-like = Wild-type like
HRD	Homologous Recombination Deficiency	HRD High, HRD Low.
AGI	Genomic Instability (combination of HRD + P53 Subtype)	AGI = any genomic instability (TP53 mut like and/or HRD high), NGI = no genomic instability (TP53-WT and HRD low)

Gene expression data for 3 variables:

Variable Name
ESR1
PGR
ERBB2