

Harmonizing Cancer Registry Data across Five States: The ENCLAVE Study experience

Sandi Pruitt
October 13 2026

Ethnicity and Nativity in Cancer - Latino & Asian Enclaves: 5R01CA237540

MPI: Salma Shariff-Marco & Sandi Pruitt



University of California
San Francisco

DREA.II.LAB

UT Southwestern

Harold C. Simmons
Comprehensive Cancer Center

Ethnic Enclaves

Ethnically/culturally distinct neighborhoods

- High concentrations of individuals of the same ethnic origin
- High linguistic isolation
- A large share of immigrants and recent immigrants
- Ethnic businesses and resources



**San Francisco Asian American enclave:
Chinatown**



**West Dallas Hispanic/Latino enclave:
Fabrication street**

ENCLAVE Study

- Population: Asian American and Hispanic or Latino patients with breast, cervical, or colorectal cancer
- Setting: Five states: CA, FL, NJ, NY, TX
- Aims:
 - Define ethnic enclaves and examine cancer incidence by enclave residence
 - Assess associations of enclave residence + nativity with stage at diagnosis
 - Assess associations of enclave residence + nativity with survival

Census tract level enclave measure

Asian American enclave

Indicator Variable	Definition (numerator/denominator)
% Asian American	Asian American count/total population
% Foreign-Born Asian American	Asian American who are foreign-born/total population
% Limited English: API languages	People who speak English “not well” or “not at all” and speak API languages /total population
% Linguistically Isolated: API languages	Households that are linguistically isolated and speak API languages /total households

Hispanic/Latino enclave

Indicator Variable	Definition (numerator/denominator)
% Hispanic/Latino	Hispanic or Latino count/total population
% Foreign-Born Hispanic/Latino	Hispanic or Latino who are foreign-born/total population
% Limited English: Spanish	People who speak English “not well” or “not at all” and speak Spanish /total population
% Linguistically Isolated: Spanish	Households that are linguistically isolated and speak Spanish /total households

Why each registry required separate work

Shared foundation

- Common NAACCR concepts and item numbers
- Comparable cancer sites and diagnosis years
- A single set of analytic questions

Registry-specific reality

- Separate approvals, request forms, extracts, and transfer procedures
- Different variable names, components, and date granularity
- Different years, derivations, update schedules, and missingness
- Geographic linkage required tract, certainty, county, state, and census-year information

Standardized definitions provided a starting point. Comparable analysis required registry-by-registry decisions.

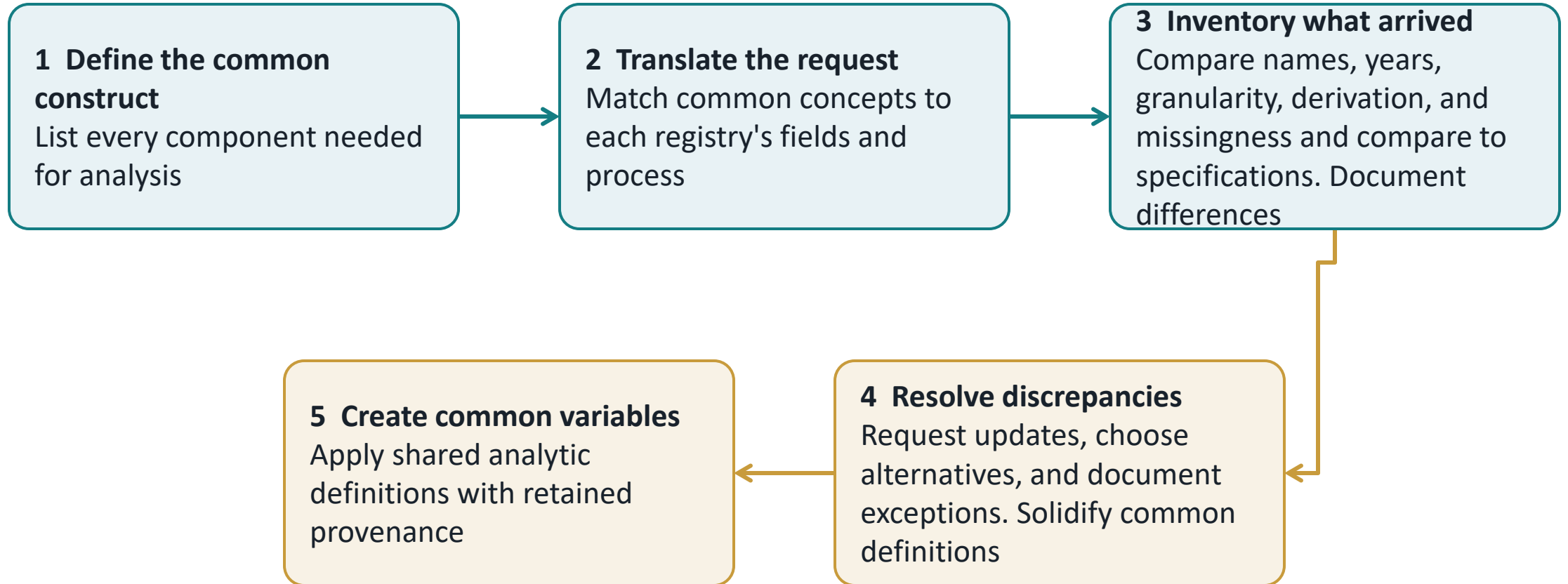
The work would not have been possible without staff support from each registry.

Five registries, five pathways for data acquisition

Registry labels are anonymized

Registry	How data were obtained	Setup required	Updates
A	Study team created the extract from an existing registry file	No inbound transfer	Later geocoding and vital-status files appended
B	Registry portal, then institutional managed transfer	Coordination across two secure systems	New registry extract replaced the earlier version
C	Institutional FTPS	Named sender and approved IP address	Same pathway reused for vital-status updates
D	Secure web-based managed file transfer	Transfer account plus separate vital-record approvals	Several supplemental files received
E	Registry-operated SFTP	Registry credentials; local IT not involved	Multiple targeted updates

The harmonization workflow



Registry_variables_received_deidentified

Search for tools, help, and more (Alt + Q)

FileHomeInsertSharePage LayoutFormulasDataReviewViewAutomateHelpDraw

CommentsCatch upEditingShare

Calibri (Body)11

Conditional FormattingCell StylesFormat As TableInsertDeleteFormat

Conditional FormattingCell StylesFormat As TableInsertDeleteFormat

E16

Received: marital_status_at_dx

Variable	NAAREg1 Item Number	NAAREg1 Field Name	Reg1 Variable Name (EPI file d	Reg2 (dataset Dec 27, 2021 + update	Reg3 (dataset Jan 25, 2022 + updated	Reg4 (dataset Mar 18, 2022 + vital sta	Reg5 (dataset Sept 11, 2023)									
	135	RACE--NAT (received Apr)	NAT	Received: nat	Received: RACE--NAT	Received: nat	n/a									
	254	Birthplace--Country	BIRTHPLACECOUNTRY	Received: birthplace_country	Received: BPCountry	Received: birthplace_country	Received: BIRTHPLACE_COUNTRY_N254									
	250	Birthplace		Received: birthplace	n/a	Received: birthplace	Received: BIRTHPLACE_N250									
	150	Marital Status at DX	MARSTAT	Received: marital_status_at_dx	n/a	Received: marital_status_at_dx	Received: MARITAL_STATUS_AT_DX_N150									
	630	Primary Payer at DX	PAYER	Received: primary_payer_at_dx	Received: PRIPAYERDX in updated dat	Received: primary_payer_at_dx	Received: PRIMARY_PAYER_AT_DX_N630									
	N/A	Secondary Payer at DX	PAYER2	n/a	n/a	n/a	n/a									
Follow-up/Vital Status	1760	Vital Status	VITSTAT	Received: vital_status	Received: VitalStatus	Received: vital_status	Received: VITAL_STATUS_N1760									
	1910	Cause of death	CDEATH	Received: primary_cause_of_death	Received: DthCause	Received: primary_cause_of_death	Received: Reg1USE_OF_DEATH_N1910									
	1750	Date of last contact	FUDATE	Received: date_of_last_contact_dd, c	Received: DateLastContact	Received: date_of_last_contact_yyyy	Received: DATE_OF_LAST_CONTACT_N1750									
	1785	Surv-Date Presumed Alive	SURV_DATELC_PSU	Received: surv_date_presumed_alive	Received: SurvDatePresAlive	Received: surv_date_presumed_alive	n/a									
	1786	Surv-Flag Presumed Alive	SURV_Reg5AG_PSU	Received: surv_flag_presumed_alive	Received: SurvFlagPresAlive	Received: surv_flag_presumed_alive	n/a									
	1787	Survival months - presumed alive	SURV_MON_PSU	Received: surv_mos_presumed_alive	Received: SurvMosPresAlive	Received: surv_mos_presumed_alive	n/a									
	1788	Surv-Date DX Recode	SURV_DATEDX_RECODE	Received: surv_date_dx_recode_dd, :	Received: SurvDateDXRecode	Received: surv_date_dx_recode_yyyy	n/a									
	1782	Surv_DateLC_Act	SURV_DATELC_ACT	Received: surv_date_active_fup_dd, :	Received: SurvDateActiveFU but data	Received: surv_date_active_fup_yyyy	n/a									
Census Tract of Residential Address at Diagnosis	1783	Surv_Flag_Act	SURV_Reg5AG_ACT	Received: surv_flag_active_fup	Received: SurvFlagActiveFU but data	Received: surv_flag_active_fup	n/a									
	1784	Surv_Mos_Act	SURV_MON_ACT	Received: surv_mos_active_fup	Received: SurvMosActiveFU but data	Received: surv_mos_active_fup	n/a									
	130	Census Tract 2000	CENSUS00	Received: dx_CenTract_2000	Received: CenTract2000	Received: census_tract_2000	Received: CENSUS_TRACT_2000_N130									
	135	Census Tract 2010	CENSUSTRACT2010	Received: dx_CenTract_2010	Received: CenTract2010	Received: census_tract_2010	Received: CENSUS_TRACT_2010_N135									
	365	Census Tr Certainty 2000	CENTRD00	Received: dx_CenTr_cert_2000	Received: CenTractCer2000	Received: census_tract_certainty_200	Received: CENSUS_TR_CERTAINTY_2000_N365									
	367	Census Tr Certainty 2010	CENSUSTRCERTAINTY2010	Received: dx_CenTr_cert_2010	Received: CenTractCer2010	Received: census_tract_certainty_20	Received: CENSUS_TR_CERTAINTY_2010_N367									
	90	County at DX Reported	COUNTY		Received: DXCOUNTY, DXSTATE in upc	Received: county_at_dx	Received: COUNTY_AT_DX_N90									
	n/a	County at DX Reported FIPS code	CNTYFIPS		Requested not received	n/a	n/a									
	95	County at DX Geocode2000	COUNTYATDXGEOCODE2000	Received: dx_countyFIPS2000	Requested not received	Received: county_at_dx_geocode_20	Received: COUNTY_DX_GEO_2000_N95									
	96	County at DX Geocode2010	COUNTYATDXGEOCODE2010	Received: dx_countyFIPS2010	Requested not received	Received: county_at_dx_geocode_20	Received: COUNTY_DX_GEO_2010_N96									
	82	State at DX Geocode 2000	DXSTATEGEO2000	Received: dx_stateFIPS2000	Requested not received	Received: state_at_dx_geocode_2000	n/a									
	83	State at DX Geocode 2010	DXSTATEGEO2010	Received: dx_stateFIPS2010	Requested not received	Received: state_at_dx_geocode_2010	n/a									
	390	Date_DX	DATEDX	Received: date_of_diagnosis_dd, dat	Received: DxDate, DxYear	No day variable. Received: date_of_d	Received: DATE_OF_DIAGNOSIS_N390									
	400	SITE	SITE_O2	Received: primary_site	Received: Psite	Received: primary_site	Received: PRIMARY_SITE_N400									
	522	Histologic Type ICD-O-3	HISTO_T3	Received: histology_icdo3	Received: HistTypeICDO3	Received: histology_icdo3	Received: HISTOLOGIC_TYPE_ICD_O_3_N522									
	523	Behavior Code ICD-O-3	HISTO_M3	Received: behavior_icdo3	Received: BehaviorICDO3	Received: behavior_icdo3	Received: BEHAVIOR_CODE_ICD_O_3_N523									

Variables (Feb 25, 2025)

Issues (Jul 21, 2025)

All formatting options

NumberAlignmentFontBorder

Horizontal AlignmentIndent

Vertical Alignment

Text control

Orientation (degrees)

Help Improve Formatting

Snapshot: translating the common request

Anonymized excerpt reconstructed from the working variable-comparison spreadsheet

Row	Common variable	Item	Registry A	Registry B	Registry C	Registry D	Registry E
4	Date of birth	240	DOB	Year only	Year, month, day	Year and month; no day	Birth date plus birth year
7	Race 2	161	Race2	Not available	race_2	race_2	Race2
8–10	Race 3–5	162–164	Race3–5	Not available	race_3–5	race_3–5	Not available
12	NHIA derived origin	191	NHIA	Not available	nhia	nhia	NHIA

Snapshot: tracking receipts and discrepancies

Anonymized excerpt reconstructed from the receipt and issues worksheets

Row	Question checked after receipt	Registry A	Registry B	Registry C	Registry D	Registry E	Team response
16	Were all race fields available?	Yes	Race 1 only	Updated after follow-up	Yes	Race 1–2 only	Use all available fields; document unequal inputs
37	Was a combined summary-stage variable supplied?	Yes	No	Received after request	No	Yes	Derive from year-specific fields where needed
41	Did follow-up extend beyond diagnosis end year?	Yes	Partial	Yes	Yes	Yes; one field initially empty	Set a common analytic cutoff and verify field meaning
68	Were day components available for key dates?	Yes	No day or month for birth	Yes	No day	Yes	Retain available precision; avoid false specificity

Snapshot: harmonizing summary stage

Anonymized excerpt reconstructed from the updated summary-stage decision table

Row	Source field	Version	Analytic years	Registry A	Registry B	Registry C	Registry D	Registry E
2	SEER summary stage 1977	1977	2000	Has	Has	Has	Has	Requested; not received
3	SEER summary stage 2000	2000	2001–2003	Has	Has	Has	Has	Requested; not received
4	Derived summary stage 2000	2000	2004–2015	Has	Missing; alternate differed	Has	Has	Requested; not received
5	Directly coded summary stage	2000	2016–2017	A-specific field	Missing	Missing	Missing	Missing
9	Combined summary stage	Across years	Final analysis	Used	Built from fields above	Used	Built from fields above	Used

Geography surprise: County!

- Geocoded data missing from one state, large missingness from another state
- NAACCR dictionary lists six different county variables
- State registries may have more, less, or different variables than NAACCR, even if the NAACCR item # is the same
- Version – counties change over time
- Surprise! Reported vs. from geocoded variable

Considerations for future multi-registry studies

- Invite, and compensate, registry staff to be part of study team
- Build the concept-to-field spreadsheet before submitting requests
- Ask about availability, derivation, years, precision, update timing, and meaning of blanks
- Request every component needed to construct geography, including certainty and county/state fields. Be mindful of version control and updates
- Plan for iterative clarification, replacement extracts, and supplemental vital status files
- Version the crosswalk, issue log, and derivation rules together
- Preserve source provenance in the analytic file and code
- Treat harmonization as an explicit part of the analytic method

Enclave Measures

- Technical report: <https://cancerregistry.ucsf.edu/resources/research-tools>
- UCSF Health Atlas – USA (50 states + DC, Puerto Rico)
<https://healthatlas.ucsf.edu>
 - 2010 and 2020 enclave quintile measures
 - View and download data directly from website
- ENCLAVE study dashboard – Five study states only; Visualize data here:
<https://cancerregistry.ucsf.edu/enclave>
 - 2000, 2010, 2020 enclave quintile and binary measures, and enclave characteristics of typology and predominant ethnic group
 - 2010 healthcare accessibility measures
 - 2010 mammography machine accessibility
 - View data online, request data from PIs

ENCLAVE study publications

- **Shariff-Marco S, Lin Gomez S, Canchola AJ, Fullington H, Hughes AE, Zhu H, Pruitt SL.** Nativity, ethnic enclave residence and breast cancer survival among Latinas: variations between California and Texas. *Cancer*. 2020;126:2849-2858.
- **Guan A, Pruitt SL, Henry KA, Lin K, Meltzer D, Canchola AJ, Rathod AB, Hughes AE, Kroenke CH, Gomez SL, Hiatt RA, Stroup A, Pinheiro P, Boscoe FP, Zhu H, Shariff-Marco S.** Asian American enclaves and healthcare accessibility: An ecologic study across five states. *American Journal of Preventive Medicine*. 2023;65(6):1015-1025.
- **Guan A, Shariff-Marco S, Henry KA, Lin K, Meltzer D, Canchola AJ, Arizpe A, Rathod AB, Hughes AE, Kroenke CH, Gomez SL, Hiatt RA, Stroup A, Pinheiro PS, Boscoe F, Zhu H, Pruitt SL.** Latino enclaves and healthcare accessibility: An ecologic study across five states. *Journal of General Internal Medicine*. 2025;40(4):739-748
- **Canchola AJ, Pruitt SL, Lin K, Boscoe FP, Guan A, Henry KA, Hiatt RA, Hughes AE, Insaf TZ, Pinheiro PS, Rathod AB, Shahan K, Stroup AM, Zhu H, Gomez SL, Shariff-Marco S.** Ethnic enclave residence and breast, cervical, and colorectal cancer incidence among Asian American adults, 2006-2017. *Cancer Epidemiology, Biomarkers & Prevention*. 2026;35(7):1139-1150.
- **Pruitt SL, Shahan KL, Canchola AJ, Henry KA, Guan A, Lin K, Boscoe FP, Insaf TZ, Hiatt RA, Hughes AE, Hyder S, Oh D, Pinheiro PS, Rathod AB, Stroup AM, Zhu H, Gomez SL, Shariff-Marco S.** Ethnic enclave residence and late-stage diagnosis of breast and colorectal cancers among Hispanic or Latino adults in four states. *Preventive Oncology and Epidemiology (in press)*



University of California
San Francisco

DREA.II.LAB

UTSouthwestern
Medical Center

THANK YOU ENCLAVE team!

DREAM Lab @ UCSF

Salma Shariff-Marco (MPI)



Alison J. Canchola



Scarlett Lin Gomez

Katherine Lin

Alice Guan

Laura Allen

Esperanza Castillo

Debby Oh

Dan Meltzer

Valerie McGuire

UCSF

Robert Hiatt

University of Texas Southwestern

Sandi Pruitt

Amy Hughes

Kathryn Shahan

Sofia Bereket

Others

Antoinette Stroup (Rutgers University)

Kevin Henry (Temple University)

Tabassum Insaf (Cancer Epi, New York State)

Paige Miller (Texas Cancer Registry)

Paulo Pinheiro (University of Miami)

Aniruddha Rathod (UAMS)

Hong Zhu (University of Virginia)

Francis Boscoe (Pumphandle Consulting)

Thank you, Cancer Registries: CA,
FL, NJ, NY, TX

NCI: R01CA237540 (Shariff-Marco
& Pruitt)