NIH Collaboratory

Health Care Systems Research Collaboratory

Demographic Information in Pragmatic Clinical Trials

Monique L. Anderson, MD
Assistant Professor of Medicine
Duke Clinical Research Institute
Duke University School of Medicine

Objectives

- Discuss importance collecting and analyzing demographic subgroup data
  - Race and Ethnicity
  - Socioeconomic Status
  - Insurance
- Discuss barriers and opportunities for demographic subgroup data collection and analysis
  - Variability and Missing Data in Electronic Health Records
  - Federal policies and Meaningful use
- Research on efforts to improve missing demographic data

Race and Ethnicity and Health Disparities

- Race and ethnicity are associated with substantial differences in access to care, utilization of health care services, and health outcomes in the US, particularly for African-Americans and Hispanics.[1]
- In America, the gap in life expectancy between Asian women and high-risk urban black men is as high as 20.7 years.[2]
- Observational studies and some prospective clinical trials show strong associations by race and ethnicity for disease prevalence, severity and outcome.[3]
  - Colon cancer, ESRD, MRSA infections, PTSD, CKD, HTN, DM,
- Associations by race/ethnicity often confounded by the complex interplay of socio-environmental factors. [3]
  - SES
  - Insurance
  - Discrimination
  - Acculturation

1. Anderson, Nahm, Califf. Under Review
2. Murray et al. Plos 2006. e260.
3. Braveman et al. JAMA 2005;294:2879-2888.

Race and Ethnicity in Clinical Research

- Significant debate still exists about the importance of documenting race and ethnicity in clinical research
  - Pros- plethora of literature demonstrating health disparities, collection better directs interventions
  - Cons- greater genetic variability within racial groups than between groups, misclassification issues that beset research, and concern that health disparities will perpetuate discrimination.[1]

- Several federal regulations to encourage the routine collection and reporting of race and ethnicity in clinical research.[2]
  - Revised Minimum OMB Categories required by the NIH and recommended by the FDA
  - Standards endorsed by IOM, AHRQ
  - Meaningful Use- >50% race/ethnicity, use OMB

- Paucity of data on the quality of demographic data in electronic health records[2]

1. Schwartz.NEJM 2001.
2. Anderson et al. Under review

Self-Reported vs Administrative Race/Ethnicity Data With Veteran Affairs

Agreement Between Self-Reported Race/Ethnicity and Administrative Data on Race/Ethnicity

	Administrative Data on Race/Ethnicity								
	Including Patients With Unknown Race/Ethnicity (n = 730 149)					Including Only Patients With Known Race/Ethnicity (n = 464 683)			
Self-Reported Race/Ethnicity	White, %	African American, %	Other, %	Unknown, %	n	White, %	African American, %	Other, %	n
Native American	49.34	4.14	**15.81**	30.71	8317	71.21	5.97	**22.82**	5763
Asian	10.73	1.13	**36.57**	51.57	4605	22.15	2.33	**75.52**	2230
African American	4.68	**60.63**	0.58	34.10	96 007	7.11	**92.0**	0.88	63 265
Hispanic	10.77	1.11	59.63	28.49	37 662	15.06	1.55	**83.39**	26 931
Pacific Islander	14.88	1.85	38.39	44.88	1727	27.00	3.36	**69.64**	952
White	**61.50**	0.44	0.88	37.17	581 831	**97.90**	.70	1.4	365 542

Note. Bold type indicates agreement between the 2 data sources.

Boehmer, U. Am J Public Health. 2003

Socioeconomic Status (SES)

- Paucity of data on the availability of SES data in the EHR
  - Experience in Duke EHR
    - Years of Education– 0% of the time
    - Occupation- 0.56% of the time
      - About 2500 patients out of >4,400,000 have this data collected.
- Paucity of availability of data in RCTs

- Increasing use of geocoded neighborhood-level SES variables in observational studies
  - More recently, the use of SES data within Medicare

Pragmatic Clinical Trials

- PCTs represent opportunities and challenges for yielding information on heterogeneity of treatment effect (HTE) for racial and socioeconomic status subgroups
- Pros-
  - Board population inclusion
  - Larger populations than clinical trials
  - Potential to yield meaningful information on populations previously not well-represented in traditional clinical trials.
  - Demonstrate the promise of EHR as robust source of data
- Cons-
  - Electronic health record data can be missing race and ethnicity information on most of its patients
  - Demographic data may have unknown degree of inaccuracy.
  - Standardization significant issue and may be a barrier to inter-institutional analyses
  - Race and SES too correlated to measure both
  - Socioeconomic status will be unavailable and/or very incomplete
  - Given clustering in trials, statistical power may not be adequate to examine subgroup differences for demographic groups.

Demonstration Projects- Race and Ethnicity Categories Data Collection

Category	Trauma*	Proven	PPACT	PTSD*	ICD-Pieces+	LIRE&	Anonymous	Anonymous
Race								
White	X	X	X	X	X	X	X	
Non-White				X				
Black or AA	X	X	X	X	X	X	X	
Asian	X	X	X	X	X	X	X	
NHOPI	X	X	X		X	X	X	
AIAN	X	X	X			X	X	
Multiple Races						X		
Hispanic	X			X				
Mexican				X				
Mexican American				X				
Chicana				X				
Cuban				X				
Spanish				X				
South American				X				
Indian					X			
Unknown			X		X	X	X	
Other	X	X	X		X			
Ethnicity								
Hispanic		X	X		X		X	

* combined race and ethnicity format

LIRE- not all sites collecting the same race categories

ICD-9 Pieces- not all sites collect Hispanic ethnicity

rials

Demonstration Projects-Data Collection on SES and Insurance Status

PCT	Socioeconomic Status*	Geocoded SES	Insurance
Trauma*	X		X
Proven			X
PPACT			X
PTSD*	X		X
ICD-Pieces+	X		X
LIRE&			X
Anonymous			X
Anonymous			X

* two PCTs listed insurance as measure of SES

Race and Ethnicity Distribution of Health Plan Membership in Kaiser Permanente Southern California

Race	Percent
Historical members up to May 31, 2011 (n=12,764,185)	
White	15.1
Hispanic	15.1
Black	4.2
Asian and Pacific Islander	2.9
American Indian and Alaska Native	0.1
Multiracial	0.1
Other	0.9
è Unknown (missing)	61.6
Active members on January 1, 2009 (n=3,323,588)	
White	25.6
Hispanic	30.1
Black	7.6
Asian and Pacific Islander	6.2
American Indian and Alaska Native	0.1
Multiracial	0.2
Other	1.9
è Unknown (missing)	28.3

Derose SF et al. Medical Care Research and Review. 2012: 70(3)330-345

Race and Ethnicity of Duke University Health System Patients

Race and Ethnicity	Percent
Unique Patient Records as of June 2008-August 1, 2014	n=4,459,451
White	53.9
Black	16
Asian	1.3
American Indian and Alaska Native	0.6
Native Hawaiian or Other Pacific Islander	0.05
Multiracial	0.2
Other	0.2
è Unknown, Unavailable, Null	23.9-25.3
Hispanic Ethnicity (distinct from race)	1.4

Provided by Duke Enterprise Warehouse, Geospatial Analyst, 8/1/2014

Indirect Estimation for Missing Race Data

- Indirect Estimation for Race and Ethnicity has been encouraged by the Agency for Healthcare Research and Quality and the Institute of Medicine[1]
- Organizations currently using these data:
  - Kaiser-Permanente Geographically Enriched Member Socio-demographics datamart (GEMS)[2]
  - Medicare[3]
  - Health plans (Aetna)
- Several methods developed to estimate missing race data in EHR and administrative records[3]
  - Surname
  - Geocoding only
  - Bayesian Surname Geocoding
  - Bayesian Improved Surname Geocoding

1. IOM Race, Ethnicity, and Language Data: Standardization for Heath Care Quality Improvement. 20
2. http://share.kaiserpermanente.org/static/cb_annualreport/reports/docs/2011_chapters/cb11_healthy_people.pdf
3. Bonito AJ, et al. Creation of New Race-Ethnicity Codes and SES Indicators for Medicare Beneficiaries. AHRQ Publication No. 08-0029_EF. January 2008

Capabilities at Duke for Indirect Estimations for Race/Ethnicity and SES

- Implemented automated address standardization and geocoding process
- SAS Data Management Studio (formally DataFlux)
  - Verify, standardize, and geocode address information coming into the Enterprise Data Warehouse.
  - USPS Quality Knowledge Base and TomTom+6 street-level geocoding
- Patients assigned rooftop coordinates (latitude/longitude)
- Addresses refresh nightly, takes 1 hour. Usually 2000-3000 addresses processed.

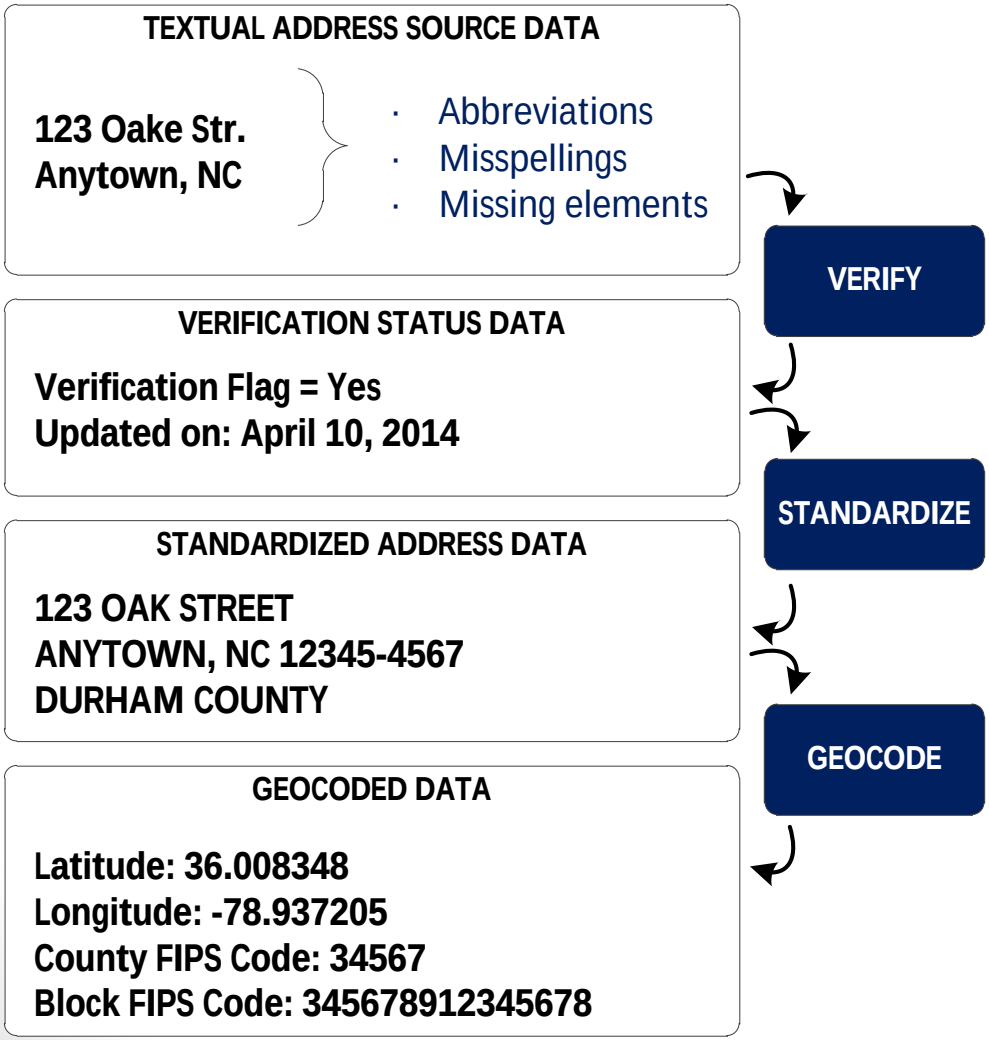
Demographic and Socioeconomic Data Linkage into Duke Medicine Enterprise Data Warehouse

- Common Fund Supplement awarded Sept 2013
  - Understand if indirect estimators for race/ethnicity and SES could be implemented in health systems
- Census data extracted, transformed, and loaded into Duke Medicine's EDW
  - US Census Bureau's 2010 Summary File 1
  - American Community Survey (2007-2011)
    - Socioeconomic status
    - Racial information
- Addresses assigned to a census block group FIPS code and combined with census race and ethnicity, as well as socioeconomic, data

Design and Implementation of an Automated Geocoding Infrastructure for the Duke Medicine Enterprise Data Warehouse

Shelley A. Rusincovitch, Sohayla Pruitt, MS, Rebecca Gray, DPhil, Kevin Li, PhD, Monique L. Anderson, MD, Stephanie W. Brinson, Jeffrey M. Ferranti, MD, MS

Duke Medicine, Durham, North Carolina



- 4,080,966 patient address records (82.2% of total) have been verified and standardized
- 87.7% of standardized patient address records have been geocoded
- Geocoded records are 72.1% of total patient address records

Bayesian Improved Surname Geocoding

- Indirect estimation approach using Baye's Theorem[1,2]
  - designed and utilized to account for missing race and ethnicity in administrative data
    - Insurance Plans
    - Kaiser Health System

- Individuals are assigned a set of probabilities for membership in each racial/ethnic group given their surname and place of residence.

- Inputs for calculation:
  - the probability of a selected race given surname
  - proportion of all people in US who self report being race *i* who reside in Census Block Group *k*
- Data Input
  - 2010 Census Data
  - 2000 Surname File
  - Electronic Health Record (name, address, race)
- Data Output
  - Set of probabilities for 6 races
  - Race cutpoint chosen if a particular probability reaches 0.50.

1.Elliot, MN et al. Health Serv Outcomes Res Method 2009. 9:69-83
2. Derose, SF. Medical Care Review and Review 2012. 70(3) 330-345.

2000 Surname File

Last Names				People with these Names		
Frequency of Occurrence	Number	Cumulative Number	Cumulative Proportion (percent)	Number	Cumulative Number	Cumulative Proportion (percent)
1,000,000+	7	7	0.0	10,710,446	10,710,446	4.0
100,000-999,999	268	275	0.0	60,091,601	70,802,047	26.2
10,000-99,999	3,012	3,287	0.1	77,657,334	148,459,381	55.0
1,000-9,999	20,369	23,656	0.4	58,264,607	206,723,988	76.6
100-999	128,015	151,671	2.4	35,397,085	242,121,073	89.8
50-99	105,609	257,280	4.1	7,358,924	249,479,997	92.5
25-49	166,059	423,339	6.8	5,772,510	255,252,507	94.6
10-24	331,518	754,857	12.1	5,092,320	260,344,827	96.5
5-9	395,600	1,150,457	18.4	2,568,209	262,913,036	97.5
2-4	1,056,992	2,207,449	35.3	2,808,085	265,721,121	98.5
1	4,040,966	6,248,415	100.0	4,040,966	269,762,087	100.0

Word et al. Demographic Aspects of Surnames for Census 2000

2000 Surname File- Probability of Race/Ethnicity for Supplement Investigators

Name	Rank	prop100k	pctwhite	pctblack	pctapi	pctaian	pct2prace	pcthispanic
ANDERSON	12	282.62	77.6	18.06	0.48	0.7	1.59	1.58
CALIFF	37688	0.21	92.61	3.06	1.62	(S)	(S)	1.26
HERNANDEZ	15	706372	4.55	0.38	0.65	0.27	0.35	93.81

2010 Census Block Group Data Sample Population

Patient	BG-FIPS	Tot Pop	White	Black	AIAN	Asian	NHOPI	Other	Multi	Hispanic
1	370690604023	1730	539	1149	4	0	0	7	31	63
2	370630020133	2030	1326	409	18	135	1	83	58	198
3	370370201031	3947	2991	323	11	101	2	440	79	849
4	370630018022	1629	205	950	12	15	0	415	32	530
5	370319702002	561	552	0	2	0	0	0	7	0
6	370339305001	1135	630	485	1	0	0	1	18	16
7	371539705002	696	369	256	17	5	0	32	17	63

Sample BISG Imputation Probabilities from Sample Population

Self-Reported Race	P_WHITE	P_BLACK	P_AIAN	P_ASIAN	P_HISPANIC	P_MULTIPLE	BISG Imputed Race
Black	0.276	0.706	0.001	0.000	0.006	0.012	Black
White	0.687	0.249	0.005	0.013	0.019	0.027	White
Asian	0.016	0.000	0.000	0.958	0.014	0.012	Asian
Hispanic	0.002	0.007	0.000	0.000	0.990	0.001	Hispanic
White	0.985	0.000	0.002	0.000	0.000	0.013	White
Black	0.496	0.491	0.000	0.000	0.002	0.010	Unassigned
Black	0.909	0.041	0.011	0.001	0.021	0.016	White
Unavailable	0.673	0.170	0.044	0.005	0.091	0.016	White

BISG Assigned Race Category from based on calculated race probability > 0.50

Pilot Testing of BISG Algorithm Using General Cardiology Clinic Patient Population

- Pilot population of 447 patients
  - Surname data available on 90.2% of patients
  - 76.7% addresses able to be assigned to block group Census Data
    - Slightly better than geocoded percentage for all of Duke patients- 72.1%
  - Imputation of Missing EHR Race/Ethnicity Data (n=14)
    - White=10
    - Black=2
    - No imputation=1 (no probability greater than 50%)

Pilot Experience with BISG in Duke Data

	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
Black	60.1	91.0	86.0	71.3
White	87.1	68.9	68.8	87.1
Hispanic	83.3	99.5	83.3	99.5
Asian	83.3	98.2	50.0	96.3

SES Index

- Utilized previously validated SES index score[1,2]
  - Created based on measure popularized by Kreiger[1]
  - Developed to help understand health and health disparities
  - Validated by AHRQ for use in Medicare Data
- SES score- multidimensional construct accounting for wealth, income, education, housing, and occupation[2]
- Assignment of SES index score at block group level
  - 211,267 block groups in US

Krieger N, et al.Am J Epideiol. 2003;57(3):186-99
Bonito AJ, et al. Creation of New Race-Ethnicity Codes and SES Indicators for Medicare Beneficiaries. AHRQ Publication No. 08-0029_EF. January 2008

Socioeconomic Index Score

Construct	Measure	Definition
Occupation		
	Unemployment	Percentage of persons aged 16 years or older in the labor force who are unemployed (and actively seeking work)
Income		
	Below US Poverty Line	Percentage of persons below the federally defined poverty line
	Median Income	Median household income
Wealth		
	Property Values	Median value of owner-occupied homes
Education		
	Low Education	Percentage of persons aged > 25 years with less than a 12th-grade education
	High Education	Percentage of persons aged > 25 years with at least 4 years of college
Housing		
	Crowded households	Percentage of households containing one or more person per room

Integration of SES index score in DEDUCE Research Portal

- SES score[1]
  - Census block group SES index scores calculated using 8 variables
  - Scores then assigned to all patients whose addresses were able to be geocoded
    - N=2,070,519 in Duke Health System
    - Range 35-78
  - SES index quartiles created for use in research
    - SES Q1- 35-48
    - SES Q2- 49-51
    - SES Q3- 52-55
    - SES Q4- 56-78

Bonito AJ, et al. Creation of New Race-Ethnicity Codes and SES Indicators for Medicare Beneficiaries. AHRQ Publication No. 08-0029_EF. January 2008

Implementation of SES Index within Duke Medicine Health System's Research Infrastructure

DEDUCE Charts

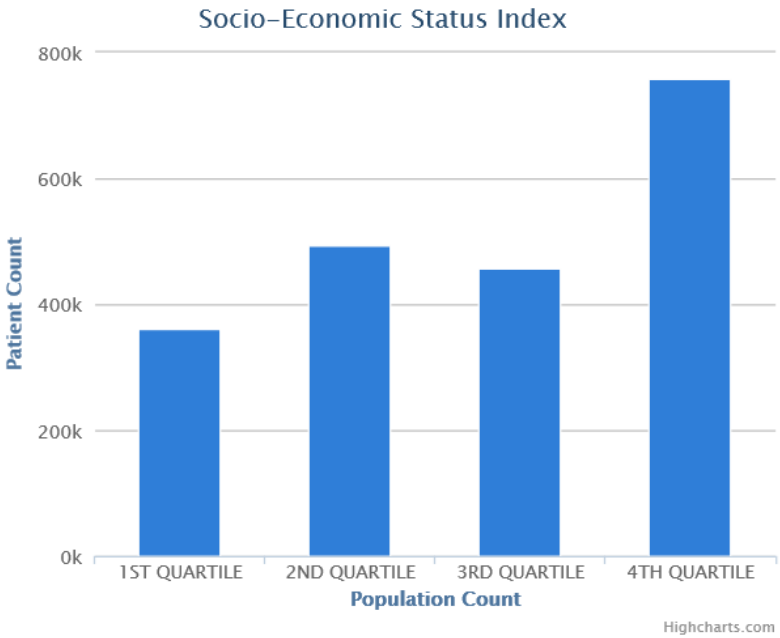
Cohort:1.2 SESINDEX Between 0,100

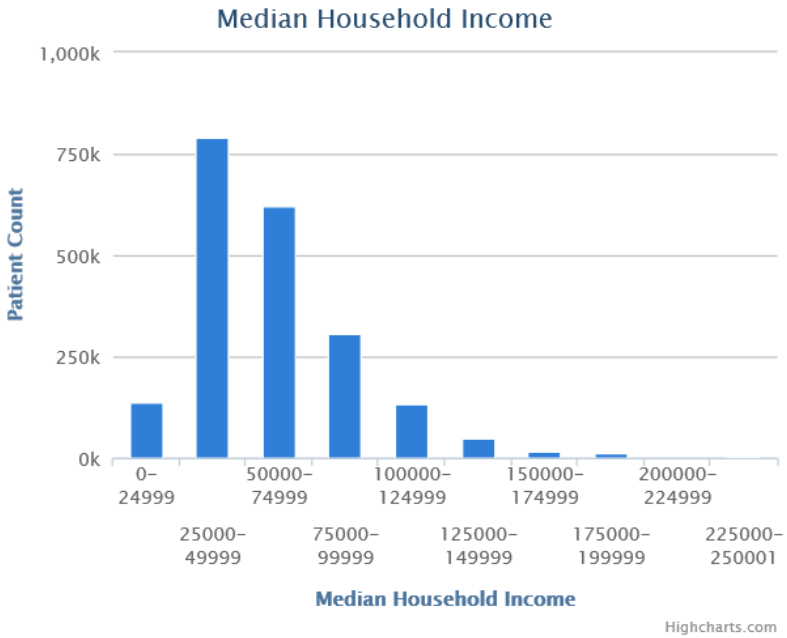
Click buttons to view additonal charts:

Patient Demographics | Inpatient | Outpatient | Patient Counts by State,Gender,Race,Death Indicator | Discharge by Gender | Discharge Date by Race | Socioeconomic Status

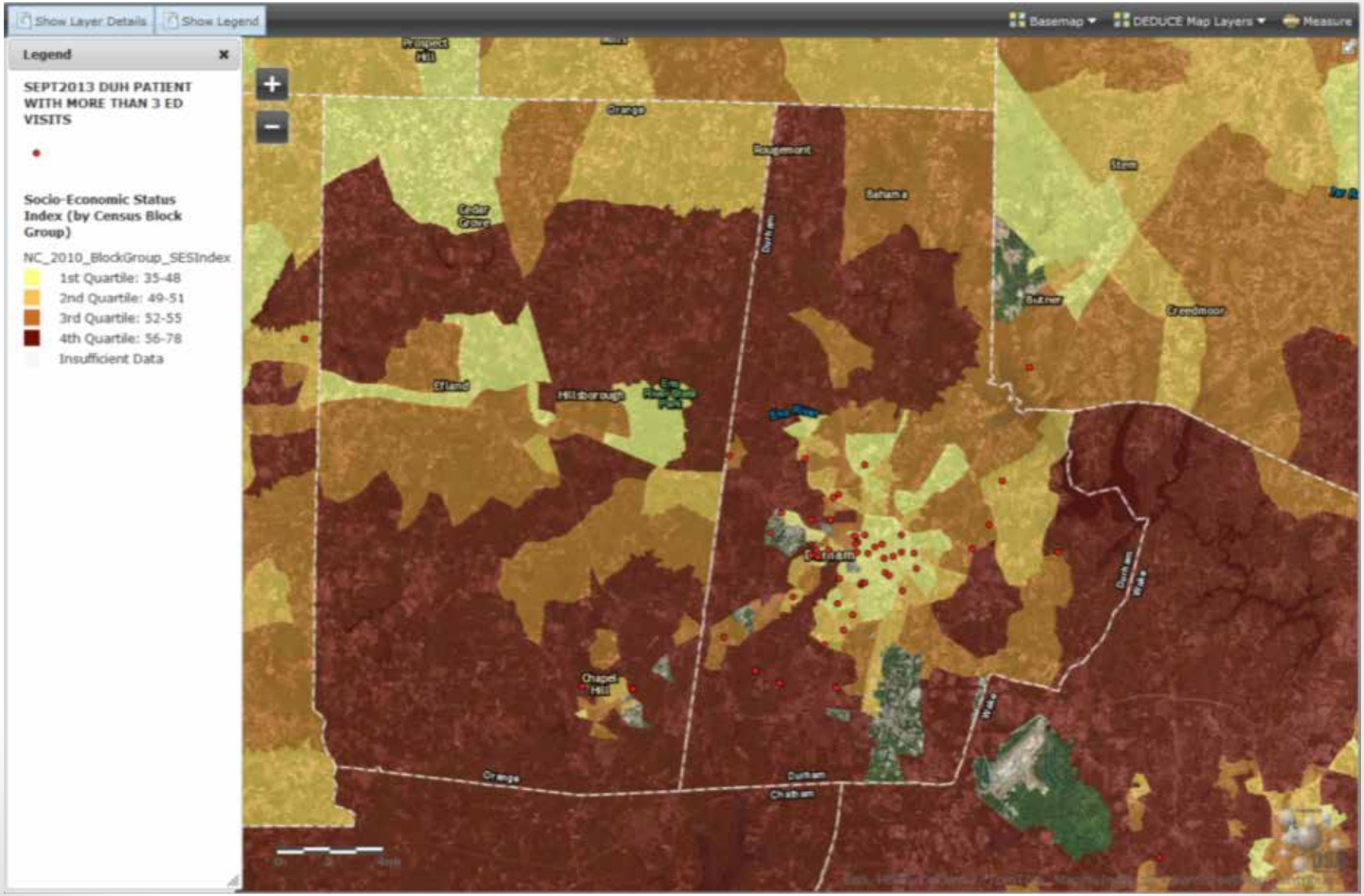
Neighborhood Level Statistics for Geocoded Patients in the United States

Source: American Community Survey, 2007–2011 5-year Estimates by Census Block Group





Socioeconomic Status and Clustering of Patients with High ED Utilization in Durham County



Limitations

- SES index
  - Validation needed for SES within respective health systems, but efforts hampered by high degree of missing data
  - Strength is that the index score would be identical across health systems.
- BISG imputation helps, but not perfect
  - No data on names occurring less than 100 times
  - Does not predict multi-racial and AIAN individuals well.
  - If block group or surname missing, will need to use less accurate imputation methods.
  - Slightly lower accuracy with women
  - Surname file due to be updated (expected in 2015)
- Automated Algorithm behind DUH firewall
  - Need to understand how to make data and algorithm generalizable

Ongoing Work

- Implementation and Validation of BISG in Duke EHR
  - Understand how to make BISG imputation and SES index variables available to demonstration projects interested
- Systematic Review to Examine Methodology currently used to examine subgroup analyses in PCTs, with a focus on CRTs
- Simulation Modeling Experiments to Better Understand how to Detect HTE for Demographic Subgroups in Cluster Randomized Trials.
  - Computer Simulations using select disease processes (CHF) with known survival outcomes and treatment effect by race and ethnicity

Conclusions

- Pragmatic clinical trials will include broader and larger populations than most phase 3 RCTs.
- These trials may yield new information for across racial, sex, and socioeconomic subgroups.
- Relative importance of demographic data collection and subgroup analysis will need to be determined for NIH and PCORI leadership and investigators.
- Investigators will likely need to plan for missing racial/ethnic and SES data.
  - Geospatially derived information could help to supplement missing race, ethnicity, and SES across health care systems.

Acknowledgements

- Robert Califf, MD
- Adrian Hernandez, MD
- Eric Peterson, MD
- Karen Chiswell, PhD
- Sohayla Pruitt, MA
- Yuliya Lokhnygina, PhD
- Meredith Nahm, PhD
- Asba Tasneem, PhD
- James Topping, MS
- Judy Stafford
- Josephine Briggs, MD
- Wendy Weber, ND, PhD, MPH
- Catherine Myers, MD
- Cheri, Tammy, Darcy, Michelle
- Jonathan McCall, Gina, Liz

Research reported in this publication was supported by the Common Fund Research Supplements To Promote Diversity In Health Related Research under Award Number 3U54AT007748-02S1 and the Health Care Systems Research Collaboratory Coordinating Center under Award Number 1U54AT007748-01 the National Center for Complementary and Integrative Health, a center of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

NIH Proposed Policy to Enhance Transparency

- Proposing to issue a policy to ensure that all NIH-funded clinical trials are registered and have summary results in ClinicalTrials.gov

- Compliance with the policy will be a term and condition in the Notice of Grant and a contract requirement in the Contract Award.

Duke Response to NPRM FDAAA- Race and Ethnicity

Results Submission- Subpart C

- **Demographic and Baseline Characteristics (FD 69639)**

"We do not propose to require the submission of information describing all of these demographic characteristics because they may not all be collected as part of a particular clinical trial, and we do not wish to impose requirements on the data that must be collected during a clinical trial. Instead, in § 11.48(a)(2)(iii), we propose as a minimum requirement that responsible parties submit information describing the age and gender of the human subjects enrolled in the clinical trial… Such information is generally collected in clinical trials and can be expected to be available for applicable clinical trials."

We request that race and ethnicity become mandatory data elements in the ClinicalTrials.gov results database, in addition to age and gender. Reporting of race and ethnicity is of primary importance both to the NIH and the FDA, as evident by multiple policies including the NIH Revitalization Act and FDA's Demographic Rule. The reporting of such information is also critical to patients, physicians, researchers, and policy makers seeking to understand how representative minority populations are in ACTs as well as whether the safety and effectiveness of medical products differ in racial/ethnic subgroups. There may be two concerns with making race and ethnicity mandatory: 1) an ACT may not collect race and ethnicity and 2) race and ethnicity may not be collected in a standard fashion. We believe such concerns related to complete demographic data collection and standardization need to be addressed, but should not be limiting factors in proceeding with making race and ethnicity mandatory.