cercit
Comparative Effectiveness Research on Cancer in Texas

Data Management of Linkage of Texas Cancer Registry with other Sources

Yong-Fang Kuo PhD
Associate Professor/Director of Biostatistics
Department of Internal Medicine
Sealy Center on Aging
University of Texas Medical Branch

Sources of other data

1. Geography data – Covered by workshop 2

2. Survey data – only available for SEER

3. Provider data – Covered by workshop 6

Geography data

Census – track, zip code, county, state

Dartmouth Atlas – primary care services areas (PCSA), hospital services areas (HSA), hospital referral regions (HRR)

Area resource files – county, state

Survey data

Medicare health outcome survey (MHOS): provides information about the health-related quality of life (HRQOL) of Medicare Advantage Organization (MAO) enrollees.

National Longitudinal Mortality Study (NLMS): provides additional opportunities for analyzing socioeconomic and demographic differentials in cancer incidence, survival, and tumor characteristics.

Provider data

1. Physician – American Medical Association (AMA) Masterfile

2. Hospital - the Healthcare Cost Report (HCRIS) and the Provider of Service (POS) survey obtained from the Center for Medicare and Medicaid Services (CMS)

AMA

1. It contains information on all physicians in the U.S. including those that are not members of the AMA.
2. A record is created for each student upon entry into a U.S. medical school. International Medical Graduates are incorporated upon entry-into a residency program accredited by the Accreditation Council for Graduate Medical Education (ACGME) or when they obtain a license from one of the 54 state licensing boards.

Available information includes:

1. Demographics: UPIN/NPI, date of birth, gender, race/ethnicity (self-report), mailing and office address, alive/dead status.
2. Medical Education: medical school, date graduated, graduate training with dates (current-only) and state licensing with dates.
3. Specialty: Board certification (current), primary /secondary specialty (self-report).
4. Hospital: (self-report/current-only): hospital identifier, percent hours at hospital.
5. Practice: (self-report): group practice identifier (current-only), type of practice, employment.

Baldwin LM et al *Medical Care* 2002;40 [supplement]: IV82 – IV95

Issues in using AMA data

1. Provider ID ??
   UPIN ~ 2007, NPI 2007 ~
2. Race/ethnicity ??
3. Missing data:
   Office address: in 2002 paper, 31% missing, 5% 1995-2007 data, 13% missing
   Type of practice: missing 7% vs. 2.2%
   including medical teaching and research
   Present employment: 28% vs. 14.6%
   44.3% - group practice; however all group ID missing
4. Specialty: 30% with 2nd specialty reported
5. Board certification
6. Current vs. Historical data

TABLE 5. Level of Agreement in Specialty Types between 1994 Medicare Physician/Supplier Claims for Washington State and 1995 AMA Masterfile*

Specialty Type	No. of Physicians Identified as this Specialty Type in Physician/Supplier Claims	% Physicians Whose AMA Masterfile Primary Specialty Type Agreed with HCFA Specialty Type	No. of Additional Physicians Identified by AMA as this Specialty Type
Medical specialties			
Hematology/oncology	58	81.0	97
Cardiology	170	85.3	65
Gastroenterology	104	92.3	33
Nephrology	37	70.3	22
Pulmonology	75	76.0	42
Surgery specialties			
Thoracic surgery	48	85.4	6
Colorectal surgery	13	76.9	3
Urology	149	96.6	6
Obstetrics/gynecology	415	95.9	29
Miscellaneous			
Radiation oncology	40	95.0	21
General surgery	332	73.2	46
Family medicine	1386	90.8	101
Internal medicine	1121	55.8	143

*This table includes only those physicians with UPINs from the physician/supplier claims that could be linked with the AMA Masterfile.

1. SEER-Medicare 1992-2002 prostate cancer patients (4.8% unlink):

   96.3% of urologists (n=4,643) based on Medicare claims had primary specialty code of urologist in AMA.

   95.9% of urologists based on AMA had specialty code of urologist in Medicare claims

2. 100% Texas 2000-2007 Medicare beneficiaries aged 66-79 who had a PCP in 2006 (1.8% unlink)

   92% of PCP (IM, FP, GP, geriatricians, n= 16,126) based on Medicare claims had primary specialty code for PCP.

3. 2006 5% Medicare data

The specialty on the Medicare claims is specific to the practice setting through which the service was billed.

Medicare claims can provide more than one specialty for an individual physician, both could be accurate.

In 2006 5% data, 8.2% of general internal medicine physicians had more than one specialty code listed in Medicare claims.

Out of 13,466 hospitalists (general internal medicine physician with >=5 E&M and 90% of E&M from inpatient services) in 2006, 4.4% of them had 2nd specialty code. Majority of them had specialty on pulmonology.

Board certification:

1. 70.3% with board certification
2. 97% of physicians had board certification correspondent to their specialty

Specialty Name	Total count	Consistent count	consistency %
Cardiovascular Disease	19570	17701	90.45
Emergency Medicine	22212	19815	89.21
Endocrinology	3844	3391	88.22
Family Practice	54451	53424	98.11
Gastroenterology	10828	10041	92.73
General Surgery	19034	18656	98.01
Hematology	1749	1518	86.79
Hematology/oncology	3424	3214	93.87
Infectious Diseases	4446	3865	86.93
Internal Medicine	66950	64784	96.76
Internal Medicine - Geriatrics	2087	2000	95.83
Nephrology	6214	5618	90.41
Neurology	9640	9454	98.07
Neuroradiology	1567	1561	99.62
Obstetrics & Gynecology	24405	24301	99.57
Oncology	4750	4342	91.41
Pediatrics	16666	16321	97.93
Radiation Oncology	3901	3875	99.33
Radiology	5878	5799	98.66
Rheumatology	3709	3361	90.62
Urological Surgery	9709	9696	99.87

Suggestions

1. AMA masterfile is the best data source on board certification.
2. Claims are the best data sources for functional definitions of physician roles.
3. A combination of specialty in both source of data may be needed for the studied of subspecialty.

Volume

Measures: the number of procedures, patients seen, or claims submitted

Issues:

Specific procedures or related procedures

Procedures – for reasons other than cancer

Non-Medicare enrollees – age, HMO penetration

Regions – SEER, boundary

	Radical Prostatectomy		Mastectomy		Breast Conserving Surgery		Colectomy	
	No. of Physicians	%	No. of Physicians	%	No. of Physicians	%	No. of Physicians	%
Relationship of volume quintile from S-M to M								
S-M 4 quintiles < M	12	1.9	77	5.0	67	4.1	69	3.7
S-M 3 quintiles < M	13	2.1	41	2.7	70	4.3	60	3.3
S-M 2 quintiles < M	15	2.4	47	3.1	96	5.9	86	4.7
S-M 1 quintile < M	75	11.9	140	9.2	310	19.0	337	18.3
S-M = M quintile	473	75.1	882	57.7	848	52.0	1015	55.1
S-M 1 quintile > M	42	6.7	340	22.2	239	14.7	271	14.7
S-M 2 quintiles > M	0	0.0	2	0.1	0	0.0	3	0.2
Total	630	100.1	1529	100.0	1630	100.0	1841	100.0
Percentage of M procedures that are captured in S-M								
0–49%	79	12.5	238	15.6	598	36.7	651	35.4
50–74%	53	8.4	153	10.0	552	33.9	699	38.0
75–89%	101	16.0	188	12.3	191	11.7	218	11.8
≥90%	397	63.0	950	62.1	289	17.7	273	14.8
Total	630	99.9	1529	100.0	1630	100.0	1841	100.0
Procedures per physician								
	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
S-M	6.9 (7.6)	0–62	4.6 (4.6)	0–35	5.6 (7.3)	0–116	7.0 (7.8)	0–81
M	8.8 (9.1)	0–80	6.0 (5.2)	0–36	11.1 (12.1)	0–142	13.5 (12.3)	0–86

S-M = SEER-Medicare; M = Medicare; SD = standard deviation.

Hospital data

HCRIS: Healthcare Provider Cost Reporting Information System

Only hospitals which submit cot report electronically are in the HCRIS.

Annul filling

Take up to 18-24 months for data available in HCRIS

contains provider information such as facility characteristics, utilization data, cost and charges by cost center (in total and for Medicare), Medicare settlement data, and financial statement data.

POS: Provider of Service

The Provider of Services (POS) Extract is created from the Online Survey and Certification Reporting System (OSCAR) database.

File contains an individual record for each Medicare-approved provider and is updated quarterly.

These data include provider number, name, and address and characterize the participating institutional providers.

There are differences between the HCRIS and POS files as to exact time periods and how selected variables are defined.

Medical school affiliation (POS): 8.9% missing, 7.1% major, 9.7 minor (limited/graduate),

Teaching (HCRIS): 47% missing, 20.9% Yes

Ownership: 49.5% for profit (POS), 26.3% for profit (HCRIS)

# of beds in different units, personnel (POS, HCRIS)

type of service provided (POS), # hospital days, # of discharge by Medicare/Medicaid (HCRIS)

Volume: Age distribution, HMO penetration, outside the region

Nationwide Inpatient Sample (INS): 20% of disharges in 1056 hospitals in 42 states

Texas discharge data

Cystectomy- SEER-Medicare data (adjusted odds ratio, low vs. high volume 1.41, 95% confidence interval, 0.89-2.23) 100% Medicare data (odds ratio, 1.82; 95% confidence interval, 1.17 to 2.84)

Hollenback BK et al *Surgical Innovation* 2007;14(3): 192 –198

AHA: American Hospital Association annual survey, detailed and comprehensive information about hospital facilities, resources, staffing, utilization, and finances. About 1000 variables

Voluntary survey, response rate ~85%, missing data is somewhat high in some items

Due to the categorization of each item are different between HCRIS and AHA, it is hard to compare.

Ownership: 96% agree

Beds: 80% agree

# of discharge: 91% agree

Schrag D et al *Medical Care* 2002;40 [supplement]: IV96 – IV103

Multilevel (hierarchical) data structures

- Naive regression:

  violate the assumption – non-independent of error terms

- Fixed effects regression

  implies that statistical inference only takes sampling error at the patient level, not sampling error at the physician/hospital level

- Regression of summary measures

  ecological fallacy: assumes that patients cared for by a physician have the *average* characteristics. Ignore the variation across patients within a physician.

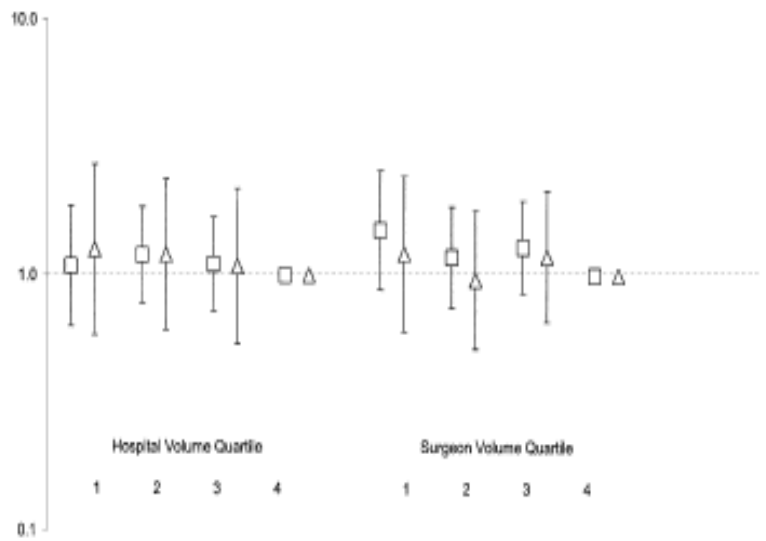
- Multilevel analyses

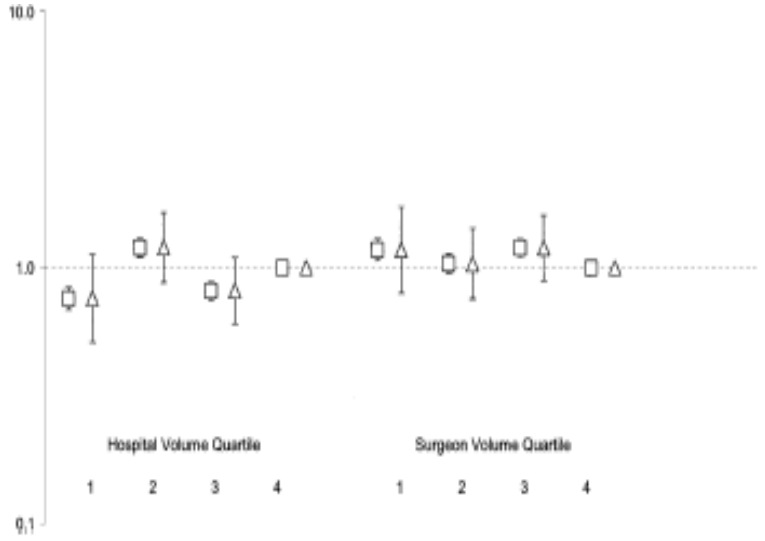
  incorporate hierarchical structure explicitly. Models referred to sometimes as mixed models, random coefficient models, and hierarchical models.

Volume – outcome research

Using native regression models

- Underestimation of the standard errors of regression coefficients and therefore an overestimation of the statistical significance
- assign provider characteristics to the patient level, implicitly treating characteristics of the provider as though they were characteristics of the patient. This results in artificially inflating the amount of real information that is available about the effect of provider characteristics on patient outcomes.
- heterogeneity of effects is not explicitly modeled , which may bias estimates of the regression coefficients





Urbach DR et al *J Clin Epidemiol* 2005; 58: 391 –400

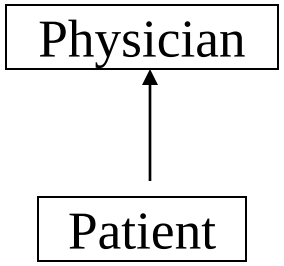
Results of multilevel and traditional analysis of TVSFP data

	Method				
	Multilevel	Naive	Fixed effects	Summary measures weighting by class size	Summary measures unweighted analysis
Model without pretest THKS					
$\hat{\beta}_1(\hat{SE}(\hat{\beta}_1))$	0.056 (0.070)	0.089 (0.047)	0.089 (0.045)	0.089 (0.067)	−0.041 (0.082)
$t_{\beta 1}$ (*df*)	0.8011 (68)	1.876 (835)	1.964 (767)	1.331 (68)	−0.498 (68)
2-tailed *P*-value	0.426	0.061	0.050	0.188	0.620
Model with pretest THKS					
$\hat{\beta}_1(\hat{SE}(\hat{\beta}_1))$	0.085 (0.061)	0.106 (0.045)	0.089 (0.043)	0.106 (0.060)	−0.018 (0.079)
$t_{\beta 1}$ (*df*)	1.340 (66)	2.369 (833)	2.057 (766)	1.727 (67)	−0.229 (67)
2-tailed *P*-value	0.168	0.018	0.040	0.082	0.819

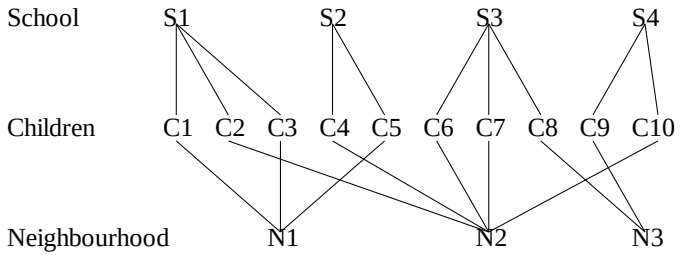
Moerbeek M et al *J Clin Epidemiol* 2003; 56: 341 –350

Classifying Structures

Simple hierarchy



Cross classifications



Multiple membership

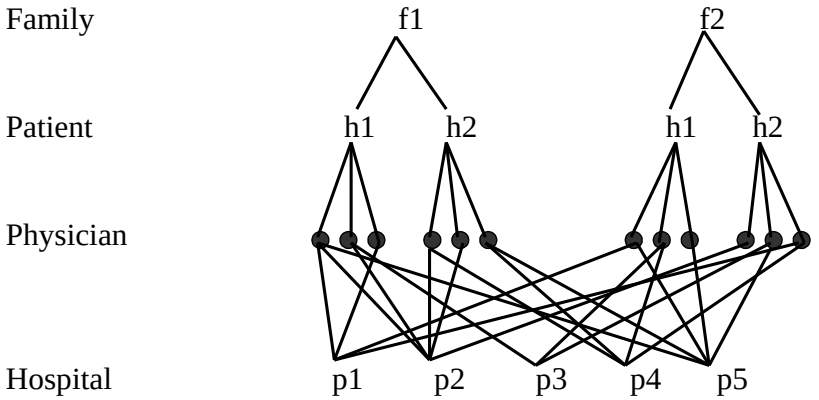


Table 3. Percentage of variance in use of androgen deprivation therapy attributable to the urologist from multilevel analyses

Characteristic	All patients			Evidence-based group (1997–1999)	Uncertain-benefit group (1997–1999)
	1992–1999	1992–1996	1997–1999		
No. of patients	61717	42425	19292	2329	16963
No. of urologists	1802	1577	1258	789	1234
Residual ICC*,†, %	20.63	20.38	26.55	29.09	25.48
Partitioned variance*,‡, %					
Stage and grade	13.06	15.40	9.71	6.63	5.34
Patient characteristics	8.99	6.58	4.29	7.27	4.99
Urologist	16.00	15.60	22.56	25.38	22.68

*Hierarchical generalized linear model with patient age; comorbidity; ethnicity; Surveillance, Epidemiology, and End Results (SEER) region; tumor stage; grade; year of diagnosis; census tract education; and census tract poverty entered as "level 1" variables and physician identifiers entered as "level 2" variables. ICC = intraclass correlation coefficient.

†The percentage of variance attributable to the urologist calculated with a threshold model, after simultaneous adjustment of all available patient and tumor characteristics. The denominator for the calculation of the percentage was composed of the variance attributable to the urologist, after adjustment for available patient and tumor characteristics, and the variance attributable to unexplained patient or tumor variables plus error.

‡The variance was further partitioned using a threshold model so that the percentages of total variance contributed by the urologist, as well as those contributed by patient and tumor variables, are presented. Results are presented as the percentage of total variance attributable to the indicated characteristic. The denominator is total variance, which is composed of the variance attributable to the urologist after adjustment for available patient and tumor characteristics, the variance attributable to available patient and tumor variables, and the variance attributable to unexplained patient or tumor variables plus error.

Shahinian VB et al *J Natl Cancer Inst* 2006;98: 839 – 45

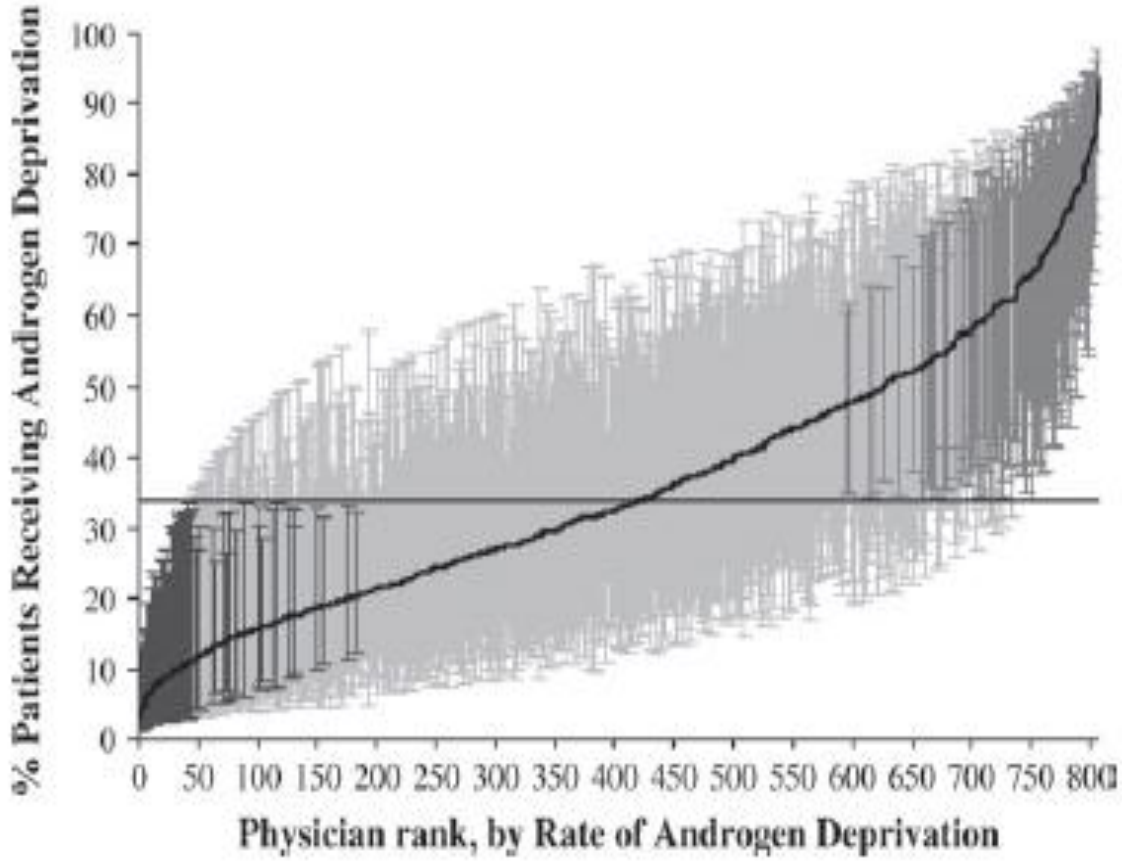

% Patients Receiving Androgen Deprivation
100
90
80
70
60
50
40
30
20
10
0
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800
Physician rank, by Rate of Androgen Deprivation

Table 2. Urologist Characteristics As Predictors of Androgen Deprivation Use in Multilevel Models

Urologist Characteristic	Adjusted ORs for Use of Androgen Deprivation*					
	Overall†		1992-1995‡		1996-2002§	
	OR	95% CI	OR	95% CI	OR	95% CI
Years since graduation						
Every 5 years	1.03	1.00 to 1.06	1.00	0.97 to 1.03	1.03	0.99 to 1.07
Board certification						
Yes	1.00		1.00		1.00	
No	1.26	0.99 to 1.60	1.16	0.88 to 1.52	1.36	1.01 to 1.83
Academic affiliation						
Major	1.00		1.00		1.00	
Minor	1.51	1.27 to 1.81	1.27	1.03 to 1.56	1.60	1.29 to 1.99
None	1.83	1.52 to 2.22	1.30	1.04 to 1.63	2.18	1.73 to 2.75
Panel size						
< 15	1.00		1.00		1.00	
15-59	1.26	1.07 to 1.48	1.10	0.89 to 1.36	1.39	1.13 to 1.71
60-119	1.41	1.18 to 1.67	1.12	0.90 to 1.39	1.69	1.36 to 2.10
≥ 120	1.44	1.19 to 1.76	1.13	0.89 to 1.43	1.76	1.38 to 2.25

Shahinian VB et al*J Clin Oncol 2007;25:5359-5365*

Overall

		Model 1			Model 2			Model 3*		
	Level	OR	(95% CI)		OR	(95% CI)		OR	(95% CI)	
Hospital Characteristics										
Total Bed	<350	1.00			1.00			1.00		
	>= 350	0.93	0.79	1.11	0.97	0.82	1.15	0.95	0.78	1.16
Cancer hospital	No	1.00			1.00			1.00		
	Yes	0.63	0.42	0.95	0.63	0.42	0.95	0.56	0.35	0.90
Teaching Hospital	Major	1.00			1.00			1.00		
	Limited/Graduate	1.17	0.92	1.49	1.21	0.95	1.53	1.21	0.92	1.60
	No affliation	1.32	1.07	1.62	1.38	1.11	1.70	1.40	1.10	1.80
Metropolitan Size	A, B	1.00			1.00			1.00		
	C, D, E, F	0.96	0.83	1.12	0.96	0.82	1.12	0.91	0.76	1.10
Urologist Characteristics										
Years Since Graduation	Every 5 years				0.97	0.94	0.99	1.00	0.97	1.03
Board Certification	Yes				1.00			1.00		
	No				1.37	1.17	1.62	1.14	0.94	1.37
Panel Size	< 15				1.00			1.00		
	15 - 49				1.10	0.95	1.27	1.17	0.99	1.38
	50 - 89				1.17	1.00	1.37	1.28	1.06	1.53
	>= 90				1.05	0.89	1.24	1.16	0.95	1.40

* Adjusted patient demographic and tumor characteristics (no SEER region)

** with SEER region

ICC	Null	model 1	model 2	model 3
Hospital	7.56%	6.75%	6.88%	9.19%
Urologist	9.53%	9.70%	9.25%	11.67%

Table 2. Association of having a primary care physician (PCP) and the racial disparity in CRC screening rates among Medicare enrollees in Texas*.

	Odds ratio	
	PCP not identified	PCP identified
Black vs. White	0.67 (0.66, 0.69)	0.83 (0.81, 0.85)
Hispanic vs. White	0.51 (0.49, 0.53)	0.57 (0.55, 0.59)

**This interaction model was adjusted for age, sex, comorbidity in the year before monitoring, and the high risk indicator for CRC of the year before monitoring.*

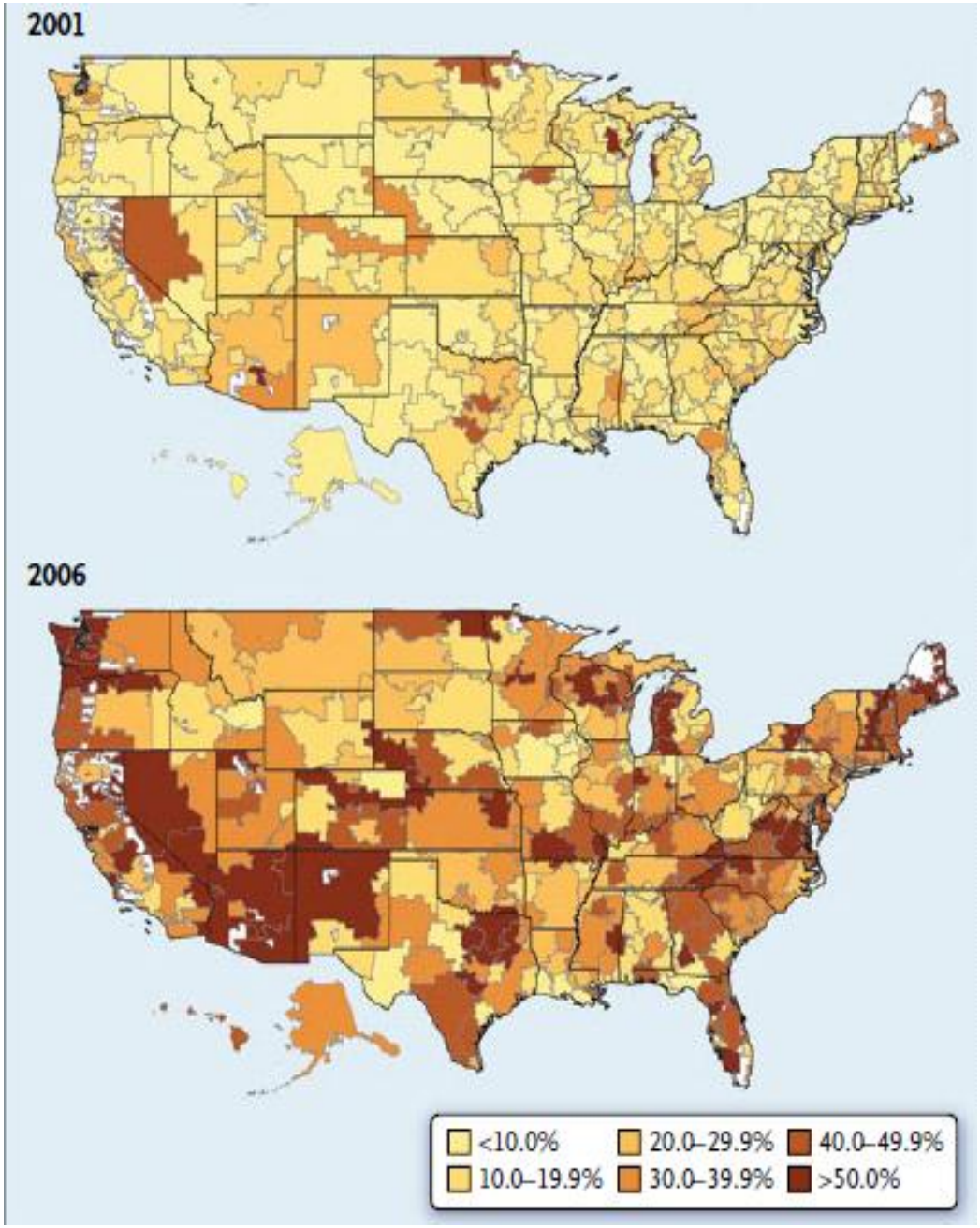
Table 4. Association of primary care physician (PCP) characteristics and the racial disparity in colorectal cancer (CRC) screening by multi-level analysis among Medicare enrollees in Texas

	Null Model	Adjusted model*
ICC of PCP (%)	7.50	5.76
		Odds Ratio
Race (Black vs. White)	-	0.98 (0.95, 1.01)
Race (Hispanic vs. White)	-	0.82 (0.80, 0.85)
PCP sex (Female vs. Male)	-	1.16 (1.12, 1.20)
UStrained (Yes vs. No)	-	1.16 (1.12, 1.20)
PCP age (≤50 vs. >50)	-	1.08 (1.05, 1.11)
PCP Specialty GP vs. IM	-	0.65 (0.59, 0.71)
FP vs. IM	-	0.72 (0.70, 0.74)
Geriatrics vs. IM	-	0.86 (0.73, 1.01)
% White patients in panel (each 1% increase)	-	1.01 (1.01, 1.01)

Singal AW et al. *Am J Public Health* under review

		Odds ratio*	
		Black vs. White[$]	Hispanic vs. White[$]
Overall colonoscopist availability	Q1 (0-4.8)	0.87 (0.84, 0.90)	0.73 (0.71, 0.76)
	Q2 (4.8-6.9)	0.82 (0.78, 0.85)	0.75 (0.72, 0.78)
	Q3 (6.9-8.5)	0.78 (0.76, 0.80)	0.52 (0.49, 0.55)
	Q4 (>8.5)	0.77 (0.74, 0.80)	0.64 (0.60, 0.68)
PCP availability	Q1 (0-34.1)	0.83 (0.80, 0.86)	0.76 (0.73, 0.80)
	Q2 (34.1-45.7)	0.81 (0.77, 0.84)	0.76 (0.73, 0.79)
	Q3 (45.7-65.5)	0.83 (0.81, 0.86)	0.59 (0.56, 0.62)
	Q4 (>65.5)	0.77 (0.75, 0.79)	0.51 (0.48, 0.54)

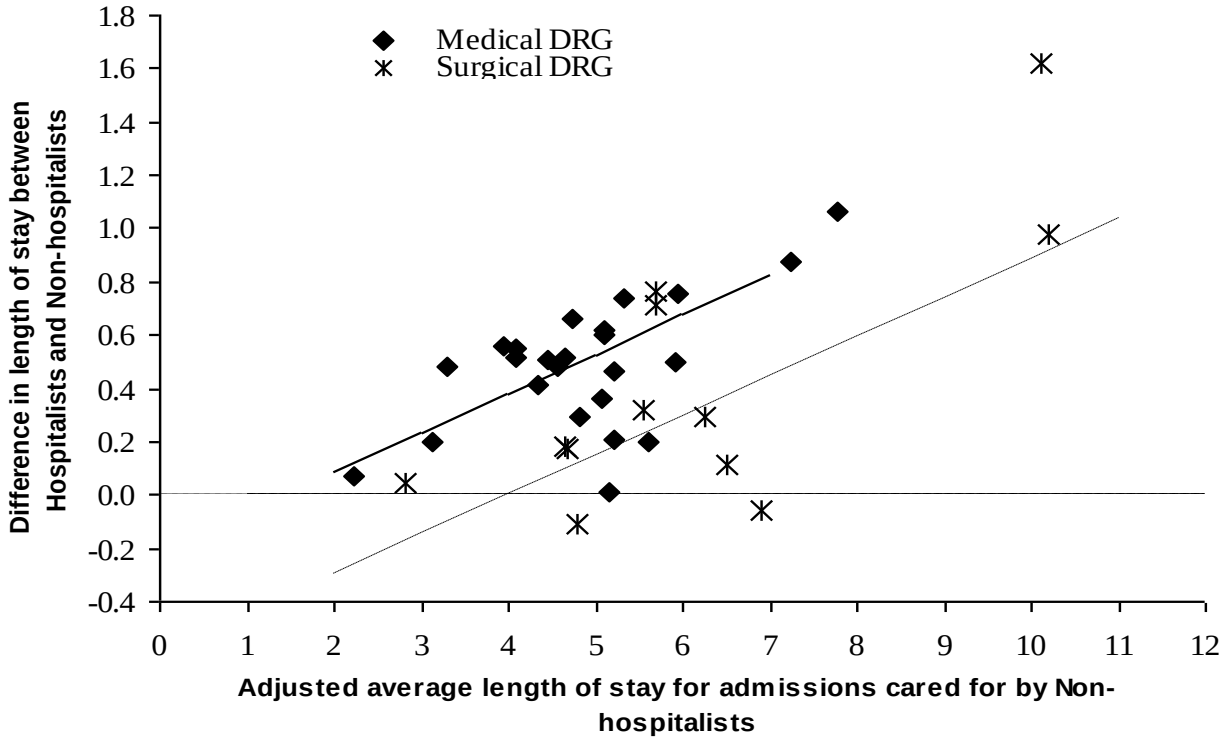
Benarroch-Gampel J. et al *Medical Care* under review



Kuo YF et al *N Engl J Med* 2009;360:1102-12.

Association of hospitalist care with medical utilization after discharge

- Patients cared for by hospitalists were less likely to have an identified primary care physician (PCP) prior to admission. Patients with an established PCP differ from those without, in that they are younger and have fewer comorbidities.



Kuo YF et al *J Am Geriatri* 2010;58:1649-57.

Selection Bias

Select a more homogeneous study cohort

patients with medical DRG and with an identified PCP
205,190 admissions at 4657 hospitals

Where the differences between treatments exist? Hospitals

What is the study focus? Within hospital differences

Select hospitals with at least 20 admissions cared for by hospitalists and at least 20 admissions cared for by PCP

58,125 admissions at 454 hospitals

We generated the propensity that an admission would be cared for by a hospitalist from a logistic regression model for each hospital.

Conditional logistic regression models, Generalized linear models with various distributions and including hospital as a fixed covariate

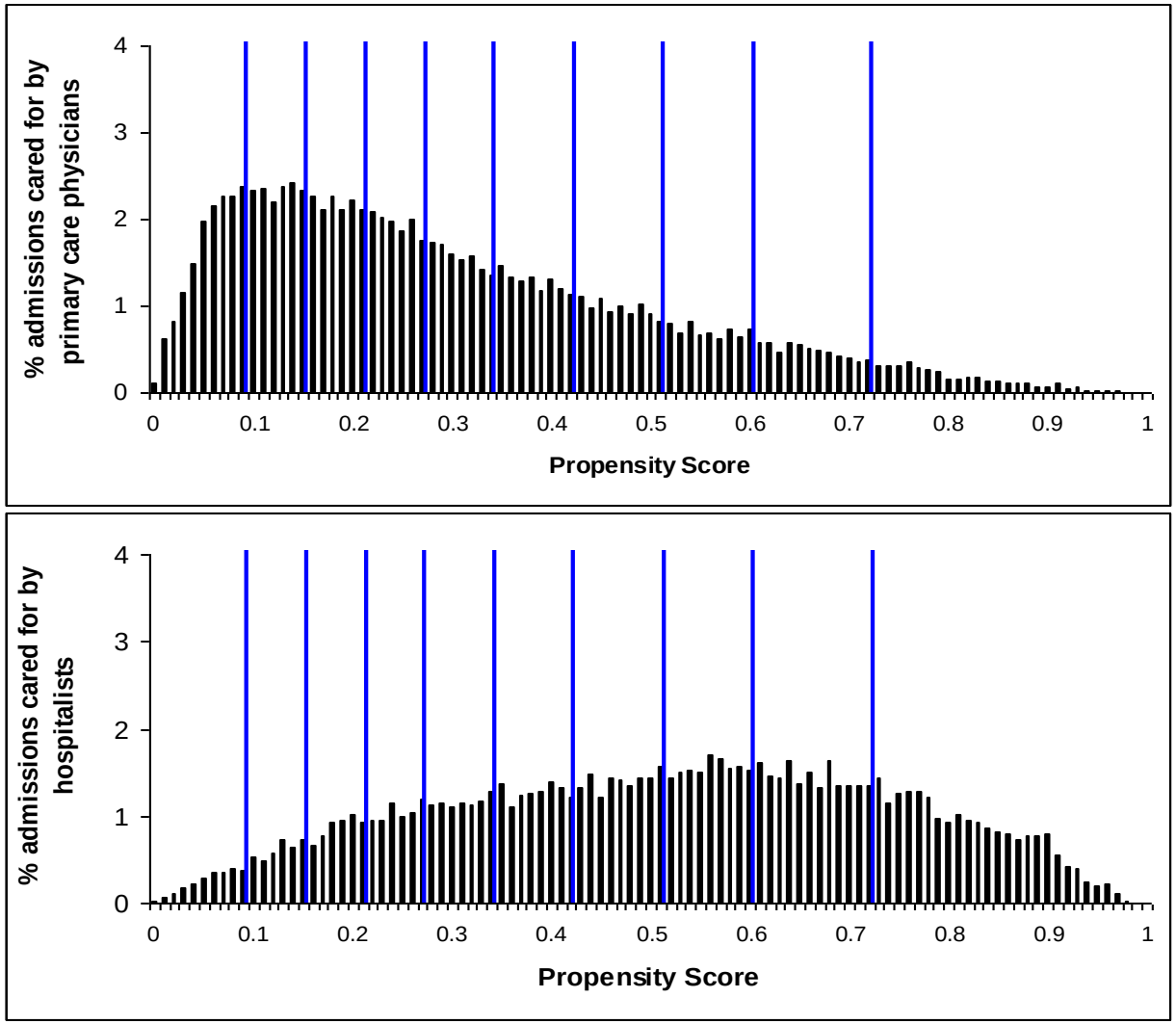

% admissions cared for by primary care physicians
4
3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Propensity Score
% admissions cared for by hospitalists
4
3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Propensity Score

Table 3: Propensity analyses estimating mean length of stay, hospital charges, and Medicare costs in the 30 days after discharge for admissions cared for by hospitalists or by their primary care physicians*

		PCP cohort	Hospitalist cohort	Difference
Average length of stay (95% CI)		5.82 (5.76 – 5.87)	5.17 (5.11 – 5.23)	-0.64 (-0.56 – -0.72) §
Average hospital charges (95% CI) †		$15301 (15188 – 15416)	$15019 (14884 – 15155)	-$282 (-115 – -451) §
Medicare costs in the 30 days post discharge†‡				
Professional Services	Percent of patients with charges	93.0%	91.8%	
	Average cost	$643	$690	
	Combined model	$598 (579 – 616)	$633 (606 – 662)	$35 (12 – 59) §
Outpatient facilities	Percent of patients with charges	37.3%	41.0%	
	Average cost	$446	$435	
	Combined model	$166 (160 – 173)	$178 (170 – 187)	$12 (3 – 21) §
Hospitalization	Percent of patients with charges	17.2%	18.0%	
	Average cost	$9008	$9412	
	Combined model	$1548 (1498 - 1601)	$1694 (1623 – 1769)	$146 (71 – 218) §
Skilled Nursing home	Percent of patients with charges	6.3%	6.6%	
	Average cost	$7755	$8147	
	Combined model	$487 (458 – 518)	$535 (494 – 579)	$48 (9 – 85)
Other facilities	Percent of patients with charges	1.0%	1.2%	
	Average cost	$12643	$11687	
	Combined model	$132 (117 – 154)	$140 (112 – 161)	$8 (-11– 25)
Total costs	Percent of patients with charges	93.6%	93.0%	
	Average cost	$3149	$3523	
	Combined model	$ 2947 (2834 - 3063)	$3279 (3106 – 3459)	$332 (160 – 486) §‖

Kuo YF et al *Ann Intern Med*. Provisional Accept

Additional analyses:

Effect of hospitalist care within each decile of propensity

Test interaction between patient/hospital characteristics and hospitalist care

Sensitivity analyses

Marginal models for entire study population

Random effect model with three level structure (hospitals nested within HRR)

Instrumental variables analyses

distance from patients' home to hospital, and % of admission cared for hospitalist in HRR ???

PCP preference in use of hospitalist care, distance from PCP's office to hospital

Takeaway Points

Understand what are available in claims.

ID (perform, refer), zip code, specialty, place of service, type of facilities

Use of supplement databases.

1. Correctly specify data structures avoid overestimate of outcomes
2. Understand between-unit variation
3. Locate significant factors operated at different levels
4. Finding instrumental variables with a good story

Available geographic and provider identifiers

Medicaid:

NPI will be included in MAX2009. Specialty codes are specific by states. Some states do not have specialty codes.

Service provider ID are available– state specific

No provider details in MAX files.

Patient Zip code is availble, not provider.

ResDAC works with vendor to give the same ID for Medicare and Medicaid in the future.

Marketscan:

Physician, specialty, place of service, type of provider, and hospital D are available.

Patient, physician, and hospital with 3 digit zip code