Paper presented at the Royal Geographic Society with Institute of British Geographers (RGS-IBG) 2005 Annual Conference, London , 30 August

Data Mining of ethnic information in patient records through birthplace & name analysis. An example in Camden PCT

Mateos, P.; Jones, CE.; Longley, PA. and Webber, R.

Centre for Advanced Spatial Analysis
University College London
1-19 Torrington Place, London WC1E 6BT

p.mateos@ucl.ac.uk

Abstract:

Understanding the health needs of patients is essential to the NHS Public Health function, but traditionally, assessment at the individual level has been carried out using segmentation of patients by age and gender solely.

There is a growing public health requirement to understand ethnic diversity of local populations to reduce health inequalities, especially in Inner cities (in Inner London ethnic minorities represent 34.3% of population vs. 13% nationally). Yet the requisite data are often incomplete, inconsistent or out of date. The most reliable small area data hitherto used to ascribe ethnicity have been census data, with (as of 2001) the output area being the finest spatial resolution obtainable.

This paper applies data mining techniques to (poor quality) patient registration data in order to classify individuals into ethnic groups. These techniques use intelligent text searching to ascribe a person's country of origin and immigration date, where available. This is combined with the analysis of birth date and the origins of surname and forenames at the address level, using fuzzy logic to allocate a degree of belonging to a particular ethnic group. These data are then spatially aggregated to neighbourhoods for reporting purposes, and the potential value of the approach is assessed.

Keywords: Ethnicity Classification, Health Inequalities, Geospatial Data Mining, Geodemographics

Camden NHS Primary Care Trust

UCL

Data Mining of ethnic information in patient records through birthplace & name analysis.

An example in Camden PCT

Pablo Mateos
Catherine E. Jones
Paul Longley
Richard Webber

Centre for Advanced Spatial Analysis (CASA)
University College London

Royal Geographic Society Annual Conference 30th August 2005

Contents

1. Objective & justification
2. Country of birth analysis
3. Name ethnicity analysis
4. Household ethnicity analysis
5. Ethnicity model compilation & evaluation
6. Future enhancements

1- Objective and opportunity

- Objective:
  - Provide a method to measure the ethnic/cultural background of the population in Inner London
    - at the individual level
    - beyond the Census ethnic classification of 16 categories
- Opportunity:
  - Underutilized “Birth Place” field in the NHS patient register (Exeter)
  - UCL research on Surname & Forename analysis

Justification: Ethnic inequalities in health

- Growing evidence on differential health needs by ethnic group
- NHS is required by law to demonstrate equity of service
- London ethnic minorities represent 39% of total population (UK 12%)
- Lack of detailed data on ethnicity
- Immigration is a hot political issue

2001 Census 16+ classification

White	**91.3%**	
British		87.5%
Irish		1.2%
Other White		2.6%
Mixed	**1.3%**	
White & Black Caribbean		0.5%
White & Black African		0.2%
White & Asian		0.4%
Other Mixed		0.3%
Black or Black-British	**2.2%**	
Black-Caribbean		1.1%
Black-African		0.9%
Black-Other (please describe)		0.2%
Asian or Asian-British	**4.4%**	
Indian		2.0%
Pakistani		1.4%
Bangladeshi		0.5%
Any other Asian background		0.5%
Chinese or other group	**0.9%**	
Chinese		0.4%
Any other ethnic group		0.4%

Total Non- White British	**12.5%**
Poorly Studied Groups	**5.1%**

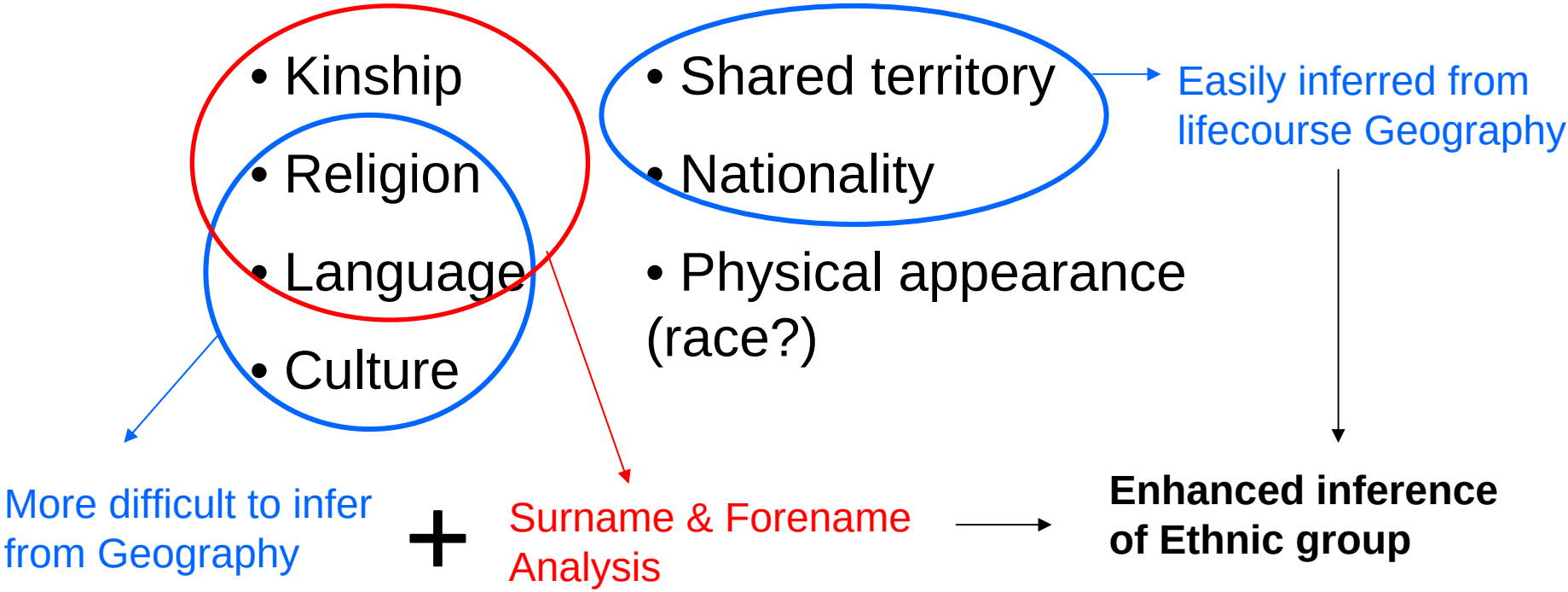
- Confusing question!
- Strongly based on a “skin colour problem”
- Represents and reproduces current crude stereotyping of ethnic minorities
- Best used in combination with Country of Birth and Religion

Source: ONS Census 2001 – Great Britain Population

UCL

From race to ethnic group: a new ontology of ethnicity

- **<u>Ethnicity</u>: A multi-dimensional concept that encompasses different aspects of identity:**



2 - Country of birth analysis

Underutilized patient birthplace field

NHS Exeter Fields

PATIENT REGISTRATION
GP_LOCAL_CODE
PERSON_TITLE
PERSON_FORENAME
PERSON_MIDDLENAMES
PERSON_SURNAME
PERSON_PREMISES
PERSON_STREET
POSTCODE
NEW_NHS_NUMBER
NHS_NUMBER
OLD_NHS_NUMBER
PERSON_DATE_OF_BIRTH
PERSON_SEX
PERSON_BIRTHPLACE
REGISTERED_DATE

Examples of “Birth Places”

PERSON_BIRTHPLACE
AFRICA
ALGIERS
ARABIA ARR 5.8.93
CAMDEN LONDON
EIRE ARR 9.10.03
ERITRIA-ASMARA
FRANCE ARR NO DATE
ISOMALIA
LUDSCHENDAM NETHERLAN
MOGDIO SOMALIA
MOGDISH.SOMALIA
MORROCCO ARR 20.9.88
PADDINGTON,LONDON
PALESTINIAN
RFH LONDON
S.AMERICA
SCOTLAND
SOUTH AMERCIA
ST MARY'S
SWANSEA
TOWER HAMLETS
UGANDA ARR 15.4.89
UNKNOWN
WEST INDIES
AT HOME

A tool has been developed to allocate a person’s Birth Place to a gecoded place within a formalized World Gazetteer

Birthplace Geocoder

PERSON_BIRTHPLACE	Continent	Subcontinent	Country, Territory,	UK_Country:	City	Local Authority	Arrival Date
AFRICA	Africa						
ALGIERS	Africa	North Africa	Algeria		Algiers		
ARABIA ARR 5.8.93	Asia	Middle East	Saudi Arabia				5/8/1993
CAMDEN LONDON	Europe	Western Europe	United Kingdom		London	Camden	
EIRE ARR 9.10.03	Europe	Western Europe	Ireland				9/10/2003
ERITRIA-ASMARA	Africa	Western Africa	Eritrea		Asmara		
FRANCE ARR NO DATE	Europe	Western Europe	France				
ISOMALIA	Africa	Western Africa	Somalia				
LUDSCHENDAM NETHERLAN	Europe	Western Europe	Netherlands		Ludschendam		
MOGDIO SOMALIA	Africa	Western Africa	Somalia		Mogadishu		
MOGDISH.SOMALIA	Africa	Western Africa	Somalia		Mogadishu		
MORROCCO ARR 20.9.88	Europe	Western Europe	Morocco				20/09/1988
PADDINGTON,LONDON	Europe	Western Europe	United Kingdom	England	London	WestMin	
PALESTINIAN	Asia	Middle East	Palestina				
RFH LONDON	Europe	Western Europe	United Kingdom	England	London	Camden	
S.AMERICA	America	South America					
SCOTLAND	Europe	Western Europe	United Kingdom	Scotland			
SOUTH AMERCIA	America	South America					
ST MARY'S	Europe	Western Europe	United Kingdom	England	London	WestMin	
SWANSEA	Europe	Western Europe	United Kingdom	Wales	Swansea		
TOWER HAMLETS	Europe	Western Europe	United Kingdom	England	London	TowerH	
UGANDA ARR 15.4.89	Africa	Western Africa	Uganda				15/4/1989
UNKNOWN	Delete						
WEST INDIES	America	Caribbean	British West Indies				
AT HOME	Delete						

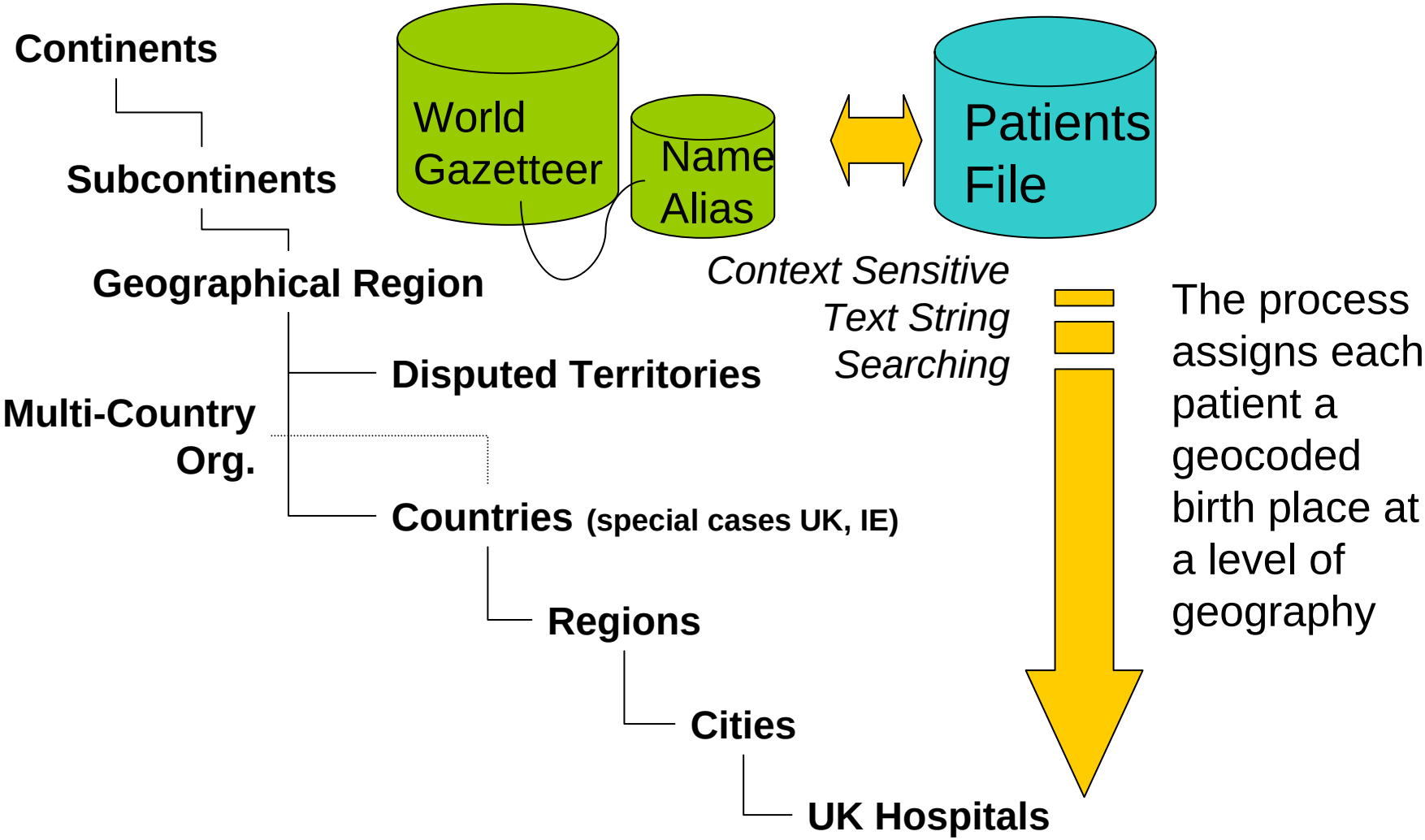
Building alias tables

COUNTRY_ALIAS	COUNTRY_CODE
BANGLEDESH	BD
BANLADESH	BD
BANGLADASH	BD
BANGALDESH	BD
BANGLDESH	BD
BANGLASDESH	BD
BANDLADESH	BD
BANGLEDASH	BD
BANGLADESHI	BD
BANDGLADESH	BD
BADGLADESH	BD
BAGLADESH	BD
BAMGLADESH	BD
BANGADESH	BD
BANGALADESH	BD
BANGDALESH	BD
BANGELDESH	BD
BANGILADESH	BD
BANGLA DESH	BD
BANGLAADESH	BD
BANGLADEESH	BD
BANGLADES	BD
BANGLA-DESH	BD
BANGLA - DESH	BD
BANGLADISH	BD
BANGLADSEH	BD
BANGLADSH	BD
BANGLAESH	BD
BANGLASESH	BD
BANGLIDESH	BD
BENGLEDESH	BD

31 ways to spell one country name!

The more records coded from other PCTs the more effective the name alias lookup tables will get

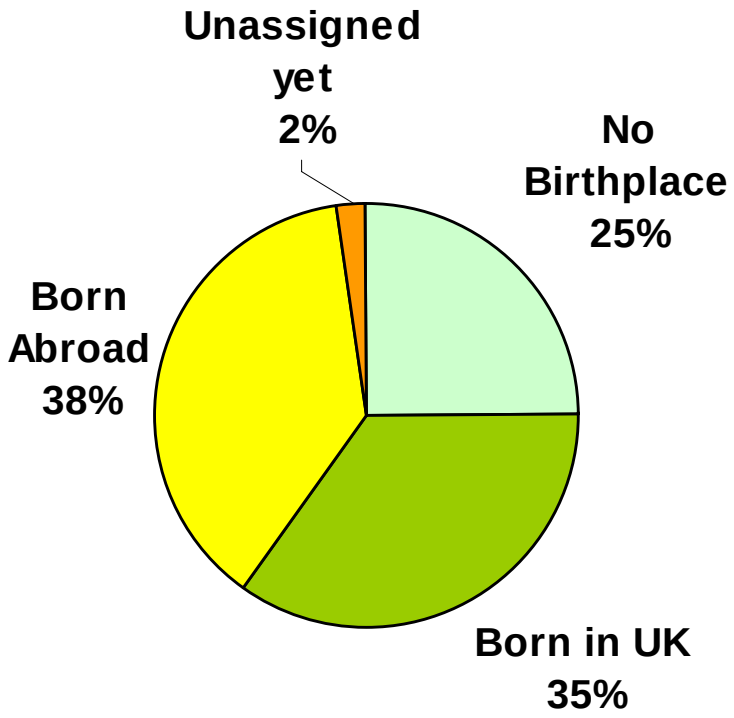
Birthplace Geocoder



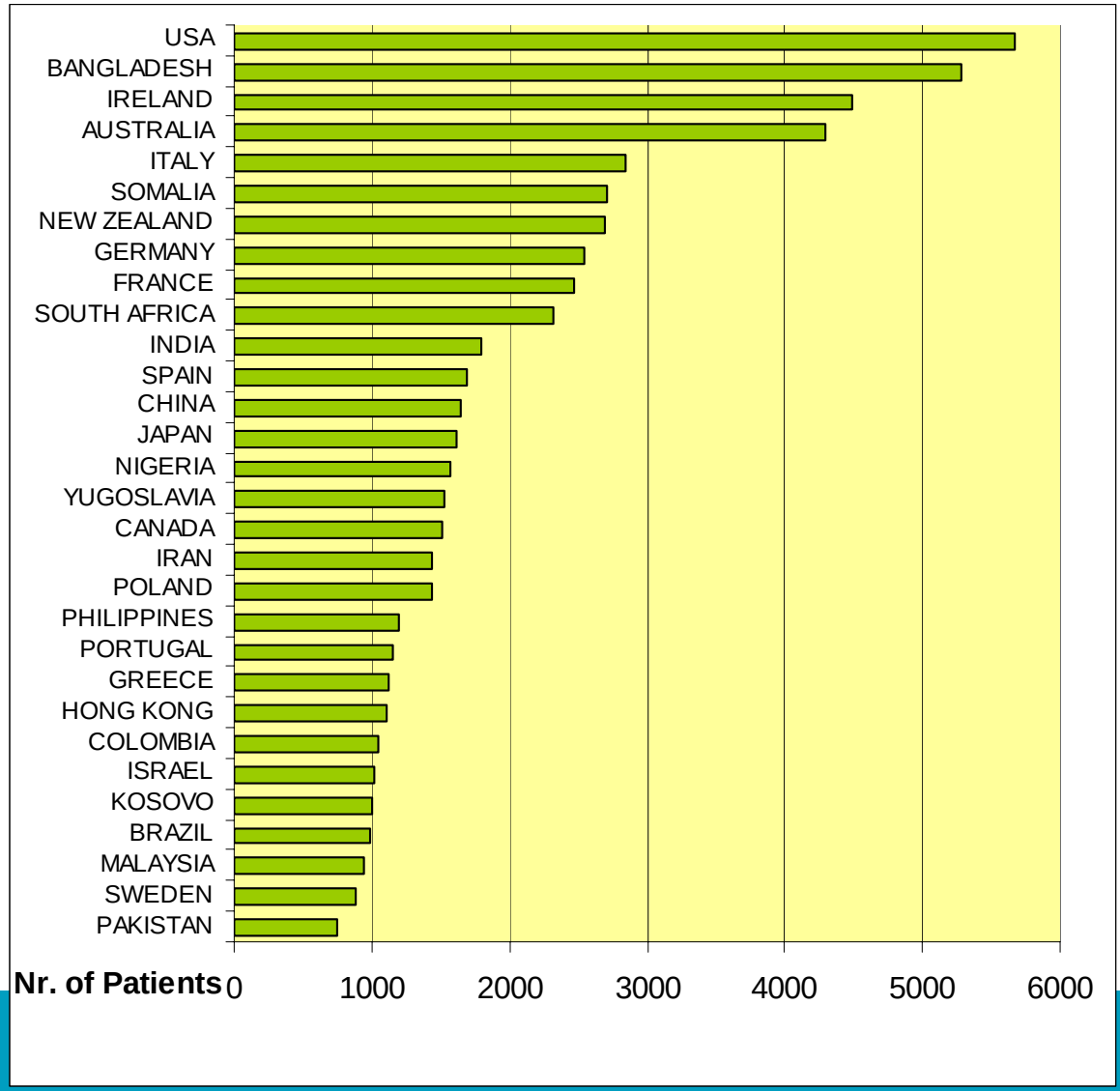
Camden PCT birthplaces

204,068 Patients (Oct 04)

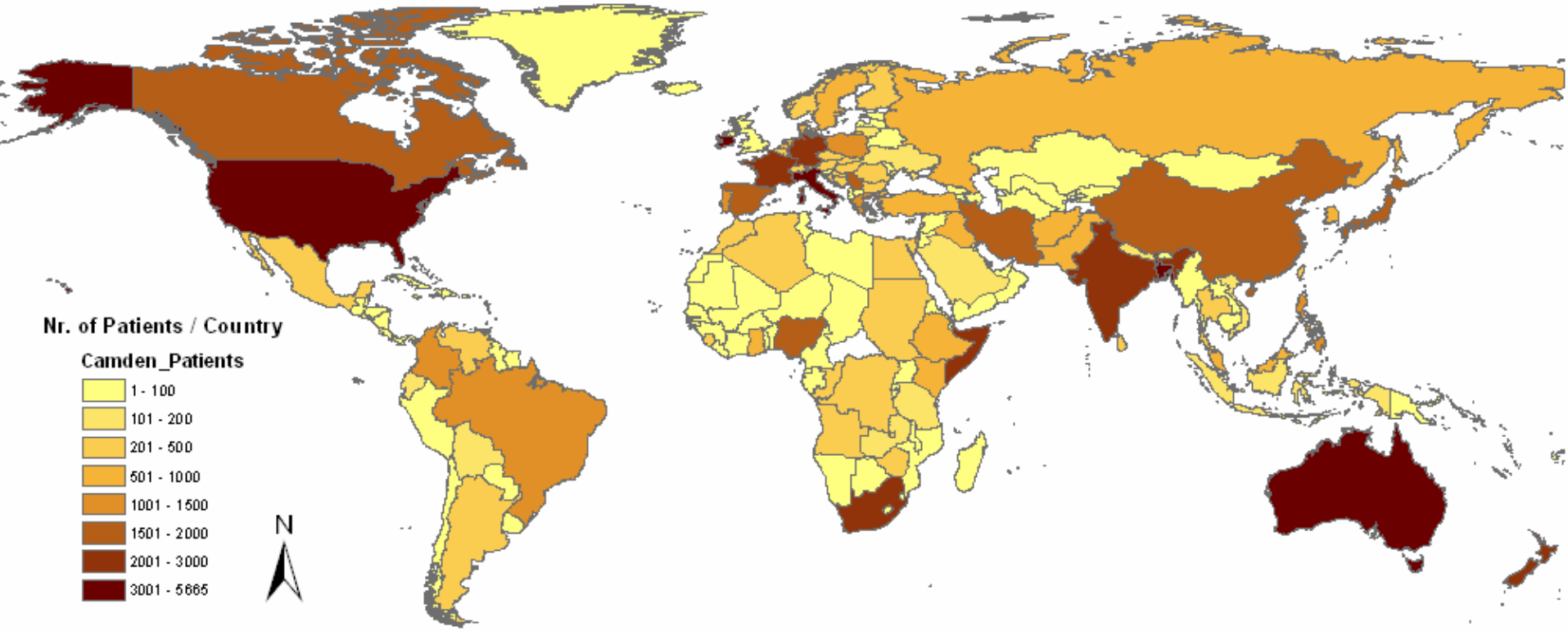
7% of Patient records have a Date of Arrival to the UK



Camden population top 30 countries of birth

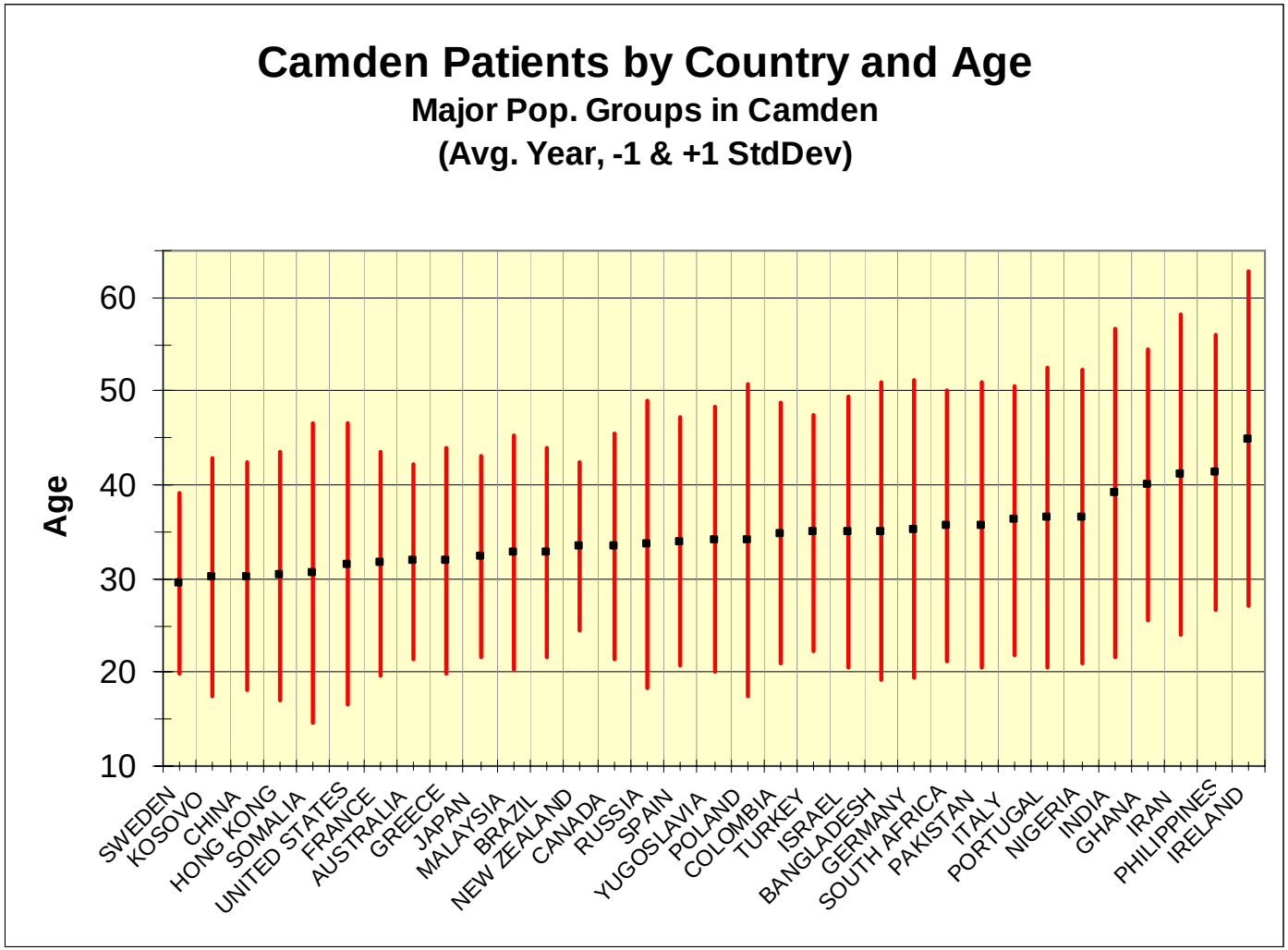


World map of Camden population

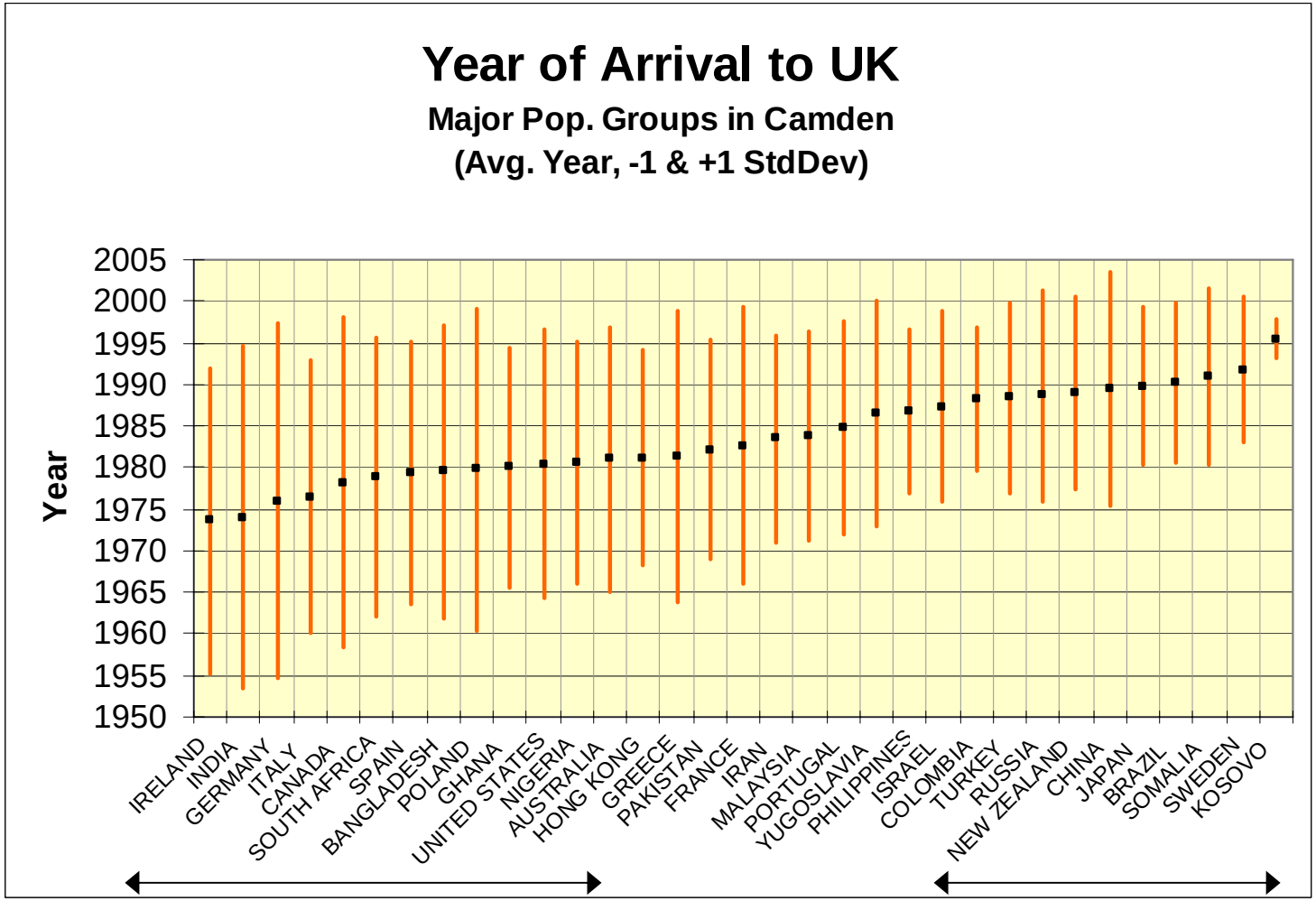


Camden 2004 population born abroad by country of birth

Analysis by country of birth (1)



Analysis by country of birth (2)



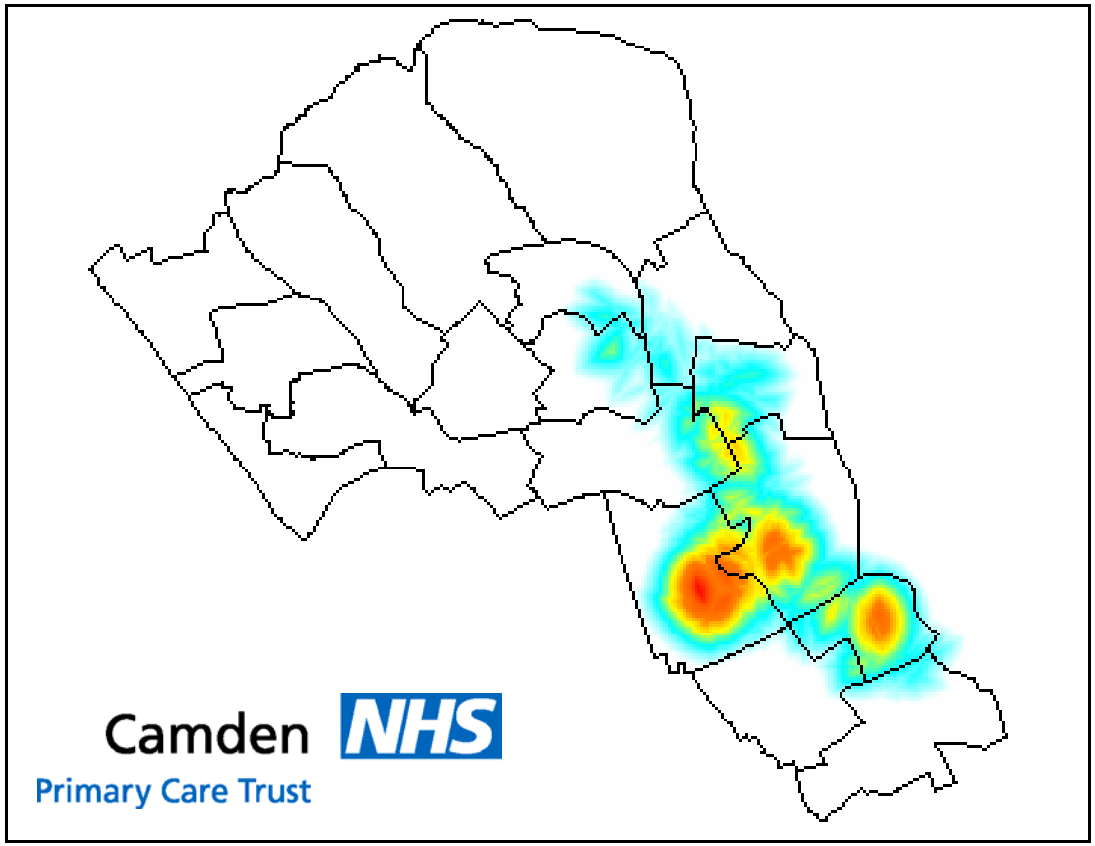
Older Immigration

Newer Immigration

COB & Public health

Non-responders to Breast Screening:

Women born in Bangladesh



Issues with COB analysis

- Miss-identification of 2nd & 3rd generation immigrants born in UK, and ‘White British’ born abroad.
- Cascade immigration (ie. Indians born in Eastern Africa)
- A large proportion of patient records does not contain COB (25% in Camden)

3 – Name ethnicity analysis

3- Name ethnicity analysis

- CASA project on quantitative name analysis
  http://surnames.casa.ucl.ac.uk/uclnames
- Names on Electoral Roll & historic Census have been assigned to a Cultural Ethnic & Linguistic group (*CEL*)

  56,000 surnames and 78,700 forenames
- Can potentially provide information about:

Language	Migration flows
Religion	Age
Geographic Origin	Gender

Literature on Names & Ethnicity

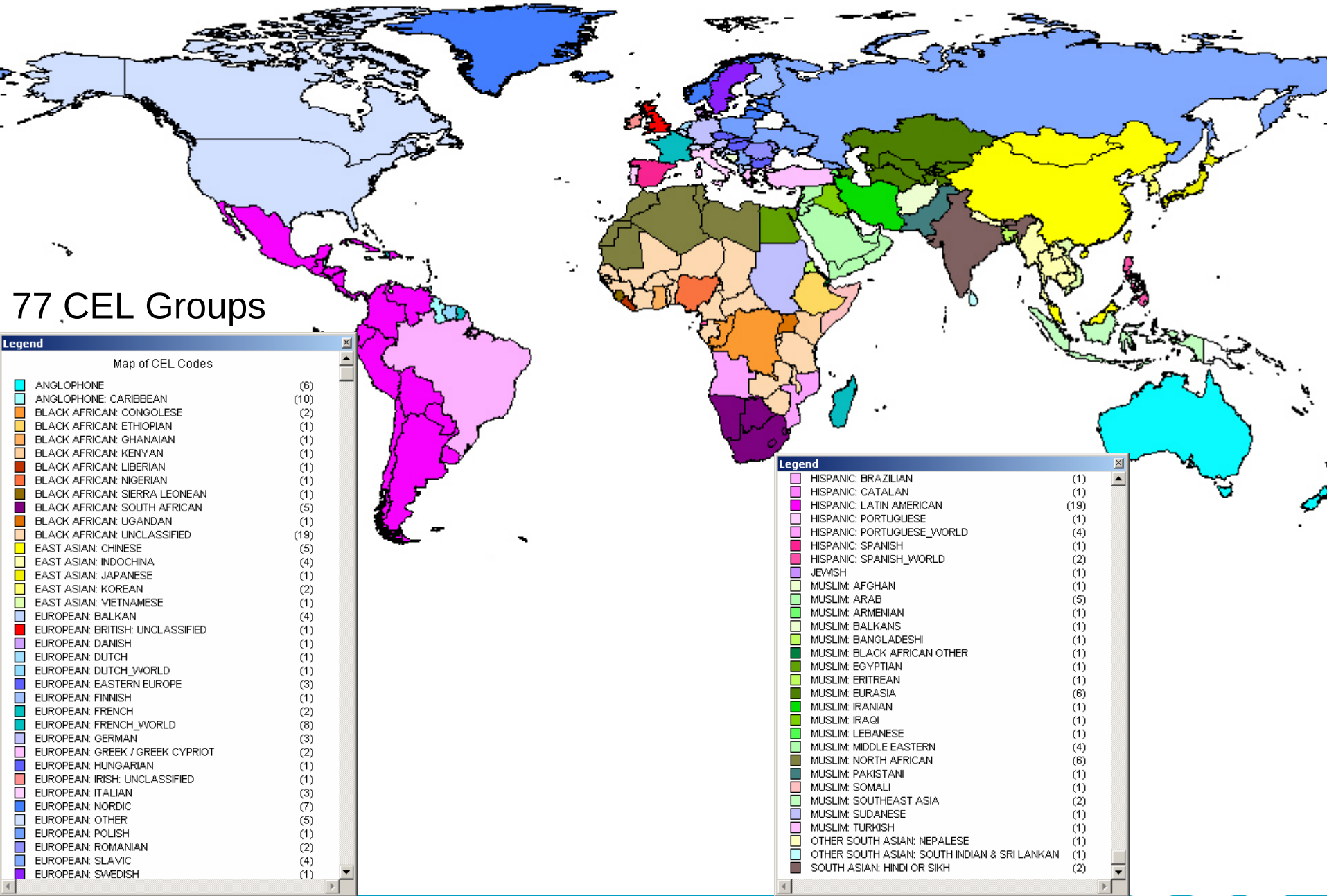
Paper	Geographical area		Ethnic Minorities (EM)	Name to CEL	Nr Surnames
	Ctry	Area of study			
Lauderdale & Kestenbaum (2000)	US	National	Chinese, Japanese, Filipino, Korean, Indian, & Vietnamese	Automatic	27,000
Word & Perkins (1996)	US	National	Hispanic	Automatic	25,276
Razum, Zeeb, & Akgun (2001)	Germany	Rhineland-Palatinate lander	Turkish	Automatic	12,188
Nanchahal, *et al* (2001)	UK	London, W.Midlands, Glasgow	South Asian	Automatic	9,422
Harding, Dews, & Simpson (1999)	UK	Bradford	South Asian + Hindu, Muslim & Sikh	Automatic	2,995
Cummins, *et al* (1999)	UK	Thames, Trent, W.Midlands & Yorkshire	South Asian	Automatic	2,995
Coldman, Braun & Gallagher (1988)	Canada	British Columbia	Chinese	Automatic	544
Choi, et al (1993)	Canada	Ontario	Chinese	Automatic	427
Hage, *et al* (1990)	Australia	Melbourne	Chinese		145
Martineau & White (1998)	UK	Newcastle (4 GPs)	Bangladeshi, Pakistani, Indian Muslims, Non-South Asian Muslims, Sikh, Hindu, White, Other	Manual Expert	N/A
Bouwhuis &. Moll (2003)	Netherland	Rotterdam (1 Hospital)	Turkish, Moroccan, Surinamese	Manual Expert	N/A
Nicoll, Bassett, & Ulijaszek (1986)	UK	Selected areas	South Asian	Manual Expert	N/A
Harland, White & Bhopal (1997)	UK	Newcastle	Chinese	Manual Expert	N/A

Name to CEL allocations

SURNAME	CULTURAL/ETHINC/LANGUAGE GROUP	Top Mosaic Type UK		Camden Top Country of Origin	Freq GB 1881	Freq GB 1998	%98/1881	GB 1881 Top area	GB 1996 Top area
WEINSTEIN	JEWISH;JEWISH	2	Cultural Leadership		22	156	709%		NW
WOOLF	JEWISH;JEWISH	1	Global Connections		893	1700	190%	E	NW
WEINER	JEWISH;JEWISH	1	Global Connections		25	260	1040%	WC	NW
WEISZ	JEWISH;JEWISH	2	Cultural Leadership		0	102			NW
GORSIA	JEWISH;JEWISH	1	Global Connections		19	218	1147%		HA
HALAI	JEWISH;JEWISH	1	Global Connections	Zimbabwe	18	161	894%		HA
BUX	JEWISH;JEWISH	3	Corporate Chieftains		28	272	971%	E	IG
JANJUA	JEWISH;JEWISH	1	Global Connections	Germany	146	635	435%	EC	WC
SAMAD	Muslim;Bangladeshi	26	South Asian Industry	Bangladeshi	0	236			NW
HUQ	Muslim;Bangladeshi	29	City Adventurers		0	141			NW
BHOJANI	Muslim;Bangladeshi	26	South Asian Industry	Bangladeshi	1	421	42100%		E
KHALIL	Muslim;Bangladeshi	26	South Asian Industry	Bangladeshi	21	104	495%		E
SAMAD	Muslim;Bangladeshi	26	South Asian Industry	Bangladeshi	0	216			E
KADRI	Muslim;Bangladeshi	26	South Asian Industry	Bangladeshi	0	115			E
KANBI	MUSLIM;Bangladeshi		#N/A		0	246			HA
MENDIS	Muslim;Bangladeshi	20	Asian Enterprise	India	2	373	18650%		HA
SALEM	MUSLIM;Egyptian	1	Global Connections	Egypt	11	394	3582%		NW
KHATRI	MUSLIM;Egyptian	1	Global Connections	Egypt	0	174			EC
BAH	MUSLIM;Egyptian	26	South Asian Industry	Egypt	3	157	5233%		N
SHABBIR	Muslim;Egyptian	1	Global Connections	Egypt	0	105			WC
BAPU	Muslim;Eritrean	26	South Asian Industry	Eritrea	0	316			IG
MURAD	MUSLIM;Iraqi	26	South Asian Industry	Iraq	0	142			NW
KHALIL	Muslim;Lebanese	26	South Asian Industry	Lebanon	0	774			NW
SOLIMAN	Muslim;NORTH AFRICAN	36	Metro Multiculture	Egypt	1	137	13700%		NW
KADRI	MUSLIM;North African	20	Asian Enterprise	Algeria	0	178			NW
SAYED	Muslim;NORTH AFRICAN	20	Asian Enterprise		0	357			NW

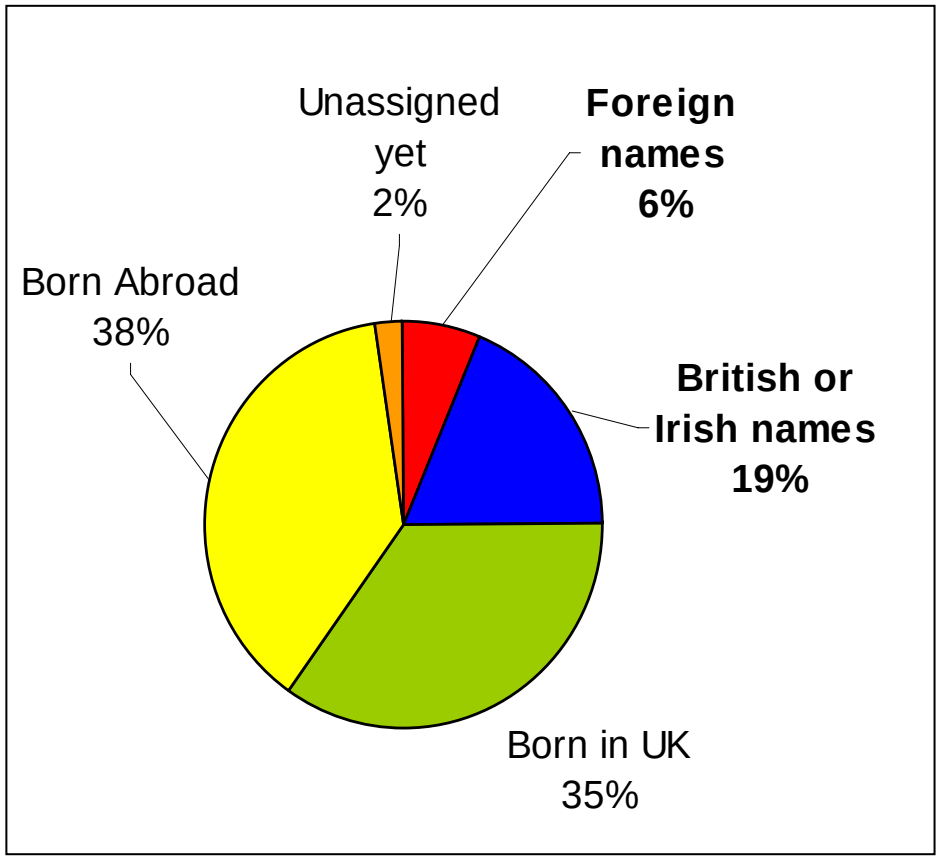
World map of CEL groups

77 CEL Groups



Initial name analysis

- Enrich patient records with no Birth Place, by assigning a probability of ethnic group through their Surname or Forename



- More correct to apply it to all records in order to:
  - Identify 2nd & 3rd generation immigrants born in UK
  - Account for ‘White British’ born abroad

4 – Household ethnicity analysis

4- Household ethnicity analysis

- Patient's Address Geocoded to a UPRN
  (Unique Property Reference Number from a Local Property Gazetteer)

UPRN	SURNAME	AGE	GENDER	COB
123456	Soandso	1	M	UNITED KINGDOM
123456	Soandso	5	F	UNITED KINGDOM
123456	Soandso	8	F	ALBANIA
123456	Soandso	33	F	ALBANIA
123456	Soandso	52	M	ALBANIA
654321	Z1	8	F	UNITED KINGDOM
654321	Z1	15	F	AUSTRALIA
654321	Z1	16	F	
654321	Z1	18	M	SUDAN
654321	Z2	40	F	SUDAN

Household Most Likely CEL

Albanian (3 out of 5)

Sudanese (2 out of 4)

5 – CEL model compilation and evaluation

5- Compiling a CEL model

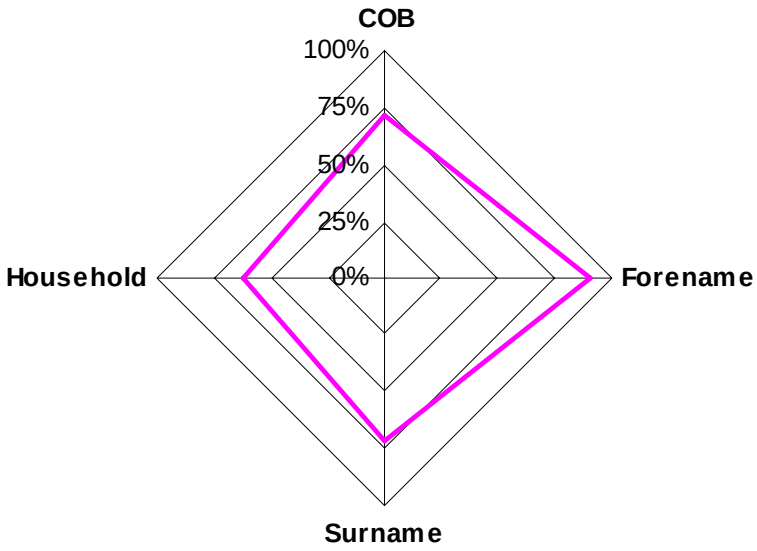
Compilation of potential CEL groups per person

- Country of Birth
- Surname
- Forename
- Household

Surname	COB_CEL	Surname_CEL	Forename_CEL	Household_CEL	Nr of coincidences	Most Likely CEL
SoandSo	British_Unclassified	Bangladeshi	Muslim: Unclassified	Bangladeshi	2	Bangladeshi

4-CELs coverage & match

CEL Matched	Total Available	Patients	%	% Cumm.
4	4	5234	2.6%	2.6%
3	4	8284	4.1%	6.6%
3	3	5421	2.7%	9.3%
2	4	40200	19.8%	29.1%
2	3	51330	25.2%	54.3%
2	2	7696	3.8%	58.1%
1	1	9128	4.5%	62.6%
1	2	41390	20.3%	82.9%
1	3	26228	12.9%	95.8%
1	4	8484	4.2%	100.0%

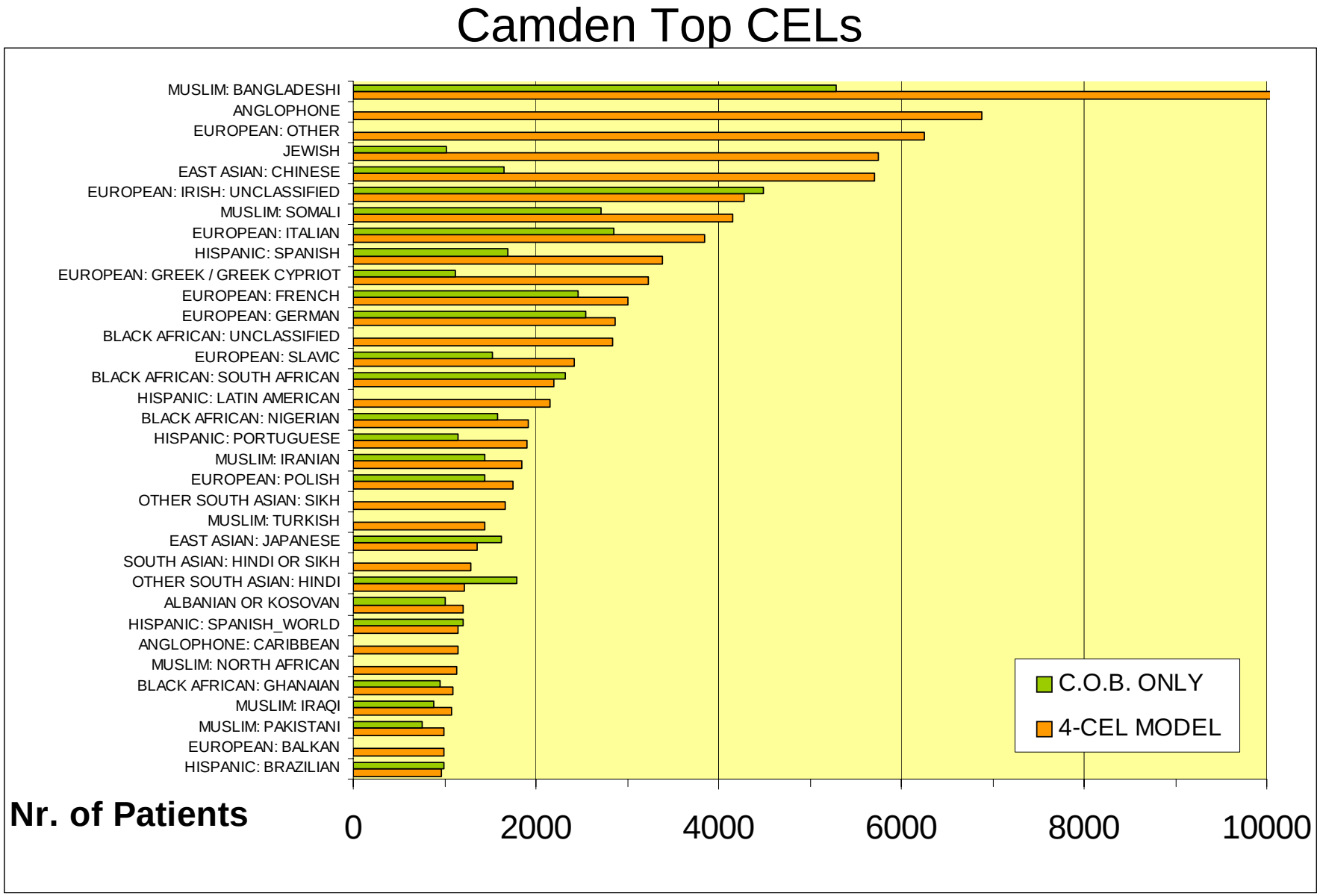


Algorithms to choose most likely CEL

Code	Description	CEL_Chosen
449	4/4 All variables match	Surname
349	3/4 Odd is NOT Surname	Surname
348	3/4 Odd IS Surname	COB
339	3/3 One missing is NOT Surname NOR COB	Surname
338	3/3 One missing IS COB	Surname
337	3/3 One missing IS Surname	COB
249	2/4 Surname = COB	Surname
248	2/4 Surname = Forename	Surname
247	2/4 Surname = Household	Surname
246	2/4 COB = Forename	COB
245	2/4 Forename = Household	Forename
244	2/4 COB = Household	COB
239	2/3 Surname = COB	Surname
238	2/3 Surname = Forename	Surname
237	2/3 Surname = Household	Surname
236	2/3 COB = Forename	COB
235	2/3 Forename = Household	Forename
234	2/3 COB = Household	COB
229	2/2 Surname = COB	Surname
228	2/2 Surname = Forename	Surname
227	2/2 Surname = Household	Surname
226	2/2 COB = Forename	COB
225	2/2 Forename = Household	Forename
224	2/2 COB = Household	COB

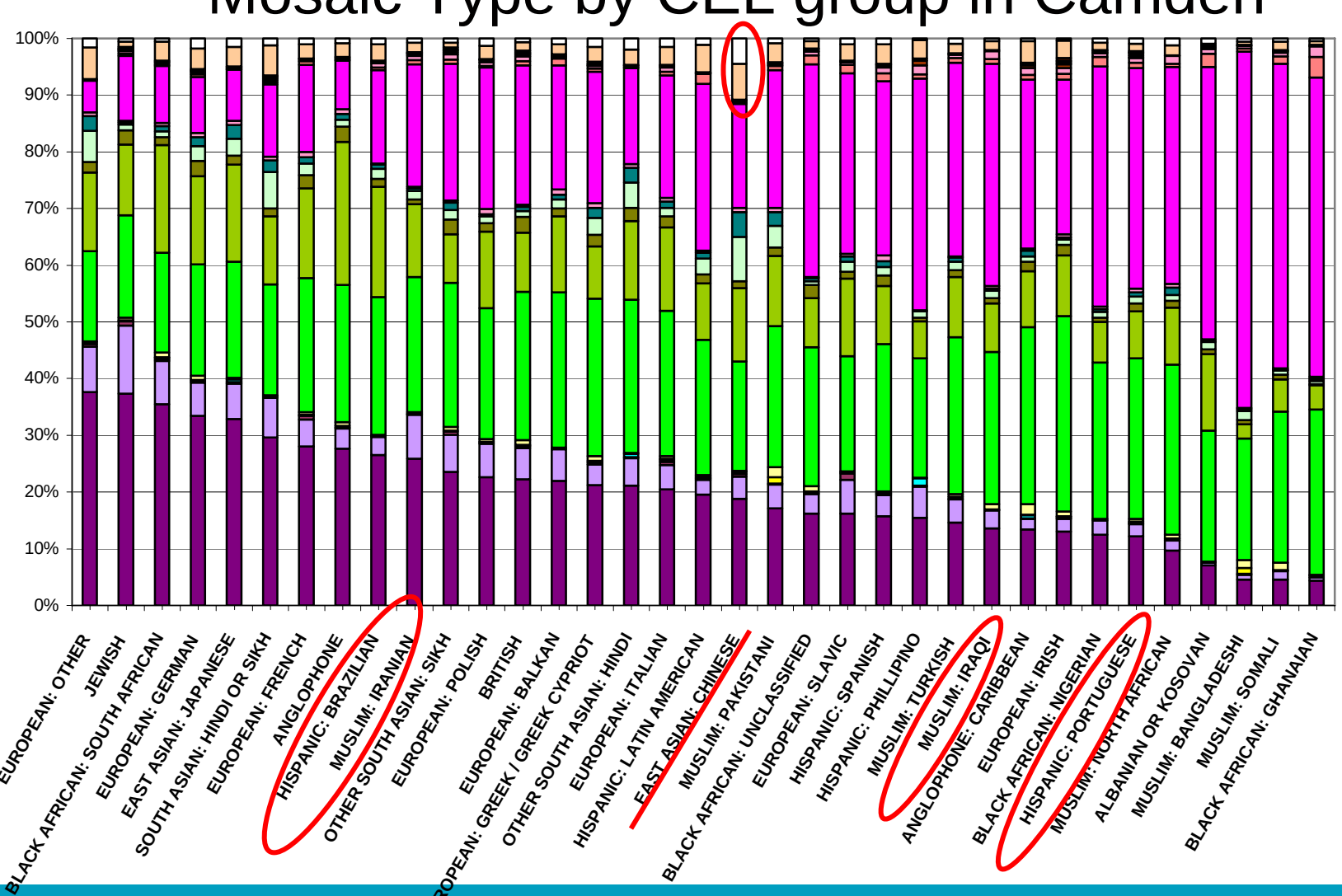
Code	Description	CEL_Chosen
149	1/4 Surname + COB = BRITISH	Surname
148	1/4 Surname + COB<> BRITISH	COB
139	1/3 Surname + COB = BRITISH	Surname
138	1/3 Surname + COB<> BRITISH	COB
137	1/3 Surname + Forename+ Household	Surname
136	1/3 COB + Forename + Household	COB
129	1/2 Surname + COB = BRITISH	Surname
128	1/2 Surname + COB<> BRITISH	COB
127	1/2 Surname + Forename	Surname
126	1/2 Surname + Household	Surname
125	1/2 COB<> BRITISH	COB
124	1/2 COB = BRITISH + Forename	COB
123	1/2 COB = BRITISH + Household	Household
122	1/2 Forename + Household	Household
119	1/1 Surname	Surname
118	1/1 COB	COB
117	1/1 Forename	Forename
116	1/1 Household	Household

COB Vs 4-CEL model



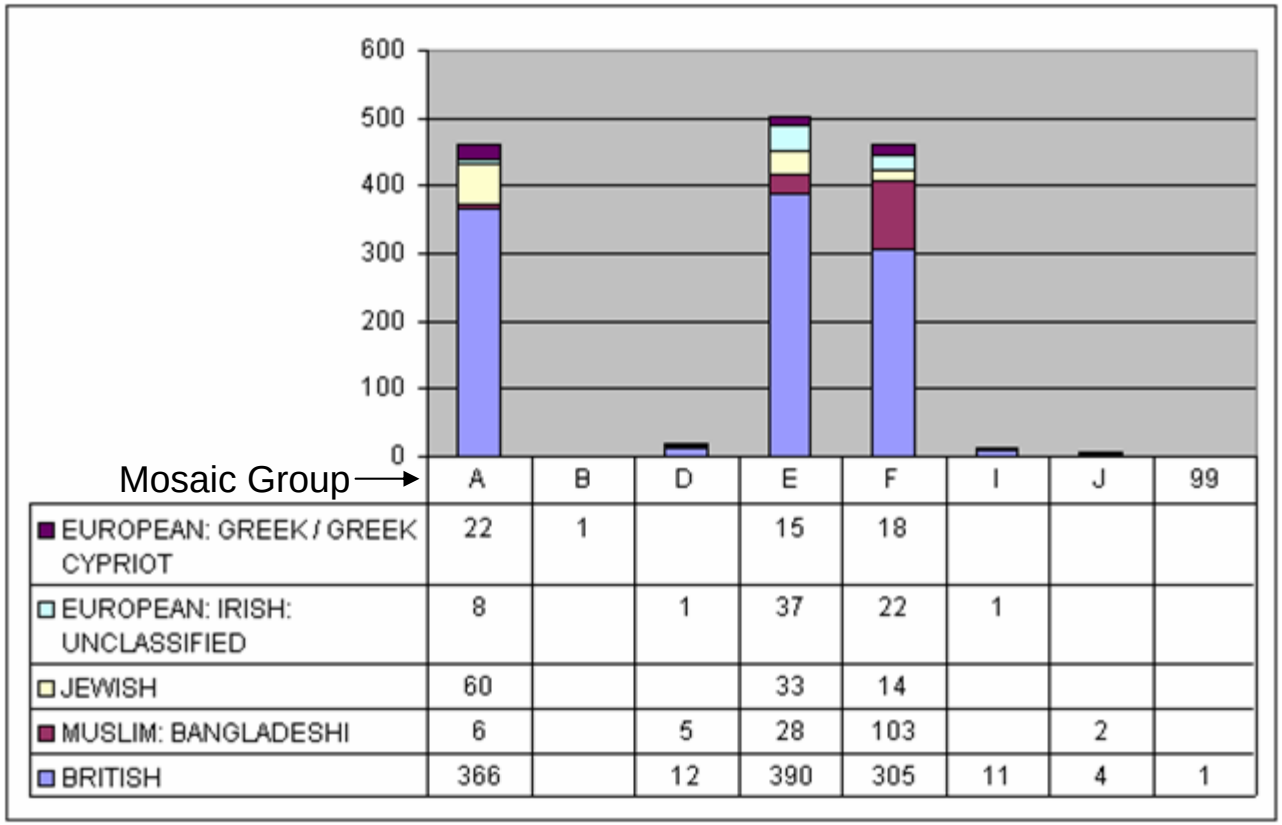
Ethnic Innequalities

Mosaic Type by CEL group in Camden



Ethnicity and breast screening

Top 5 Ethnic Groups who do not respond to breast screening invites



Mosaic Group →	A	B	D	E	F	I	J	99
EUROPEAN: GREEK / GREEK CYPRIOT	22	1		15	18			
EUROPEAN: IRISH: UNCLASSIFIED	8		1	37	22	1		
JEWISH	60			33	14			
MUSLIM: BANGLADESHI	6		5	28	103		2	
BRITISH	366		12	390	305	11	4	1

Evaluating the Model

- Evaluation of the CEL model through self-reported ethnicity from Hospital Episode Statistics
  - 40,714 patients (20% of total) matched to a unique true ethnic code (1991 Census categories)

		Actual Ethnicity from HES data											
Predicted by CEL		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**	**Sensitivity**	**Specificity**	**PPV**
0	White	**24,656**	624	652	331	88	23	388	46	2,499	29,307	0.92	0.67	0.84
1	Black - Caribbean	35	**147**	3	15	3			1	35	239	0.17	1.00	0.62
2	Black - African	385	44	**1,948**	174	47	11	22	5	438	3,074	0.67	0.97	0.63
3	Black - Other										0	0.00	1.00	
4	Indian	426	15	17	8	**333**	16	12	2	150	979	0.13	0.99	0.44
5	Pakistani	19	1	3		22	**75**	11		29	160	0.32	1.00	0.47
6	Bangladeshi	96	5	59	37	132	75	**2,672**	1	292	3,369	0.84	0.98	0.79
7	Chinese	126	2	12	2	6	1	1	**272**	94	516	0.73	0.99	0.53
8	Any other ethnic group	1,046	19	196	64	67	36	87	44	**1,511**	3,070	0.30	0.96	0.49
	Total	26,789	857	2,890	631	698	237	3,193	371	5,048	**40,714**			

Issues with the model

- Fails to measure mixed ethnicity
- Problems with 2nd or 3rd generation migrants
- Changes of name (marriage, other)
- Should ‘UK ethnicity’ be treated separately?
- Need to introduce probabilistic and fuzzy CEL allocations

6- Future enhancements to the ethnic classification model

- Improve household structure and overall model algorithms
- Expand name analysis
  - Introduce language and religion at subnational geographies
  - Introduce probabilistic and fuzzy CEL allocations
- Involve other London PCTs and maybe NHS nationally in the analysis
  - Broaden the surname/forename base & placename alias tables
  - Disseminate the methods & tools

    www.casa.ucl.ac.uk/geonom

Questions?

Centre for Advanced Spatial Analysis (CASA) University College London

www.casa.ucl.ac.uk/geonom