Machine Made Sampling Designs: Applying Machine Learning Methods for Generating Stratified Sampling Designs

Trent D. Buskirk, UMass Boston Center for Survey Research

Todd Bear, University of Pittsburgh School of Public Health

Jeff Bareham, Marketing Systems Group

BIGSURV18 ⚙ Barcelona ⚙ October 25, 2018

UMASS BOSTON

University of Pittsburgh

Marketing Systems Group

1

Background and Goals

- Study Goals
- Landline RDD Sampling Number Generation

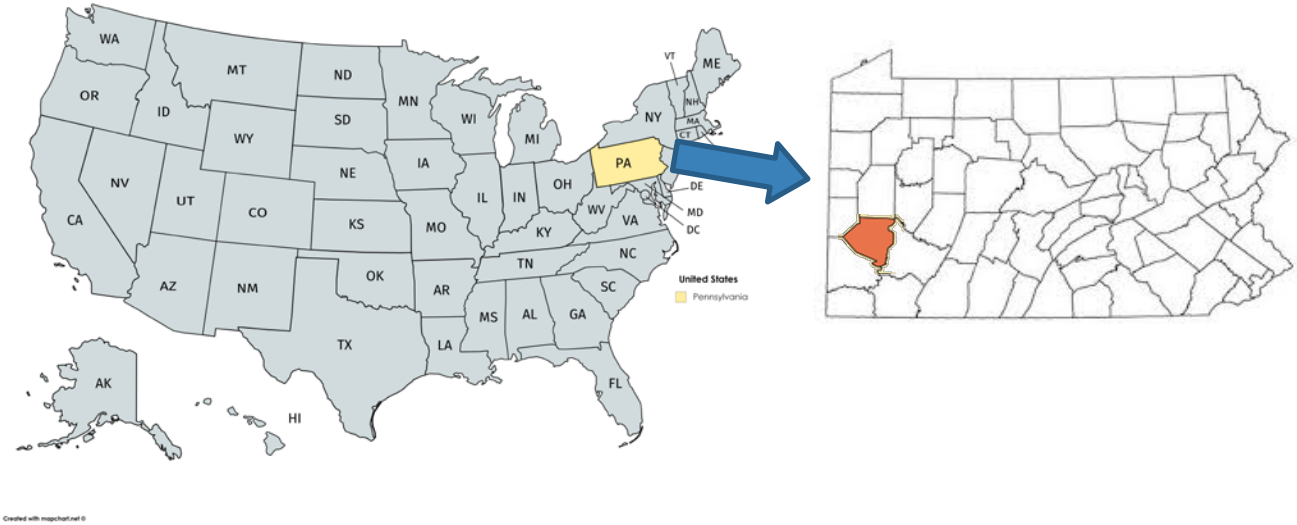
Survey Background and Goals

- The Allegheny County Health Survey (ACHS) is a population based, telephone survey of health related behaviors for adults living in Allegheny County, PA.
- The survey is modelled after the Centers for Disease Control and Prevention's BRFSS
  - More information about the survey can be found here: http://bit.ly/ACHS-2016
- Goal was to tie health information to political units (council districts defined by geopolitical boundaries)
- Sampling from 13 council districts would provide higher resolution than previous surveys that simply stratified by RACE or INCOME

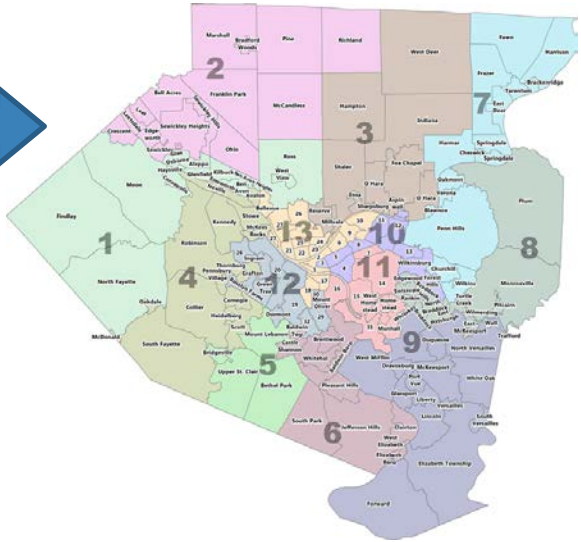
UMASS BOSTON

Study Goals and Background

- **Main Task:** Develop a stratified RDD sampling plan to select households from the 13 council districts in Allegheny County, PA.



Allegheny County, PA



13 Council Districts each defined to have roughly the same population

2

Methods

Clustering (not cluster sampling)

Our Approach…

Landline Numbers organized into 1K Banks

Admin. data available for some landline numbers allowing identification of local geography (LLKGs)

Council District assignments appended based on known boundaries

Machine learning algorithm (k-means clustering) applied to determine strata

UMASS BOSTON

Telephone Number Generation in the U.S.

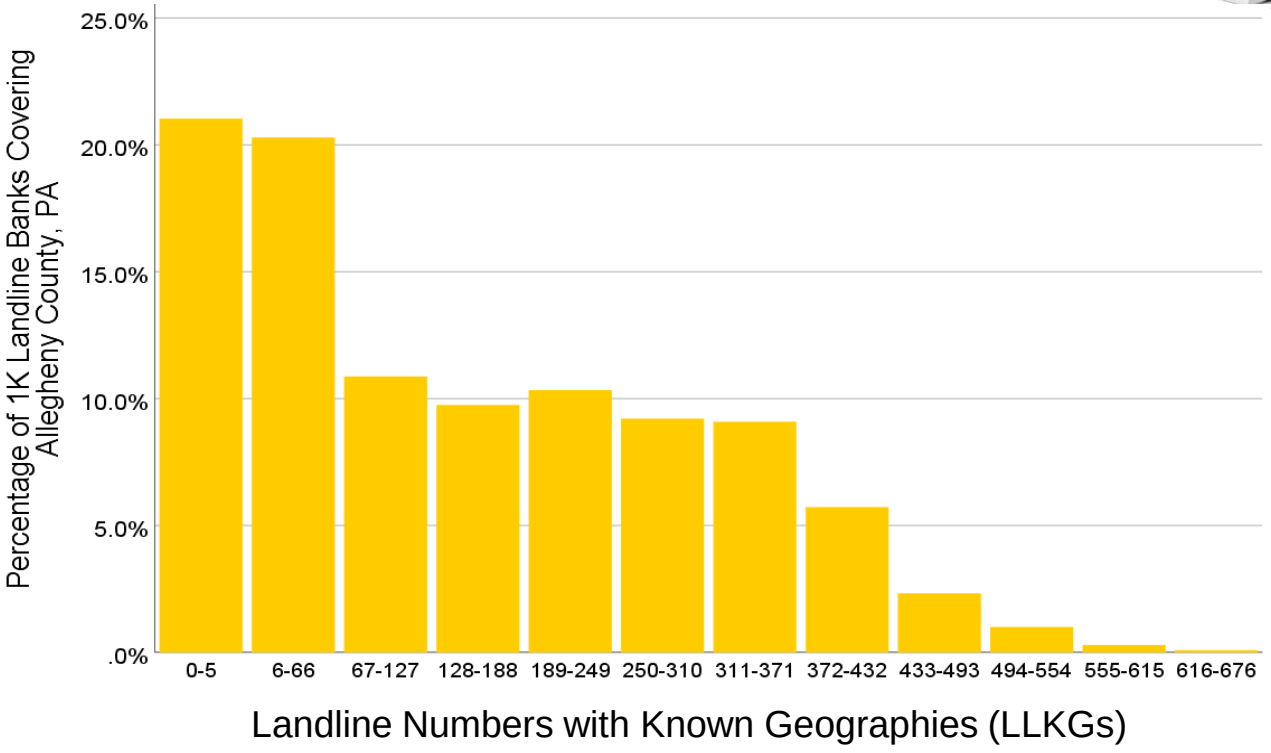
Area Code | Prefix | 10K Bank | Last 2 of Phone #

(ABC)-DEF-GH XY

1000 Bank

(1K Bank)

Distribution of 1K Banks by Number of LLKGs



A closer look at the LLKG's

2,410
1K banks

507
with 5 or fewer LLKGs

1,903
1K banks

367,085
LLKGs

988
from eliminated 1K banks

366,097
LLKGs

Data used in k-means clustering

$\tau_{1,1}$	$\tau_{1,2}$	$\tau_{1,3}$	▪▪▪	$\tau_{1,13}$
$\tau_{2,1}$	$\tau_{2,2}$	$\tau_{2,3}$	▪▪▪	$\tau_{2,13}$
⋮				
$\tau_{1903,1}$	$\tau_{1903,2}$	$\tau_{1903,3}$	▪▪▪	$\tau_{1903,13}$

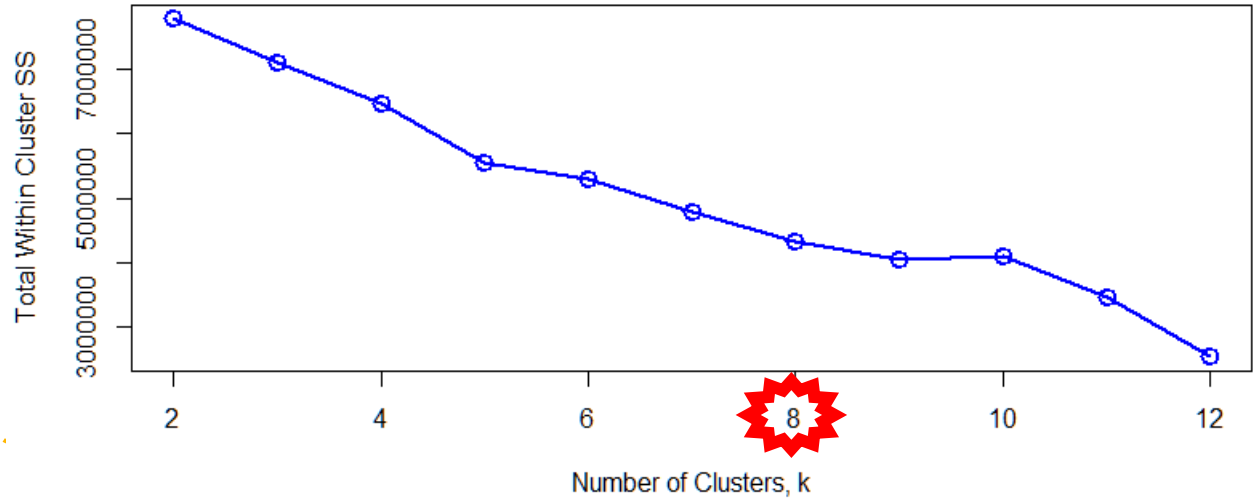
$\tau_{i,j}$ = proportion of LLKGs in 1K bank i within CD j

ω_i = total # of LLKGs in 1K bank i falling in CDs

UMASS BOSTON

Identifying the Clusters

- The k-means clustering algorithm optimizes groupings based on a within cluster SS criterion based on a multivariate set of covariates.
- We used a scree type plot created using results of a grid search for k possible clusters ranging from 2 to 12.



From Clusters to Strata

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8

HD 1
HD 2
HD 3
HD 4
⋮
HD 12
HD 13

Stratum 1
Stratum ?
Stratum ?

Using counts of the LLKGs within each cell of the cross tabulation.

UMASS BOSTON

3

Results

From Clusters to Strata…

Distribution of LLKGs by Cluster and Health District (HD)

Cluster	CD1	CD2	CD3	CD4	CD5	CD6	CD7	CD8	CD9	CD10	CD11	CD12	CD13
3	9	2	8	85	341	20,404	4	6	342	2	14	4,183	8
5	1,504	36	29	22,354	24,258	4,115	14	6	14	36	1,535	18,234	2,440
1	2	3	28	-	2	1,223	5	17	18,147	2	9	2	4
2	22,632	21,460	24,237	2,215	18	78	142	68	33	32	33	47	10,954
4	6	3	11	2	10	30	7	389	2,169	47	12,024	17	31
6	9	6	12	7	4	6	8,551	17,928	1,882	936	25	4	26
7	20	23	104	21	31	18	5,506	2,234	15	15,377	5,579	353	5,191
8	-	2	689	-	3	-	11,158	300	-	4	-	-	1

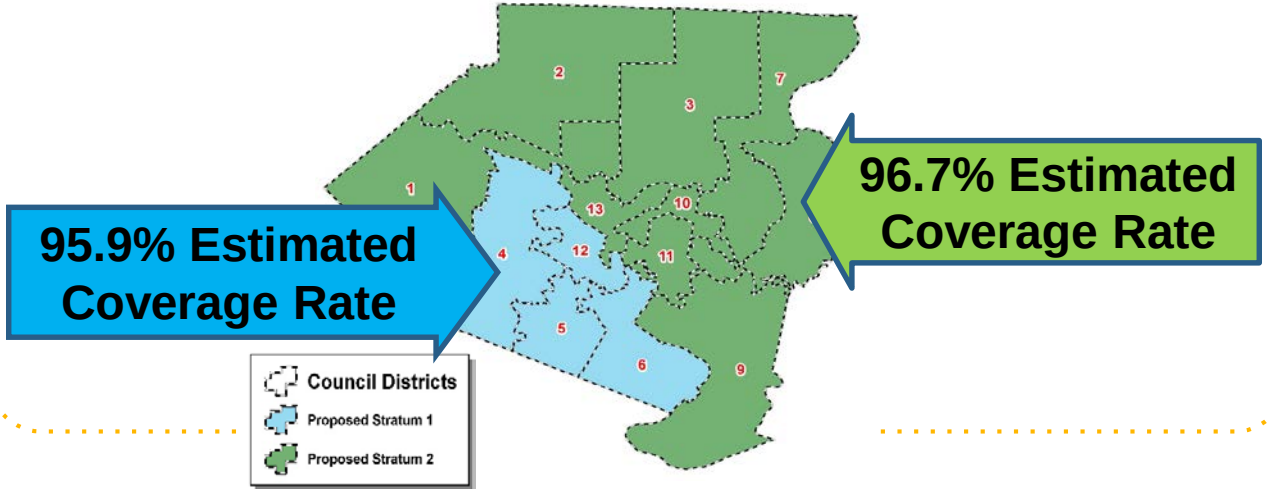
UMASS BOSTON

Final Distribution of LLKGs by Strata and Council District

Council District -->	CD4	CD5	CD6	CD12	CD1	CD2	CD3	CD7	CD8	CD9	CD10	CD11	CD13
LLKGs in Stratum 1	22439	24599	24519	22417	1513	38	37	18	12	356	38	1549	2448
LLKGs in Stratum 2	2245	68	1355	423	22669	21497	25081	25369	20936	22246	16398	17670	16207
Percentage of LLKGs in Council Districts Covered by Stratum 1	**90.9%**	**99.7%**	**94.8%**	**98.1%**	6.3%	0.2%	0.1%	0.1%	0.1%	1.6%	0.2%	8.1%	13.1%
Percentage of LLKGs in Council Districts Covered by Stratum 2	9.1%	0.3%	5.2%	1.9%	**93.7%**	**99.8%**	**99.9%**	**99.9%**	**99.9%**	**98.4%**	**99.8%**	**91.9%**	**86.9%**

From Clusters to Strata…

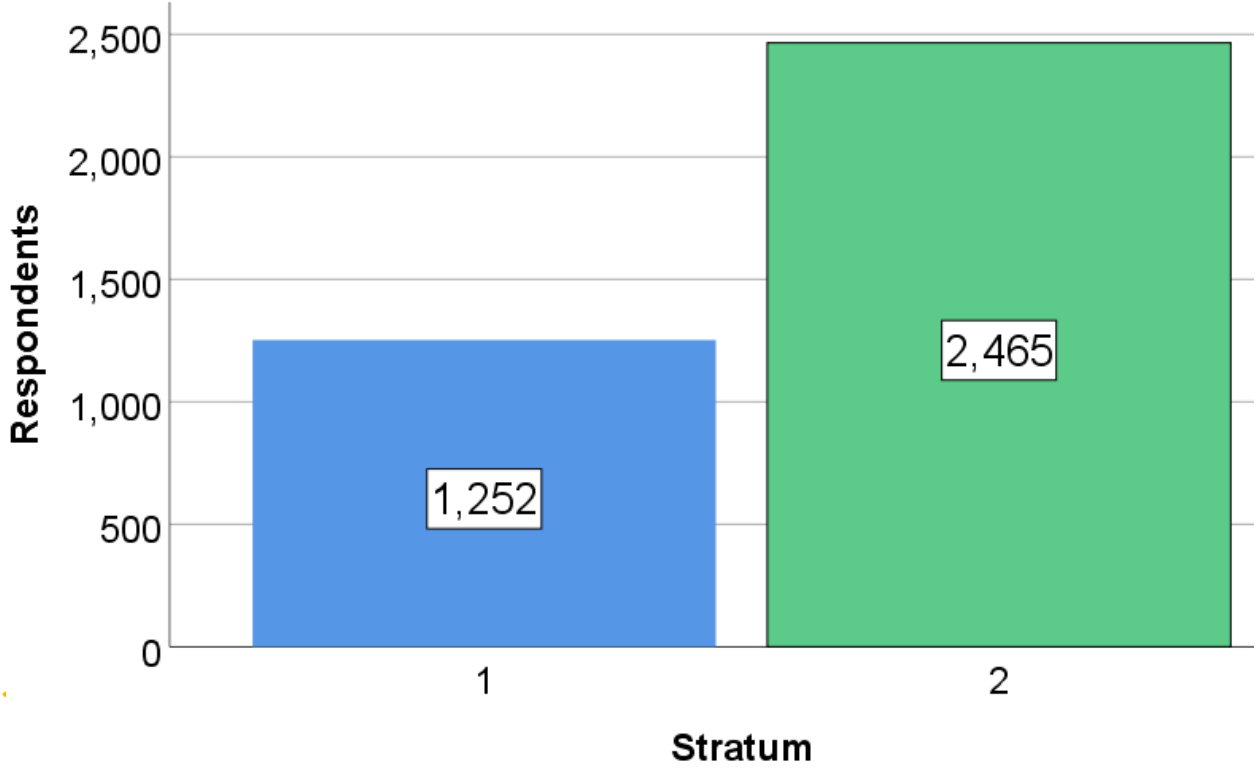
- Stratum 1 combined two of the 8 clusters to cover Council Districts 4, 5, 6 and 12
- Stratum 2 combined the remaining 6 clusters to cover Council Districts: 1,2,3,7,8,9,10,11,13
- The two strata were further partitioned into Listed LL and Unlisted LL substrata.



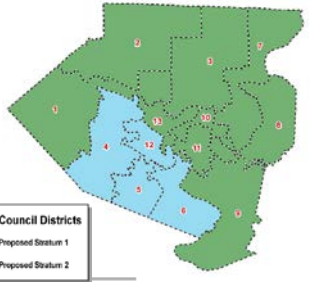
Universe and Sample Sizes & Yield

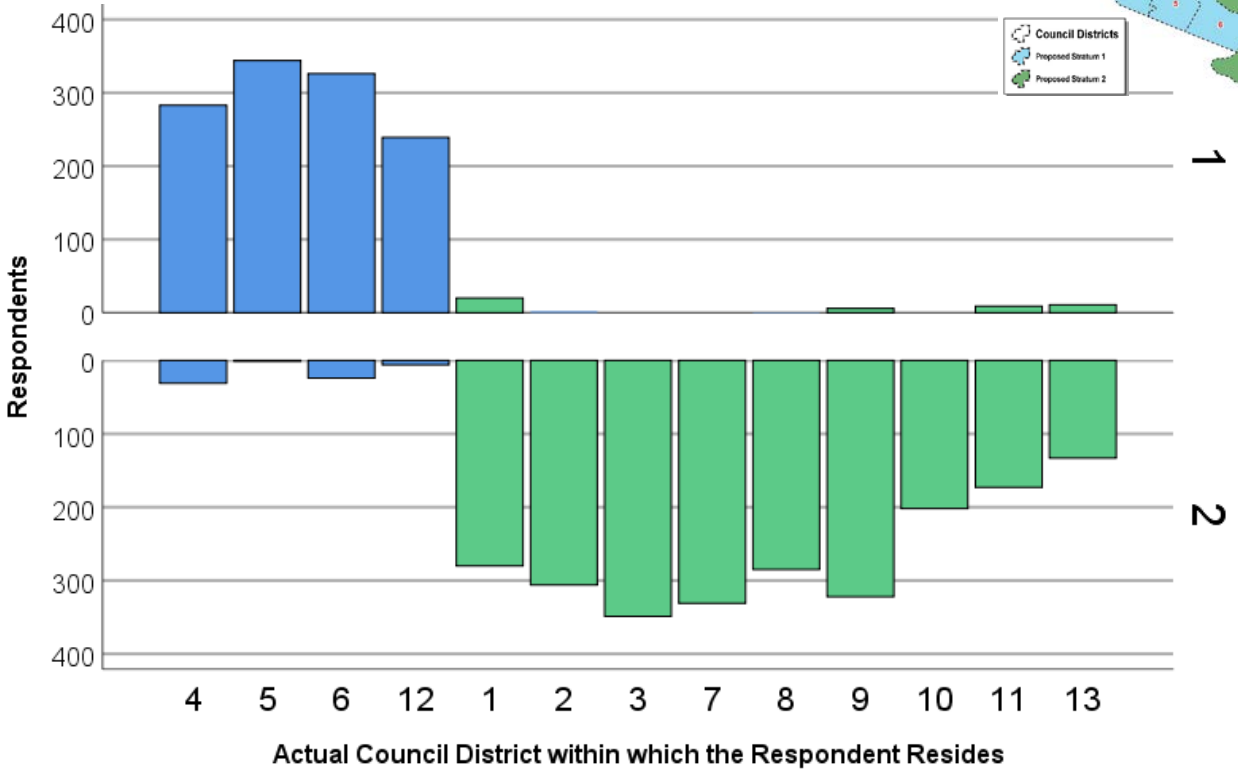
Stratum	Sub-stratum	Landline Universe	Sample Size	Respondents
1	Listed	122,858	9,277	1,155 (12%)
	Unlisted	288,593	14,543	97 (1%)
	Overall	411,450	23,820	1,252 (5%)
2	Listed	292,867	23,660	2,217 (9%)
	Unlisted	733,151	39,310	248 (1%)
	Overall	1,026,018	62,970	2,465 (4%)
Cell	N/A	1,892,417	79,200	5,315 (7%)

Overall Sample Yield by Assigned Stratum

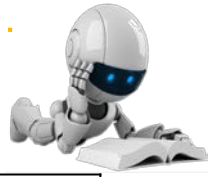


Geographic Coverage of Strata





Accuracy/ Coverage of Strata



Overall Accuracy based on 3,684 Respondents on LL Frame		Actual Stratum	
		1	2
Assigned to Stratum	1	1192	49
	2	62	2381

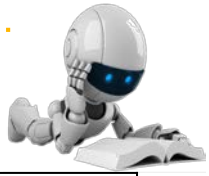
P(Assign to Stratum 1 | Stratum 1 Resident) = 95.1%

P(Assign to Stratum 2 | Stratum 2 Resident) = 97.8%

Overall Accuracy: Assigned vs. Actual Stratum = 97.0%

UMASS BOSTON

Accuracy/ Coverage of Strata: Listed



Overall Accuracy based on 3,341 Respondents on LL Frame from Listed Sub-Strata		Actual Stratum	
		1	2
Assigned to Stratum	1	1103	42
	2	54	2142

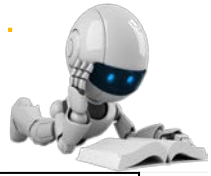
P(Assign to Stratum 1 | Stratum 1 Resident) = 95.3%

P(Assign to Stratum 2 | Stratum 2 Resident) = 98.1%

Overall Accuracy: Assigned vs. Actual Stratum = 97.1%

UMASS BOSTON

Accuracy/ Coverage of Strata: Unlisted



Overall Accuracy based on 3,341 Respondents on LL Frame from Listed Sub-Strata		Actual Stratum	
		1	2
Assigned to Stratum	1	89	7
	2	8	239

P(Assign to Stratum 1 | Stratum 1 Resident) = 91.8%

P(Assign to Stratum 2 | Stratum 2 Resident) = 97.2%

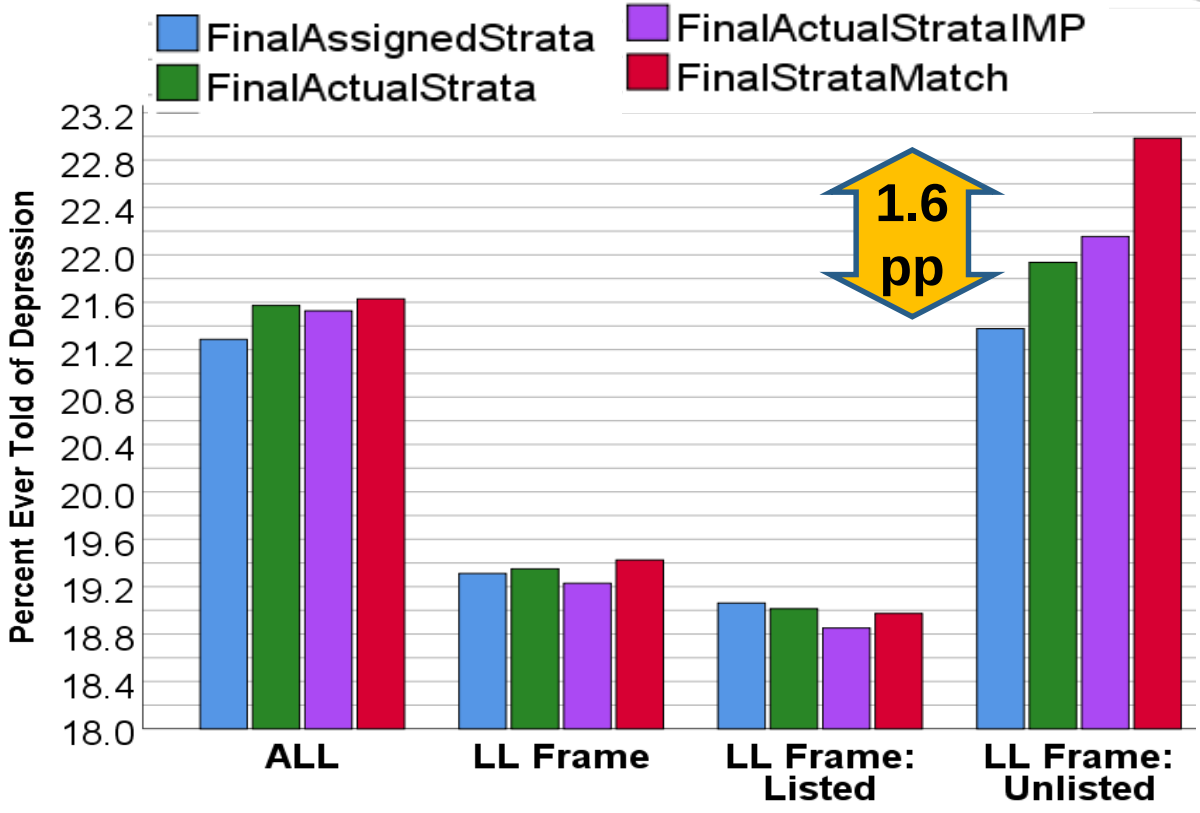
Overall Accuracy: Assigned vs. Actual Stratum = 95.6%

UMASS BOSTON

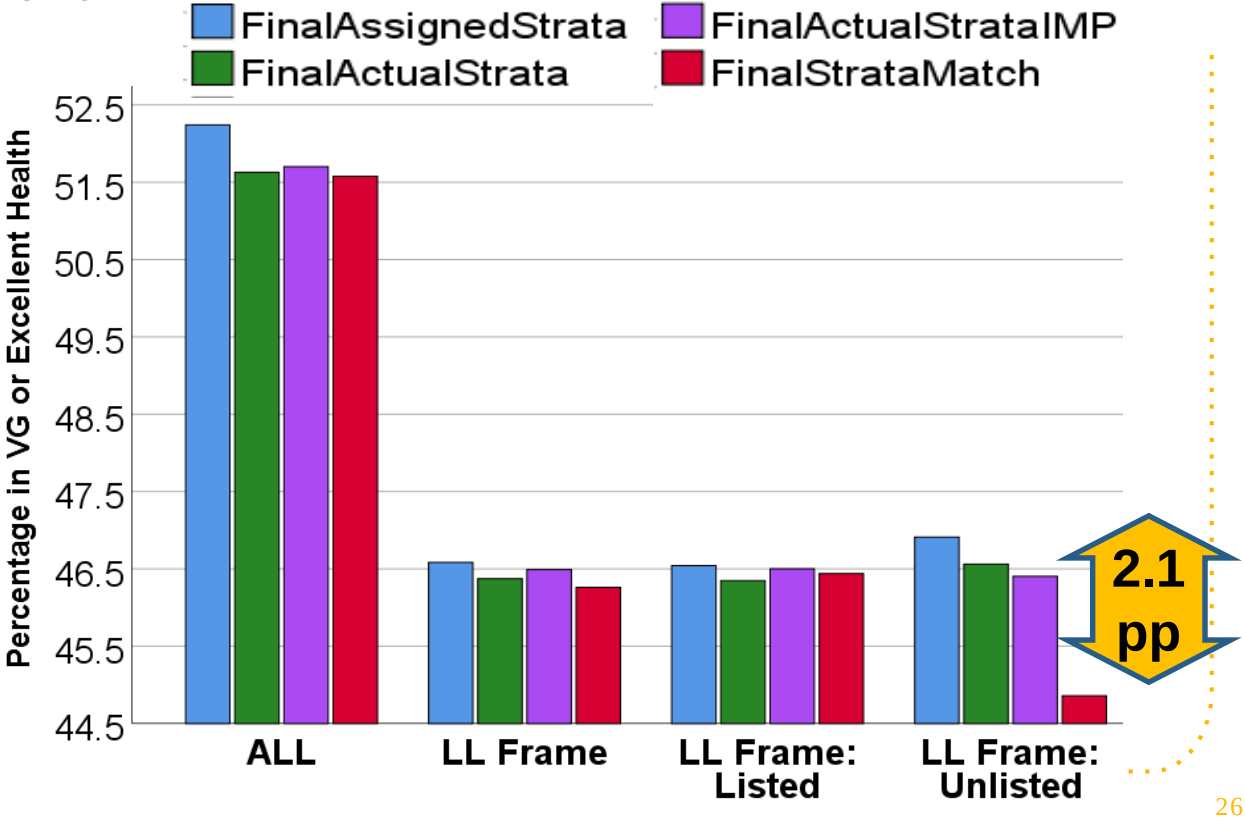
Implications

What about misclassification and survey estimates?
Percentage Ever told of Depression
Percentage in Very Good or Excellent Health
Mean BMI

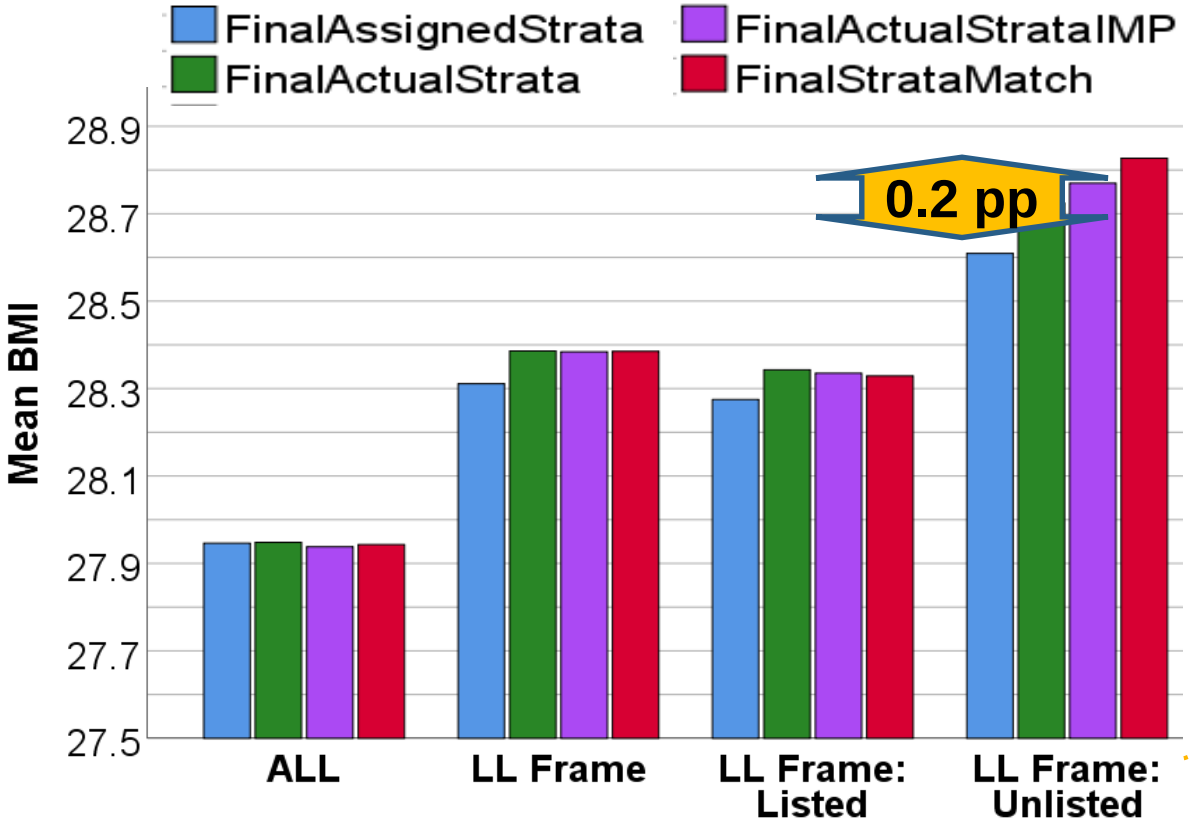
Percentage Ever Told of Depression



Percentage in Very Good/ Excellent Health



Average Body Mass Index (BMI)



5

Discussion and Future Research

Conclusions, Limitations and Next Steps…

Discussion

- The number of clusters that are deemed optimal using the k-means algorithm and criterion may not be optimal for the number of sampling strata for theoretical/ estimation and practical/ field concerns.
- The variables available for clustering may not result in strata that are efficient from a sampling design perspective – i.e. being close on the clustering variables may not imply homogeneity in strata on outcomes of interest, especially if multiple clusters need to be combined.

UMASS BOSTON

Limitations/ Future Work

- Using simple GIS information and landline listed number density we were able to successfully cover subsets of council districts whose geographies don't conform to 1K bank boundaries.
  - Unfortunately, we were not able to separate or create RDD sampling frames that could cover each council district or smaller groups of them.
  - We were also not able to apply this approach to the cellular frame because of limited geographical auxiliary information. Consumer cell sources with GIS information are rapidly improving so this might be possible in the future.

UMASS BOSTON

Thank You!

tdbuskirk@gmail.com

+1-781-964-4997

#trentbuskirk

www.linkedin.com/in/tbuskirk/