Estimating the Determinants of Health Literacy for Policy Prioritisation

UCL CDS Symposium on Data Science in Public Health

Nathan Green
Department of Statistical Science, UCL

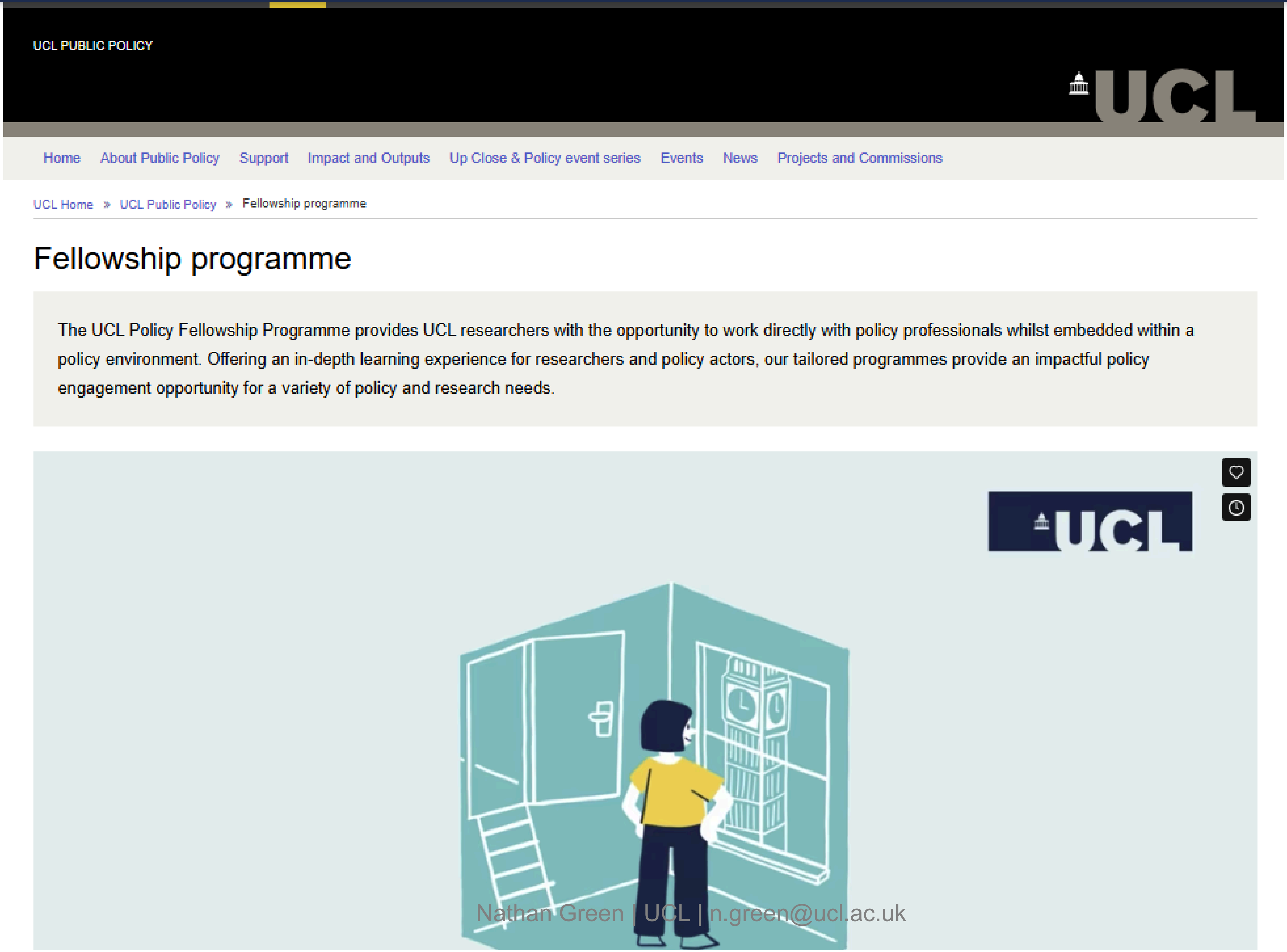
Outline

- Background
- Problems
- Solutions
  - Local level estimation
  - Predictive comparisons
  - Prioritisation
- Conclusion

ⓘ **Resources**

Slides and code here: `github.com/n8thangreen/data-science-in-health-talk`

Background



UCL PUBLIC POLICY

Fellowship programme

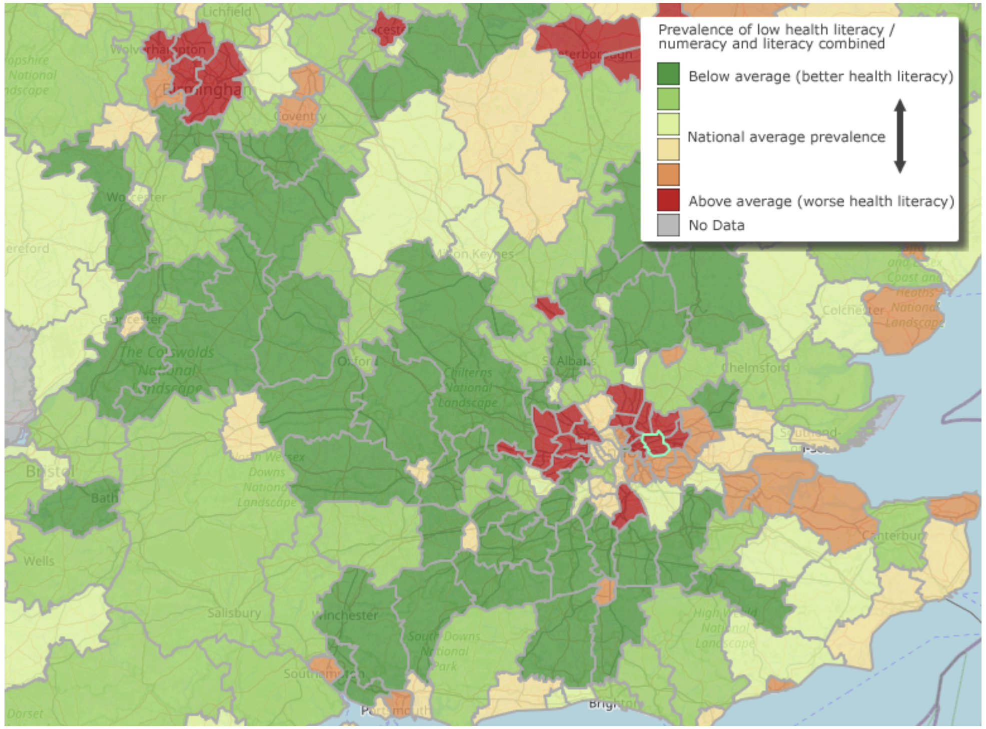
The UCL Policy Fellowship Programme provides UCL researchers with the opportunity to work directly with policy professionals whilst embedded within a policy environment. Offering an in-depth learning experience for researchers and policy actors, our tailored programmes provide an impactful policy engagement opportunity for a variety of policy and research needs.

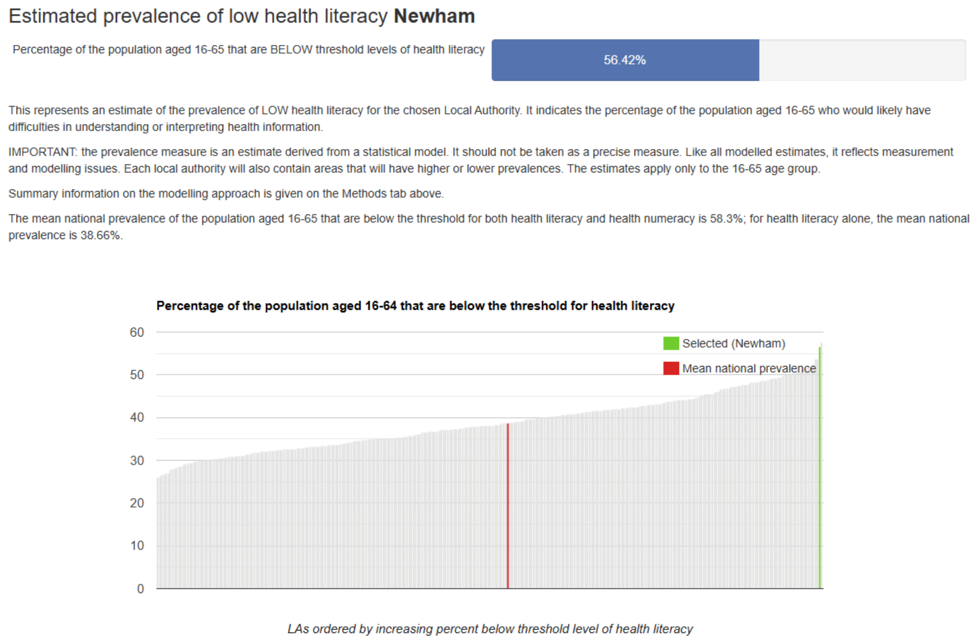
Project outline

- Title: **Assesses the factors that determine health literacy and the size of their influence/impact for Newham**
- Health literacy is broadly defined as the ability to access, understand, appraise, and communicate health information, enabling individuals to engage in healthcare and maintain good health throughout their lives.

Determinants of Health Literacy

- Focusses on Newham, a diverse borough in East London that faces unique challenges
- Identified as having some of the lowest levels of health literacy in the UK by University of Southampton (https://healthliteracy.geodata.uk/)


Prevalence of low health literacy / numeracy and literacy combined
🟩 Below average (better health literacy)
🟩
🟨 National average prevalence
🟨
🟧
🟥 Above average (worse health literacy)
⬜ No Data
Lichfield
Wolverhampton
Birmingham
Coventry
Peterborough
Worcester
Hereford
Milton Keynes
Gloucester
The Cotswolds National Landscape
Oxford
St Albans
Chilterns National Landscape
Chelmsford
Colchester
Southend
Bristol
Bath
Wells
Salisbury
Winchester
South Downs National Park
Southampton
Portsmouth
Brighton
Canterbury
High Weald National Landscape
Dorset



Estimated prevalence of low health literacy **Newham**

Percentage of the population aged 16-65 that are BELOW threshold levels of health literacy 56.42%

This represents an estimate of the prevalence of LOW health literacy for the chosen Local Authority. It indicates the percentage of the population aged 16-65 who would likely have difficulties in understanding or interpreting health information.

IMPORTANT: the prevalence measure is an estimate derived from a statistical model. It should not be taken as a precise measure. Like all modelled estimates, it reflects measurement and modelling issues. Each local authority will also contain areas that will have higher or lower prevalences. The estimates apply only to the 16-65 age group.

Summary information on the modelling approach is given on the Methods tab above.

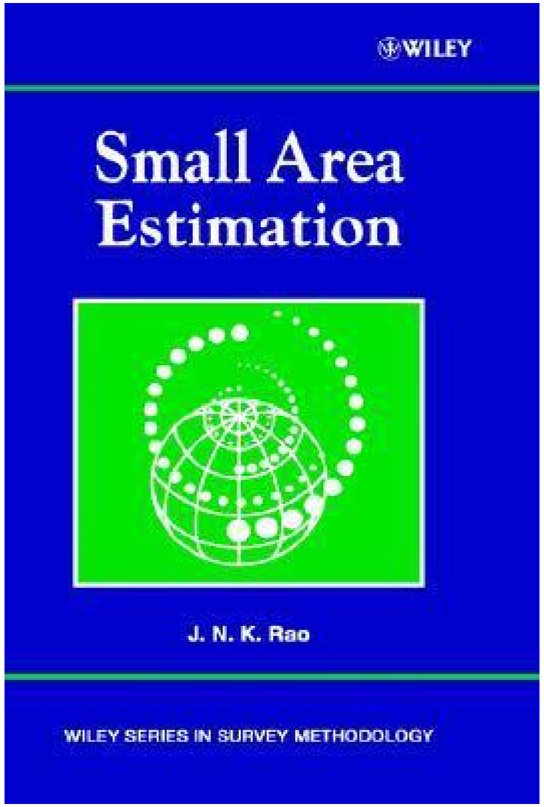
The mean national prevalence of the population aged 16-65 that are below the threshold for both health literacy and health numeracy is 58.3%; for health literacy alone, the mean national prevalence is 38.66%.

Percentage of the population aged 16-64 that are below the threshold for health literacy

60
50
40
30
20
10
0
🟩 Selected (Newham)
🟥 Mean national prevalence
LAs ordered by increasing percent below threshold level of health literacy

Previous method: Synthetic estimation

- Weighted Logistic Regression with Synthetic Estimation (Laursen et al. (2016))
  - Frequentist single-level regression with poststratification
- Used in Small Area Estimation (SAE) (Gonzalez (1973); Rao and Molina (2015))
- Can be viewed as the simpler predecessor to Multilevel Regression with Post-stratification (MRP)
  - Ignores any unique local factors
  - MRP includes shrinkage via random effects



Talking a different language 🗣

Small Area Estimation (SAE) (Previous)		**HTA / Statistics Method (Me)**
Weighted Logistic Regression with Synthetic Estimation	⟶	Multilevel Regression with Post-stratification (MRP)
Linear Plug-In Model *(Equivalent to Regression-Synthetic Estimator at Unit Level)*	⟶	Simulated Treatment Comparison (STC)
Residual-adjusted synthetic estimation	⟶	Targeted Maximum Likelihood Estimation (TMLE) *(in causal inference)*

Problems

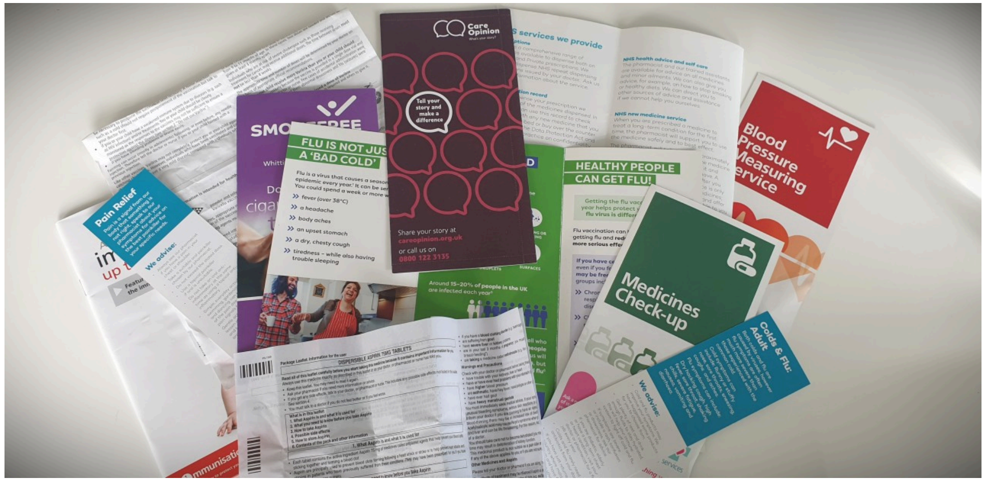
1. What is health literacy level specific to Newham?
2. What are the 'drivers' of health literacy?
3. How should we intervene to effect health literacy outcomes?

Available data

- **Skills for Life (SfL) Survey 2011** [MRP]
  - Comprehensive computer-based assessment conducted by the ONS to evaluate the skills of literacy, numeracy, and ICT
  - Total 7230 adults in England
- **Newham Residents Survey 2023 (NRS)** [MRP]
  - Periodic survey, usually every two years
  - Detailed information on views, experiences, and needs of Newham residents
  - Covers satisfaction with local services, community safety, health and well-being, housing, and employment
- Additional data
  - UK Programme for International Assessment of Adult Competencies (PIAAC) 2023 [MRP]
  - Skills for Life Survey 2003 [MRP]
  - Labour Force Survey (LFS) / Annual Population Survey (APS) [MRP]
  - UK Census 2011, 2021 [MRP]

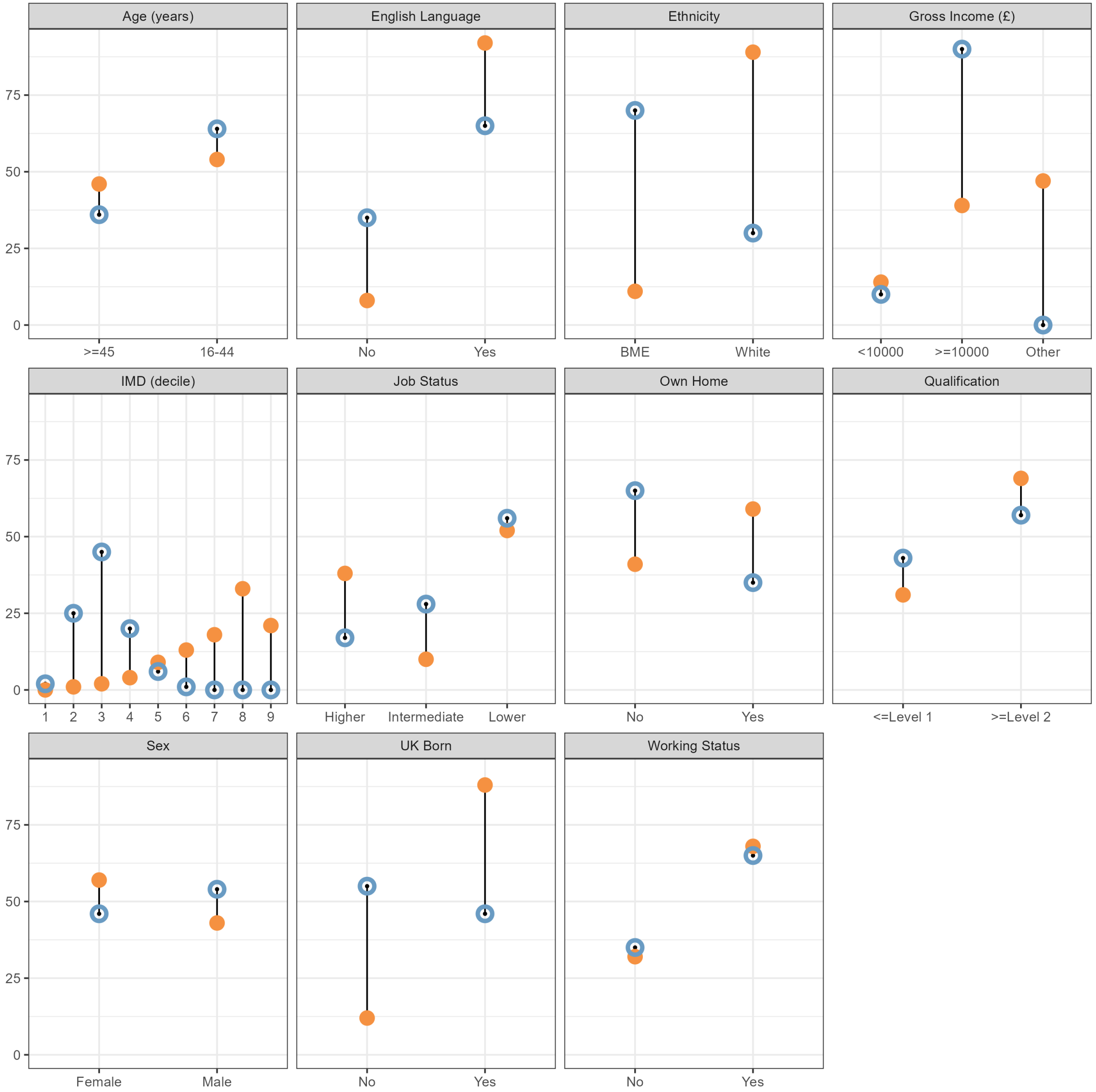
From Skills for Life to Health literacy

- From Rowlands et al. (2015)
- Sample of health materials, including
  - medicine labels
  - booklets
  - application forms
- Covered themes of health promotion, managing illness, systems navigation and disease prevention
- Assessed for literacy and numeracy complexity by education experts
- SfL responses were mapped to the binary health literacy scale according to whether they are above or below threshold


Care Opinion
FLU IS NOT JUST A 'BAD COLD'
HEALTHY PEOPLE CAN GET FLU!
Blood Pressure Measuring Service
Medicines Check-up
Pain Relief
Colds & Flu: Adult

Problem 1: Newham specific health literacy estimates? 🤔

Newham vs Skills for Life profiles



Mutlilevel Regression and Post-stratification

The predicted probability defined as:

$$\hat{\pi}_i = \text{logit}^{-1}\left(\hat{\beta}_0 + \sum_x \hat{\beta}^x_{\gamma_x[i]}\right)$$

- $\hat{\beta}_0$ is the intercept, $\hat{\beta}^x_{\gamma_x[i]}$ are coefficients for covariates x
  - age, sex, English language, white ethnicity, UK born, qualifications, income, job status, work role, home ownership
- $\gamma_x[i]$ represents the level or category for covariate x for individual i
- IMD is included as multilevel random effects $\beta^{\text{IMD}}_j \sim \text{N}(\mu_{\text{IMD}}, \sigma^2_{\text{IMD}})$
- Prior distributions for fixed effects normal distributions centered at zero with modest variance
- Half-normal priors are used for random effect standard deviations

Mutlilevel Regression and Post-stratification

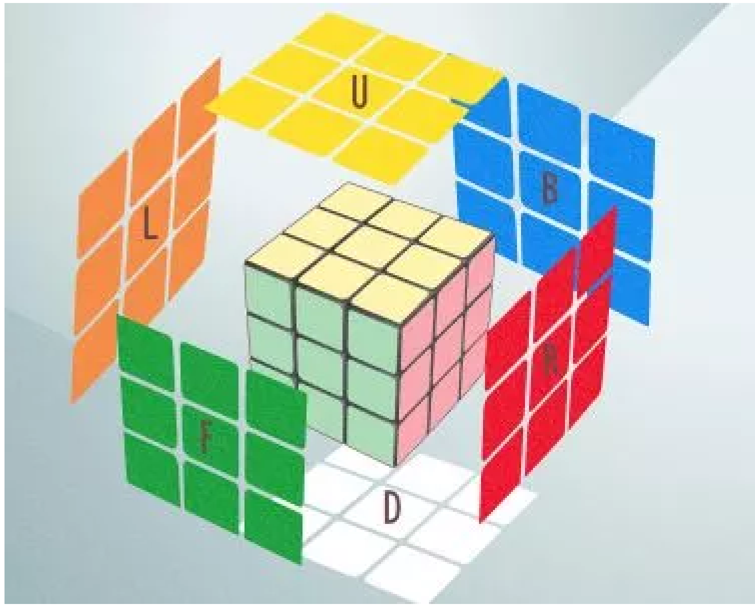
- The health literacy probabilities for each demographic category (cell c) are weighted by their proportion in the actual Newham population
- 11 covariates $\longrightarrow$ 13,824 cells
- Post-stratified estimate is:

$$\hat{\pi}^{\mathrm{mrp}} = \sum_{c=1}^{|\mathcal{S}|} \frac{N_c \hat{\pi}_c}{N}$$

- $\mathcal{S}$ is the set of all covariate combinations
- N_c is the population frequency for cell c
- N is the total population size

Missing joint distributions

- Raking / Iterative proportional fitting (IPF)
    - Adjust survey weights so that the sample distribution matches known population control totals (margins)
- Census data → Marginals
- Labour Force Survey (LFS) → Covariance structure
- Overlap issues, non-representative
    - Data augmentation *before* IPF
    - Laplace smoothing *after* IPF
        - Like “zero cell” problem in meta-analyses
- Copula method alternative



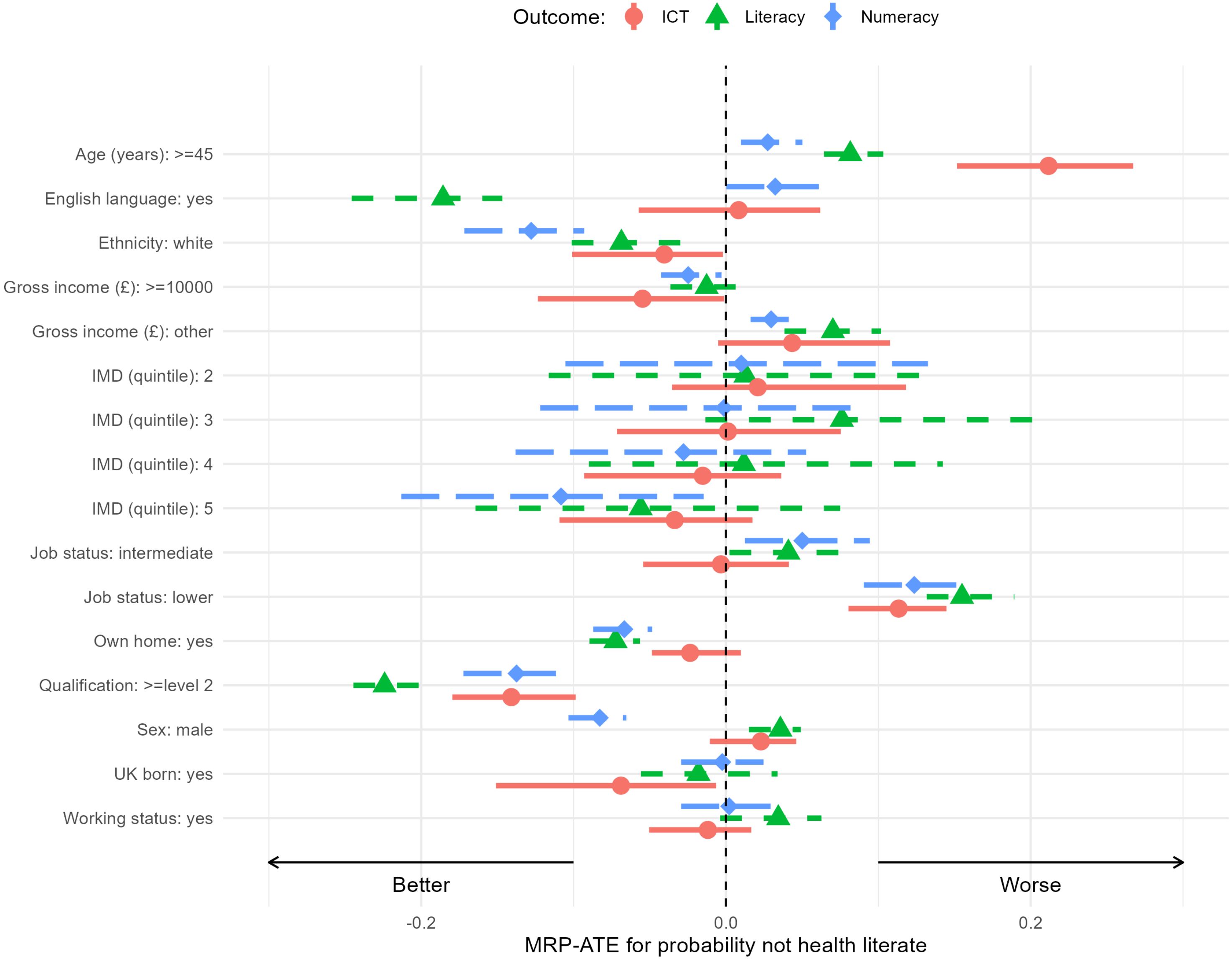
Problem 2: What are the ‘drivers’ in Newham? 🤔

Predictive comparisons

- Terminology borrow from Gelman and Pardoe (2007). Also called predicted change in probability
- Previously, crops up in other fields e.g. Lee (1981) (covariance adjustment mean difference)
- Like average treatment effects without the causal interpretation

$$\delta_u(u^{(1)}, u^{(2)}) = \frac{\mathbb{E}(y \mid u^{(2)}) - \mathbb{E}(y \mid u^{(1)})}{u^{(2)} - u^{(1)}}$$

MRP-PC results



Problem 3: How should we intervene? 🤔

Priority ranking

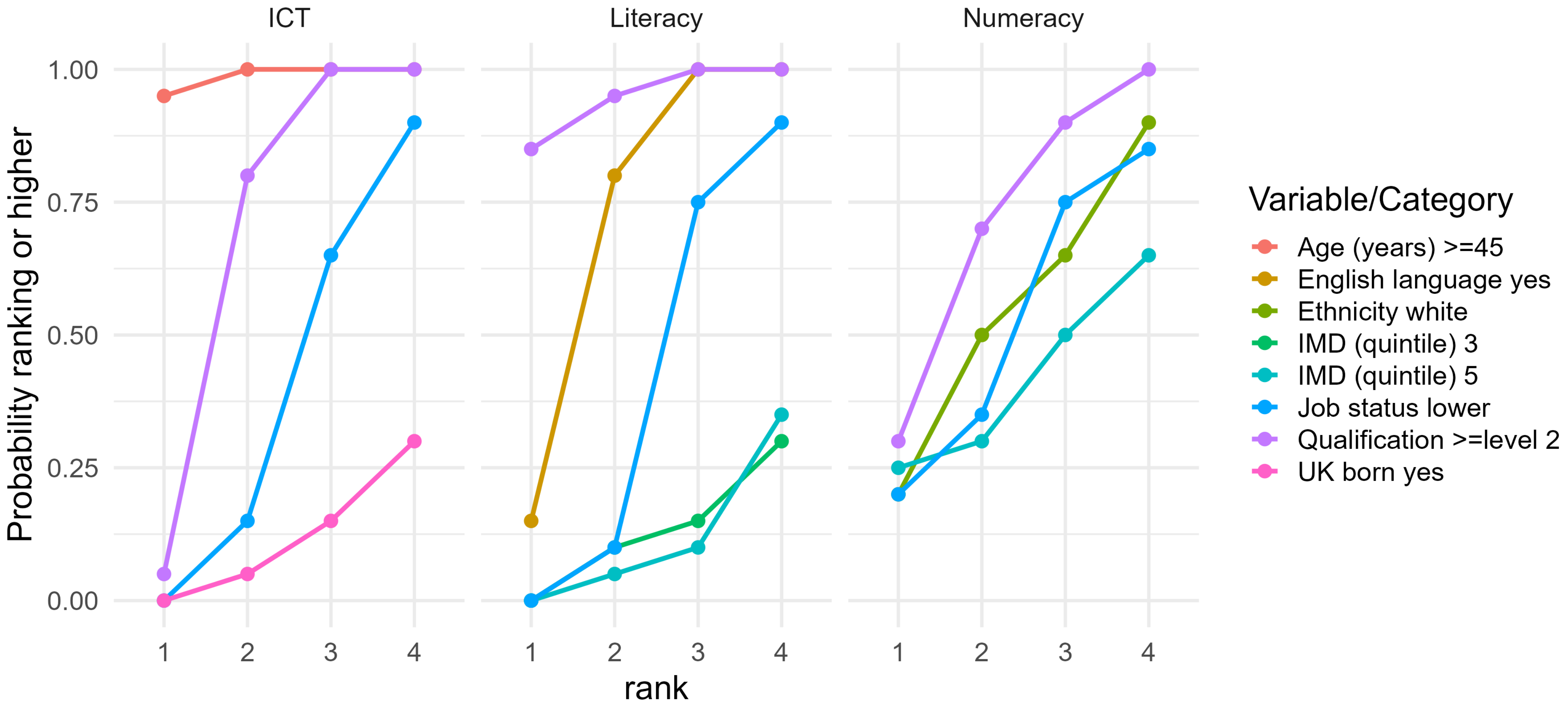
- Adopt Surface Under the Cumulative Ranking Curve (SUCRA)
  - Common in multiple-treatment meta-analysis
- Percentage of the maximum possible cumulative rank an intervention can achieve
- Providing a single value where a higher SUCRA indicates a better overall rank relative to others

$$\text{SUCRA}_{ij} = \sum_{r=1}^{n-1} P_{ijr}/(n-1),$$

- where P_{ijr} is the cumulative probability for variable i at level j and rank r
- Mean rank is

$$\mathbb{E}[\text{rank}(i,j)] = n - \sum_{r=1}^{n-1} P_{ijr}$$

Ranking results



		SUCRA			$\mathbb{E}$[rank]		
Variable	**Category**	**ICT**	**Literacy**	**Numeracy**	**ICT**	**Literacy**	**Numeracy**
Age (years)	≥45	100	66	27	1	6	12
English Language	Yes	33	93	34	11	2	11
Ethnicity	White	46	56	88	9	8	3
Gross Income (£)	≥10000	55	11	28	8	14	12
	Other	49	55	32	9	8	11
IMD (quintile)	2	34	35	47	11	11	9
	3	36	51	40	11	8	10
	4	39	40	45	10	10	9
	5	43	49	74	10	9	5
Job Status	Intermediate	31	30	49	11	12	9
	Lower	85	85	87	3	3	3
Own Home	Yes	36	58	62	11	7	7
Qualification	≥Level 2	92	99	93	2	1	2
Sex	Male	32	28	70	11	12	5
UK Born	Yes	63	16	13	6	14	14
Working Status	Yes	26	26	11	12	12	14

Conclusions

- The job is not done with the modelling ⛔
- Borrow methods from other fields
- Data issues are inevitable…deal with it
- Clear communication of results for SME and decision-maker is crucial
- It’s an iterative, team effort from project inception to decision

Thanks 🙏

References

Gelman, Andrew, and Iain Pardoe. 2007. “Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components.” *Sociological Methodology* 37 (1): 23–51. https://doi.org/10.1111/j.1467-9531.2007.00181.x.

Gonzalez, Maria E. 1973. “Use and Evaluation of Synthetic Estimates.” In *Proceedings of the Social Statistics Section, American Statistical Association*, 33–42. American Statistical Association.

Hutcheon, Jennifer A, Arnaud Chiolero, and James A Hanley. 2010. “Random Measurement Error and Regression Dilution Bias.” *BMJ* 340. https://doi.org/10.1136/bmj.c2289.

Laursen, Kamilla R., Paul T. Seed, Joanne Protheroe, Michael S. Wolf, and Gill P. Rowlands. 2016. “Developing a Method to Derive Indicative Health Literacy from Routine Socio-Demographic Data.” *Journal of Health Care Communications* 1 (4): 1–9. https://doi.org/10.4172/2472-1654.100033.

Lee, James. 1981. “Covariance Adjustment of Rates Based on the Multiple Logistic Regression Model.” *Journal of Chronic Diseases* 34 (8): 415–26. https://doi.org/10.1016/0021-9681(81)90006-4.

Rao, J. N. K., and Isabel Molina. 2015. *Small Area Estimation*. 2nd ed. Wiley Series in Survey Methodology. John Wiley & Sons.

Rowlands, G, J Protheroe, J Winkley, et al. 2015. “A Mismatch Between Population Health Literacy and the Complexity of Health Information: An Observational Study.” *British Journal of General Practice* 65 (635): e379–86. https://doi.org/10.3399/bjgp15X685285.