Frontiers in Social Statistics Methodology,
Feb 2017

# Small Area Estimation for Health Data

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

Dr Diane Hindmarsh NSW Bureau of Health Information

# Introduction

"*The client will always require more than is specified at the design stage*"

Fuller (1999) in Rao (2003)

# Estimation at small area important

- Surveys designed to provide robust results at level of health administrative unit
- Results at small area level required
  - To help identify and target areas that are worse off
  - To help measure the health "gap" between areas
  - To provide local health information for health monitoring, for local needs-assessment and in planning and supporting new services

Finding sources of data that provide robust estimates at the small area level is ~~difficult~~ sometimes **impossible**

# Basic idea behind SAE methods

Assume the response of the outcome
at the LGA level to
Age
Sex
Marital status...
Is the same in all areas across the state

And...perhaps, that each area might be a little different, on average
(random effect at area level)

Model is assumed to also apply to POPULATION

# The NSW Population Health Survey

- Ongoing CATI survey commenced 2002
- ~1,000 people per health administrative "unit" – 8 to 17 areas
- Weighted to 5-year age, sex profile of stratum
- 95% able to be coded to an LGA
  - Respondents x LGA x year ~0 to 240+
  - Median 2002 to 2008 combined: 328 (from 32 to 1955)

# Aim

Obtain estimates at LGA level using the NSW Population Health Survey data.

- Easy to calculate, including estimate of precision (practical – use SAS procedures)
- Show improvement over direct estimates
  - Validation
- Can a *single method* be used for a range of outcome variables

# Early decisions:

- Use a unit-level model
- Model separately by sex
  - use age group as a covariate (limitation of Census data)
- Ignore survey design
  - Include variables used in the survey design in model
- Use frequentist framework for estimation
- Fit linear and logistic models

# The Methods

# Unit level linear model – EBLUP

$$y_{ig} = \mathbf{x}_{ig}^T \boldsymbol{\beta} + v_g + e_{ig}$$

$$v_g \overset{iid}{\sim} \left(0, \sigma_v^2\right) \quad e_{ig} \overset{iid}{\sim} \left(0, \sigma_e^2\right)$$

$$i = 1, \ldots n_g \quad g = 1, \ldots G$$

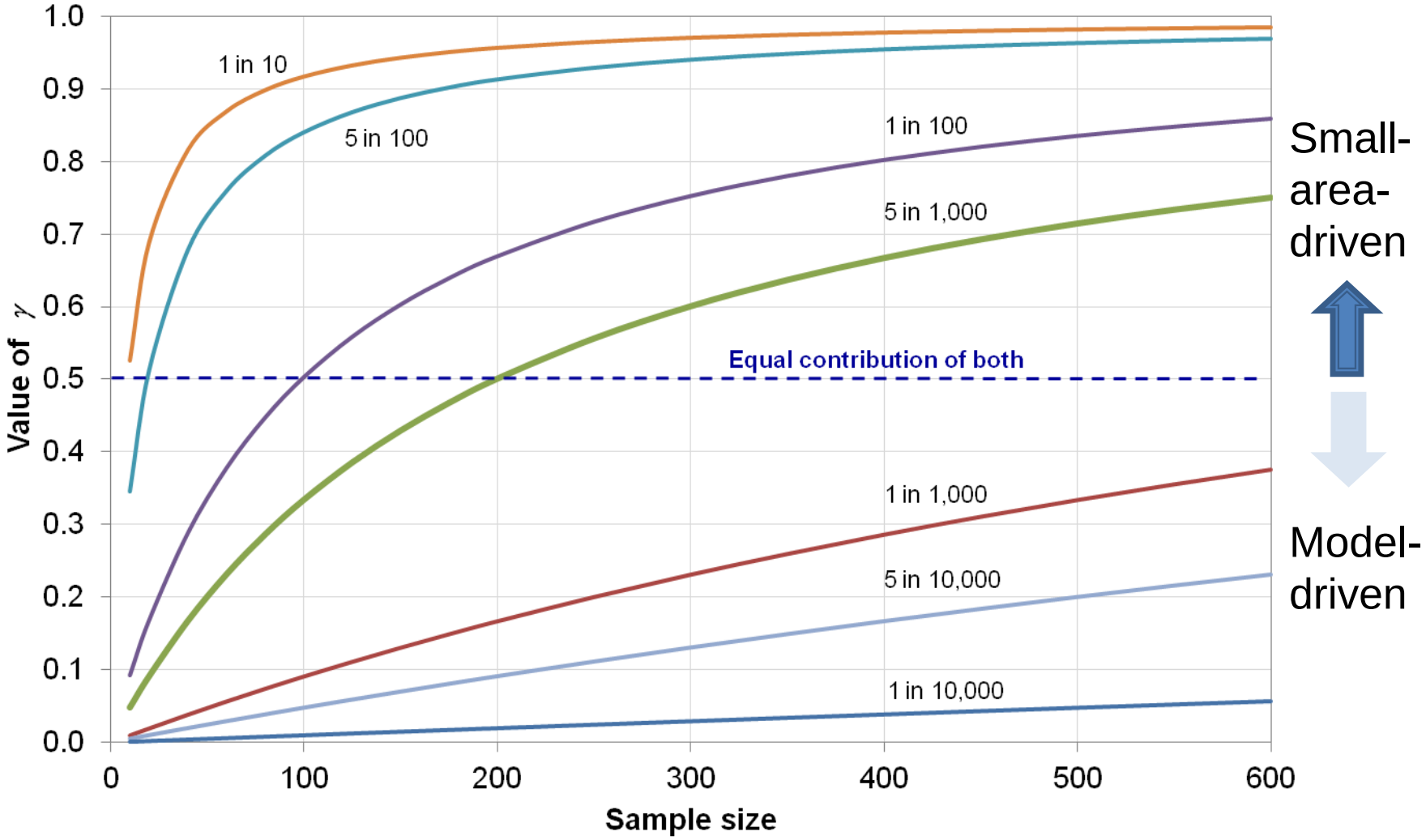
A **COMPOSITE** estimator

$$\begin{aligned} \hat{\bar{Y}}_g^{EBLUP} &= \bar{\mathbf{X}}_g^T \tilde{\boldsymbol{\beta}} + \hat{v}_g \\ &\approx \bar{\mathbf{X}}_g^T \tilde{\boldsymbol{\beta}} + \hat{\gamma}_g \left( \bar{y}_g^{\text{SurvReg}} - \bar{\mathbf{X}}_g^T \tilde{\boldsymbol{\beta}} \right) \end{aligned}$$

$$\text{where} \quad \hat{\gamma}_g = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_g}$$

$$\tilde{\boldsymbol{\beta}} = \text{BLUE of } \boldsymbol{\beta}$$

# *It doesn't take much of an area effect to be useful: effect of ICC*



# Unit level GLMM model – EBP

$$\text{logit}\left(y_{ig}\right) = \mathbf{x}_{ig}^{T}\boldsymbol{\beta} + v_g$$

$$v_g \overset{iid}{\sim} \left(0, \sigma_v^2\right)$$

$$i = 1, \ldots n_g \quad g = 1, \ldots G$$

$$\hat{\bar{Y}}_g^{EBP} = \frac{\exp\left(\bar{\mathbf{X}}_g^{T}\tilde{\boldsymbol{\beta}} + \hat{v}_g\right)}{1 + \exp\left(\bar{\mathbf{X}}_g^{T}\tilde{\boldsymbol{\beta}} + \hat{v}_g\right)}$$

# Many model options considered

- 4 outcome variables x sex
  - predictions for 3 years for each
- 4 sets of model-based estimates
  - Linear and logistic
  - With and without random effect term
- 6 covariate patterns

# Potential Covariates for NSW Health Survey data

- Marital status ☺
- Number of residents in household 😐
- Sex ☺
- Age ☺
- Number of children – not all years
- Country of Birth ☺
- Other language spoken at home ☺
- Level of schooling 😐– highest level of qualification
- Aboriginal status ☺ - *limited accessibility*
- Employment status ☺
- Pension status
- Private health insurance
- Household income (broad bands) ☹ - quality very poor

☺ = can be compared with Census

😐 = Not exactly same between survey and census

# Covariate specifications tested

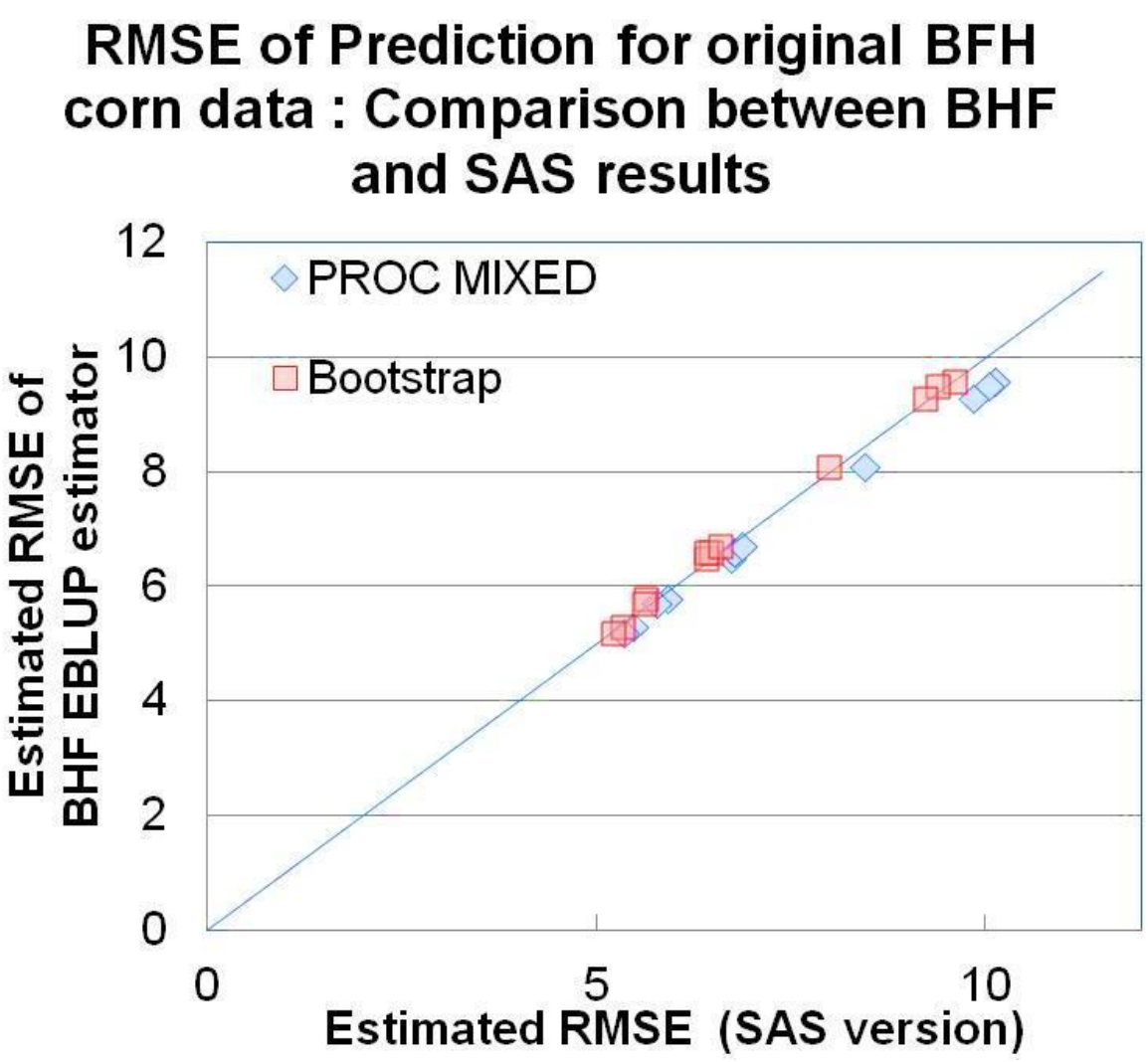
| Abbrev | Covariates included |
|---|---|
| Null | Intercept only model |
| Age | 10-year age group only |
| ONS (UK) | 10-year age group, health area and quintile of disadvantage |
| Outcome-specific | Covariates that are significant in the majority of years for that sex/outcome |
| Global | All covariates included in any of the Common models for the four outcome variables |
| Specific | Covariates that are significant following stepwise regression for the specific sex/ outome/ year/ model |

# Observed values of ICC

| Person-level data, 2006 | Null model | Age, sex | Age, Sex, Health area | Age, sex, AHS+ other covars |
|---|---|---|---|---|
| Risk Alcohol | 1.5% | 1.9% | 1.1% | 0.4% |
| Current smoking | 0.9% | 0.9% | 0.5% | 0.1% |
| Difficulties getting health care when needed | 8.6% | 8.7% | 4.4% | 2.5% |

# SAS-based RMSEs compare well

- Validated against Battese Harter Fuller (1988) results
- Parametric bootstrap also works and may be better when estimating for unsampled areas (another talk)



# Results

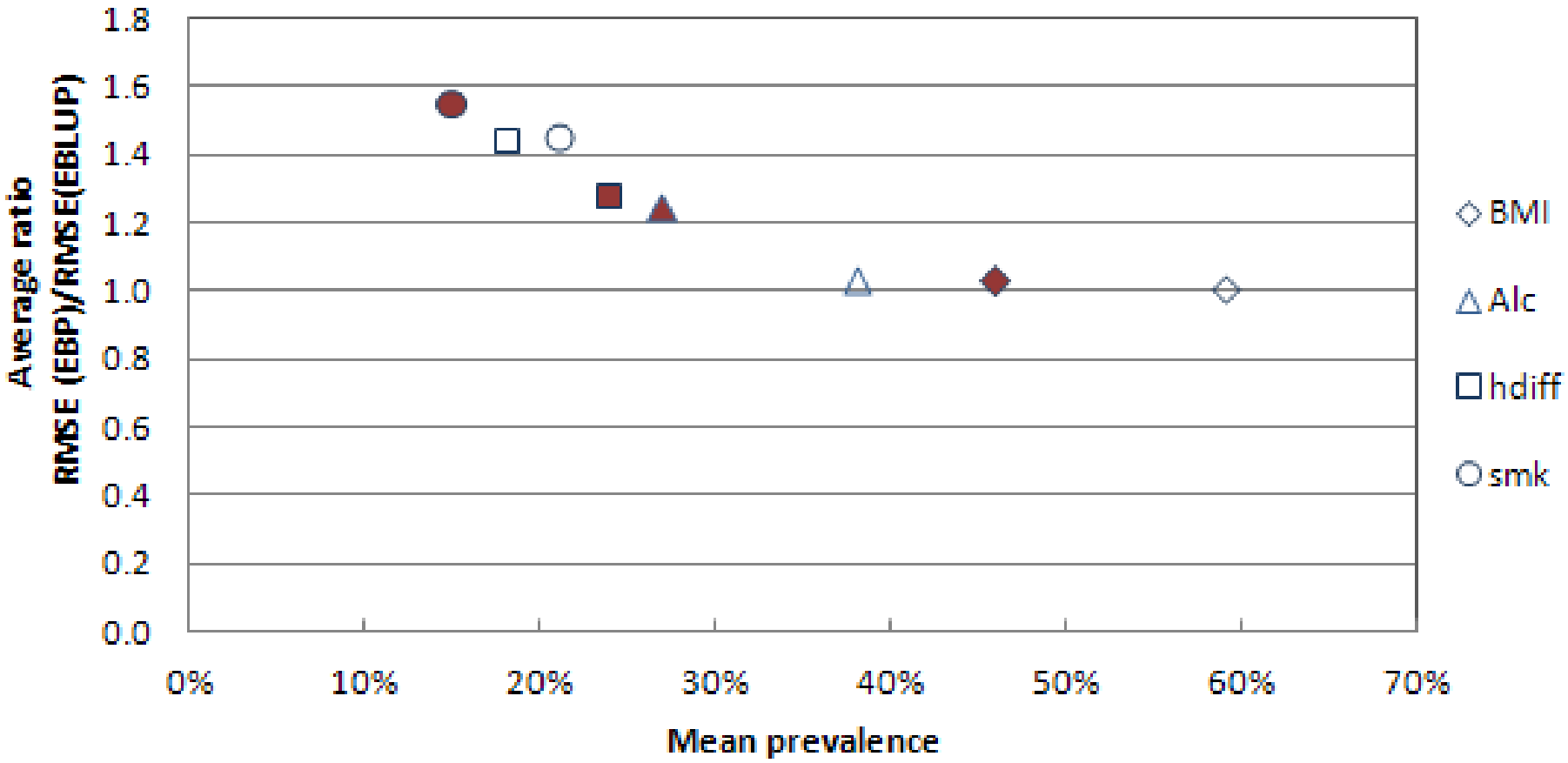
Allowing the data to speak through judicious use of models

# Linear vs Logistic

- Linear and Logistic give similar *estimates*
- Estimated RMSE: depends on prevalence
  - Linear approx to logistic works well for estimated prevalence between 30% and 70% - use either
  - Below 30% (also above 70%) more important to use logistic – linear model may underestimate RMSE

- Use of Logistic for all variables will ensure more appropriate estimated RMSEs but provide similar estimates

# Logistic not necessary 30-70%




**Ratio of max estimated RMSE of logistic EBP to max estimated RMSE of linear EBLUP vs prevalence, by outcome and sex**

Note: solid symbols denote female, open symbols denote male.

# Validation

- What to compare with?

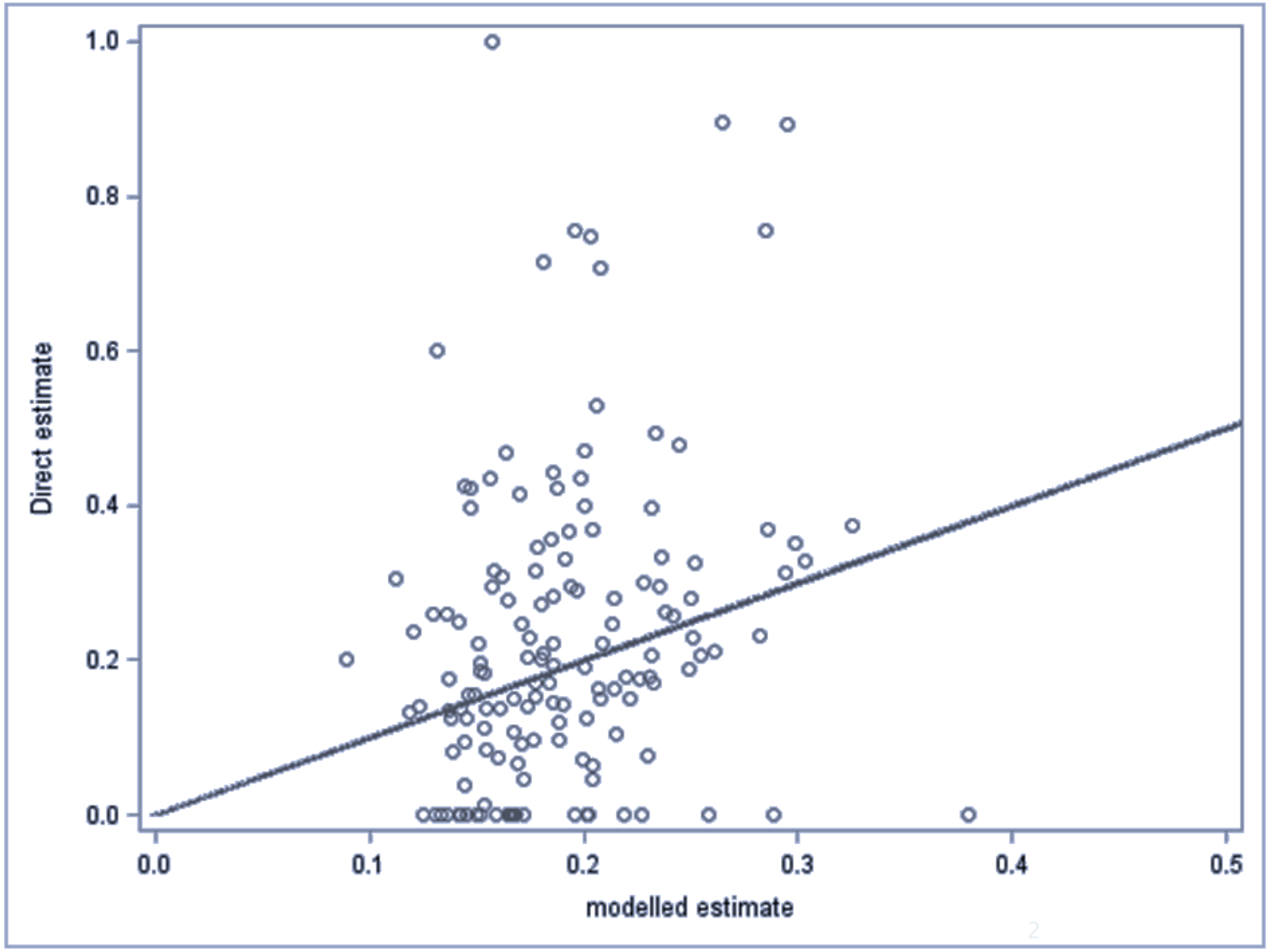
There is no GOLD standard

- *Model-based* estimate (X) vs direct estimates (Y) should be consistent with line of identity
- Difference between modelled and direct estimates should decrease as n increases

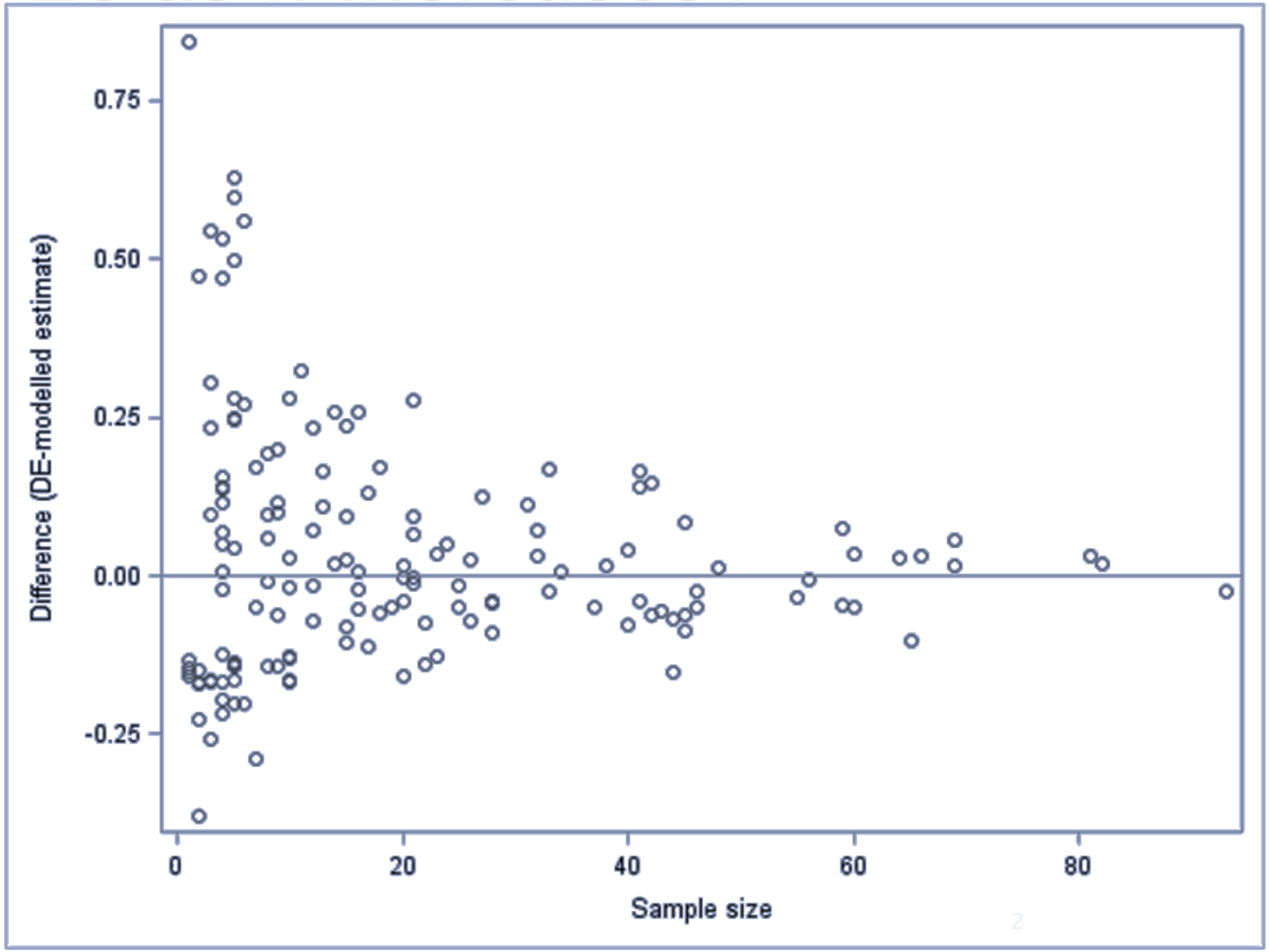
(Brown et al, 2001)

- Weighted mean of the model-based estimates compare well against direct estimates at higher aggregation levels

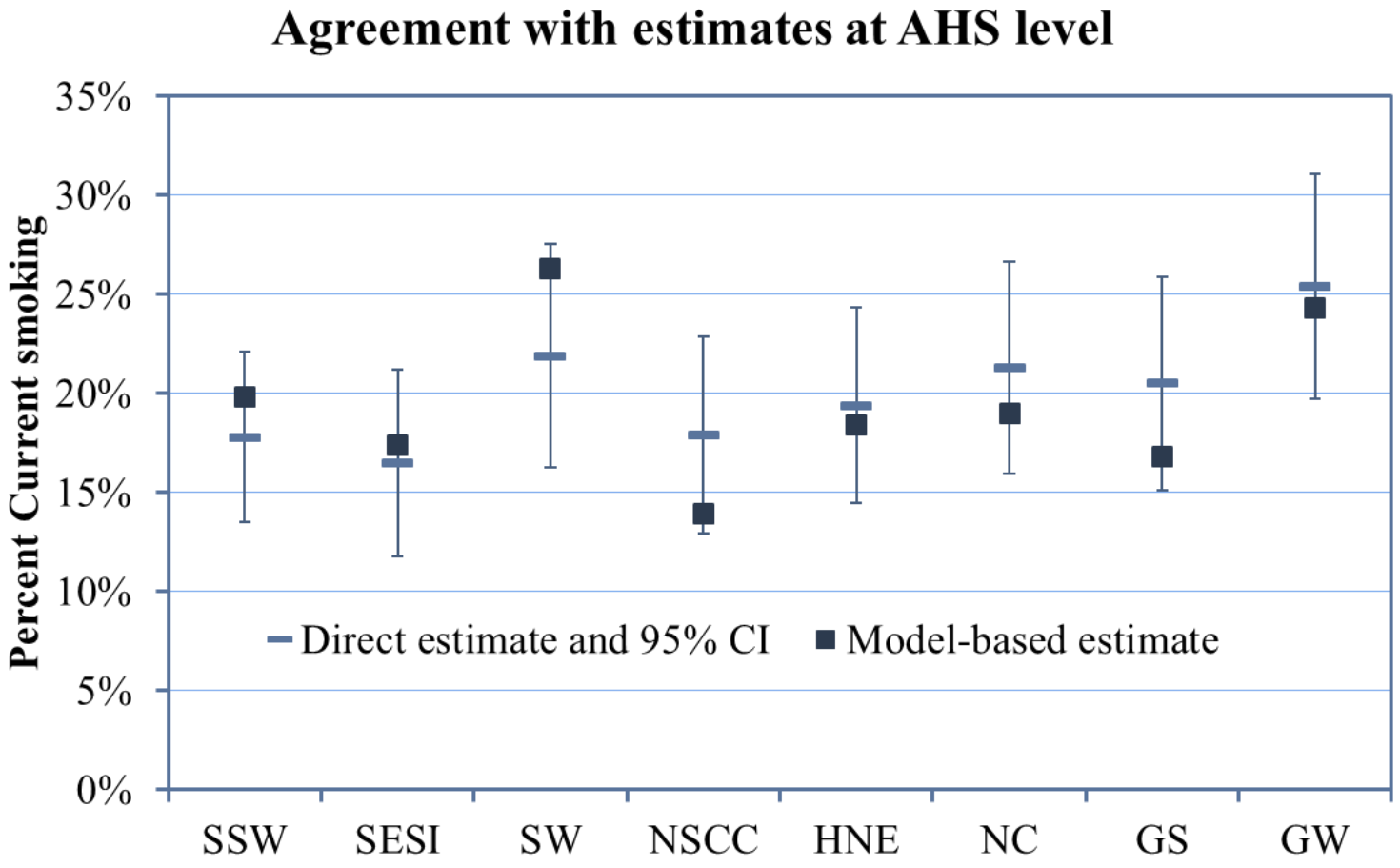
# Model-based vs direct estimate consistent with identity line



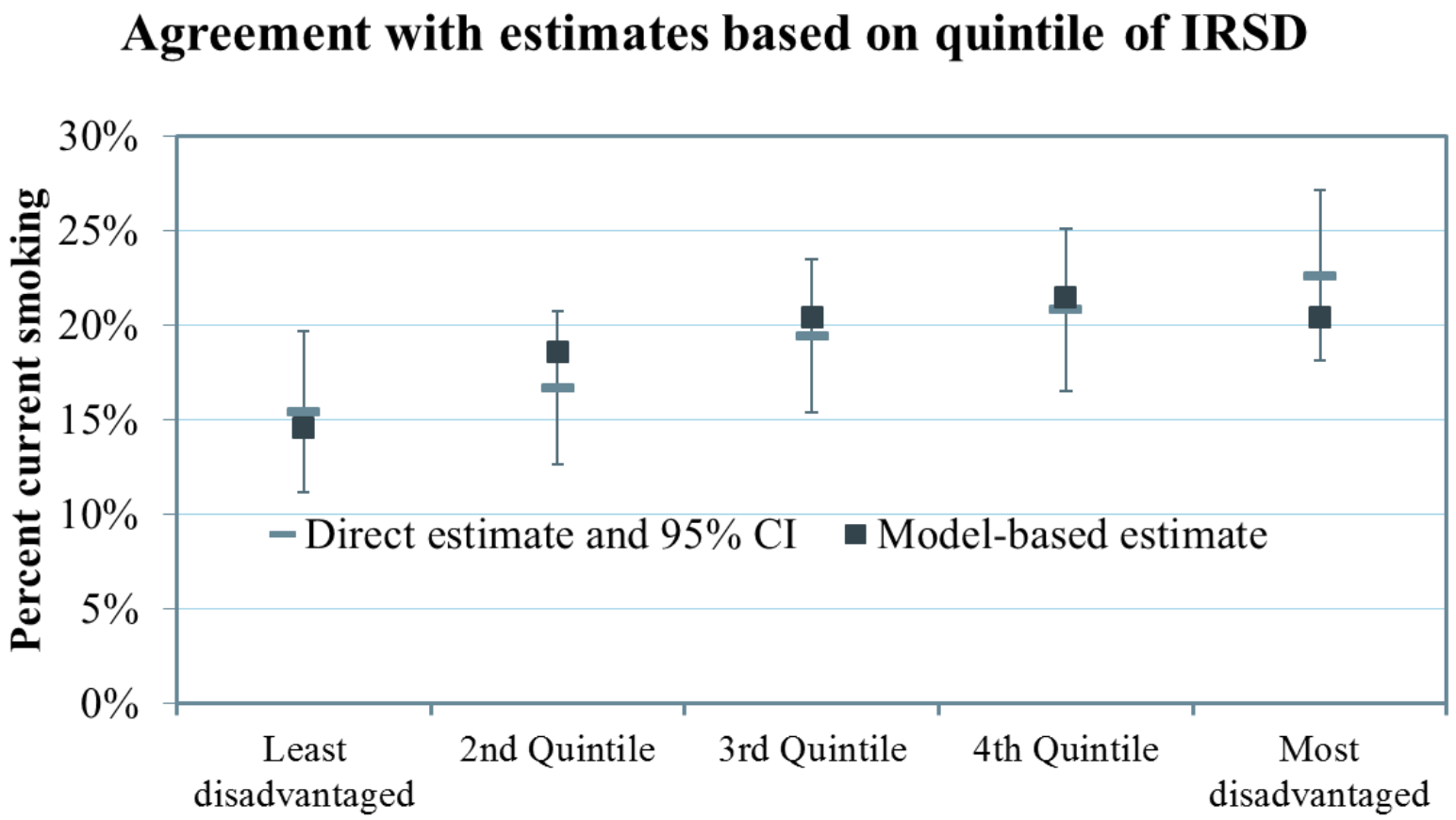
# Model-based estimates approach DEs as *n* increases?



# Agreement with estimates at Health Area level



# Agreement with estimates based on quintile of Relative Socioeconomic Disadvantage



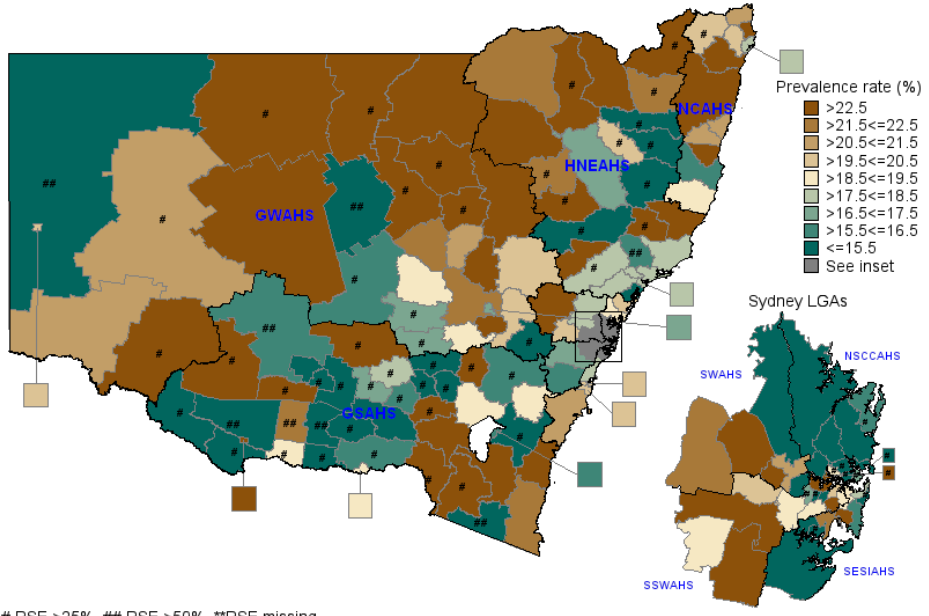
# Conclusion

"The demand for SAEs is strong and increasing, and models are needed to satisfy that demand in many cases" (Kalton, in Rao (2003))

# Conclusion

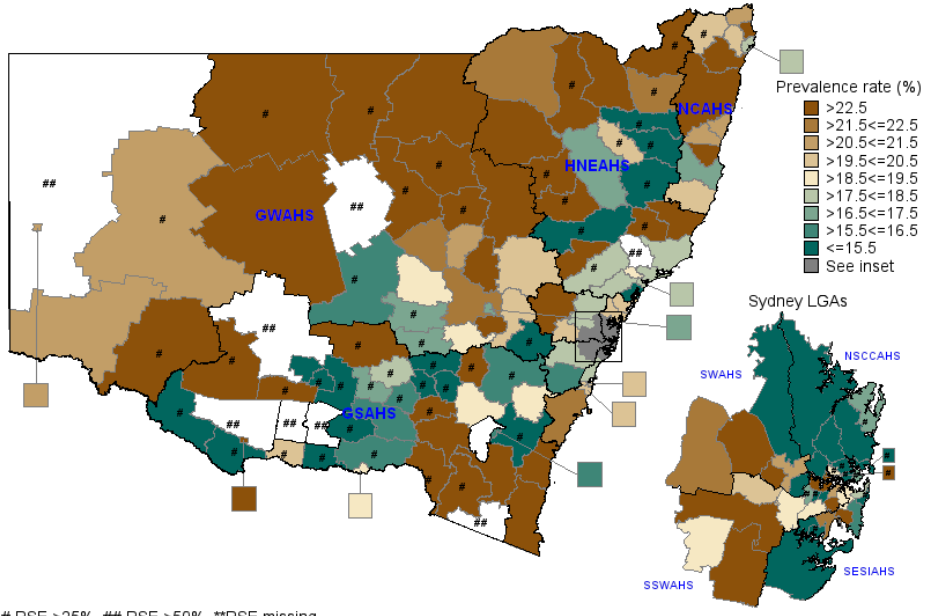
- Model-based methods create **reasonable estimates despite** difficulties in finding covariates that are **good predictors of the response variable**
- **EBP estimates** are most appropriate for binary outcome variables
- Model-based estimates from one year of data have lower maximum RMSE and RRMSE than direct estimates using multiple years of data.

## Current Smoking, 2002-2008, Male
Direct estimates



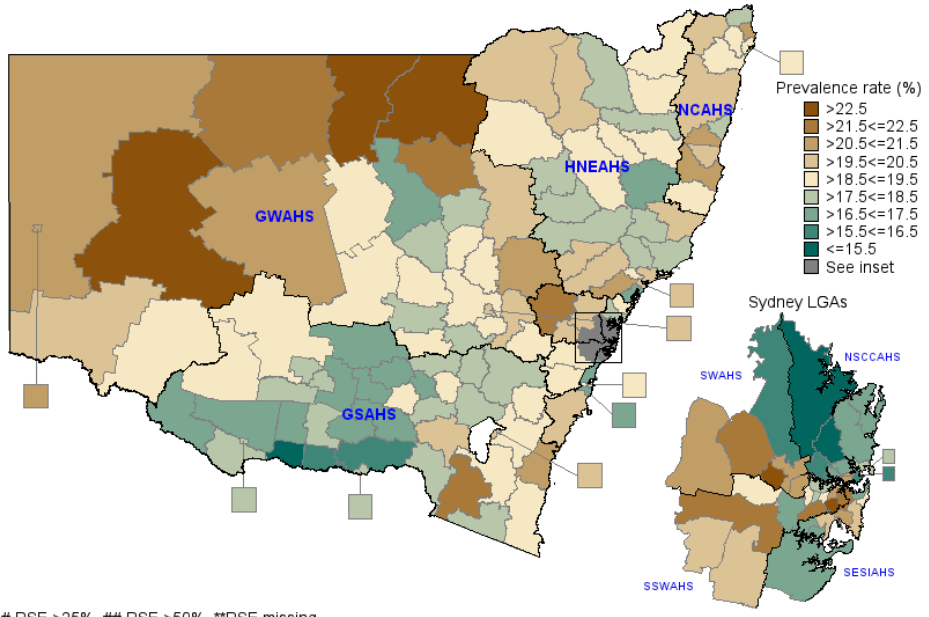

# RSE >25%, ## RSE >50%, **RSE missing

## Current Smoking, 2006, Male
Direct estimate based on 2002-2008 combined, where RSE < 50%



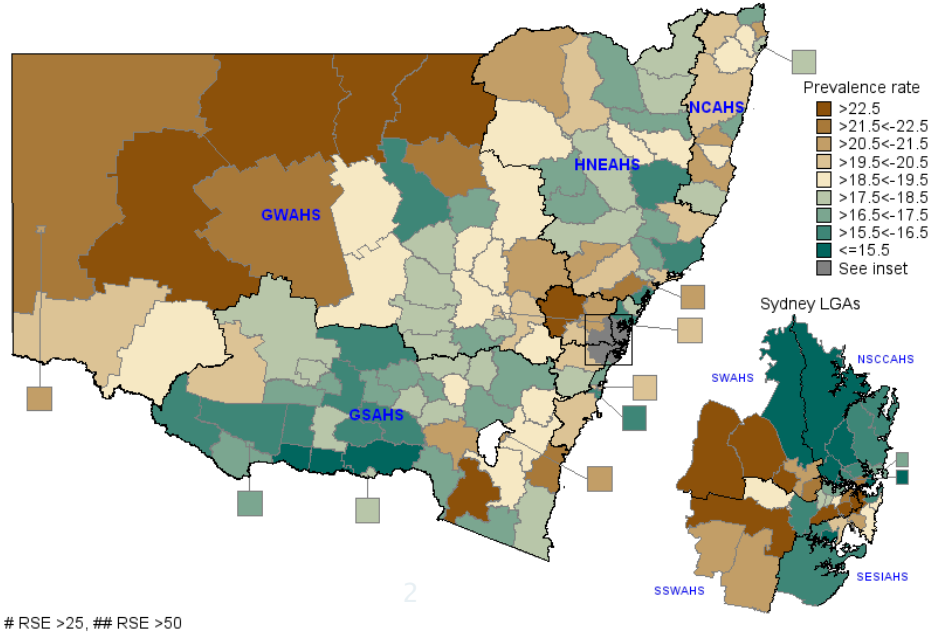

# RSE >25%, ## RSE >50%, **RSE missing

## Current Smoking, 2006, Male
Linear EBLUP




# RSE >25%, ## RSE >50%, **RSE missing

## Current Smoking, 2006, male
Calibrated Logistic with Random effect




# RSE >25, ## RSE >50

# Thanks

Sincere appreciation to

- NSW Population Health Survey Program, Centre for Epidemiology and Research, NSW Health for allowing access to NSW Health Survey data for this project.
- ARC, NSW Health, NZ Ministry of Health and ABS for funding this project through an ARC Linkage Grant
- Professor David Steel
- Professor Ray Chambers.

# Contact details

diane.hindmarsh@health.nsw.gov.au

# References


- **Battese, GE, Harter, RM and Fuller, WA (1988)**, *Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data* Journal of the American Statistical Association **83** 28-36
- **Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001)** Evaluation of small area estimation methods - an application to unemployment estimates from the UK LFS. In Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Proceedings of Statistics Canada International Symposium. StatisticsCanada.
- **Department of Human Services (2005)** Victorian Burden of Disease Study. Mortality and Morbidity in 2001. Public Health Group, Department of Human Services, Victoria, Australia.
- **Mukhopadhyay, PK and McDowell, A (2011)** *Small Area Estimation for Survey Data Analysis Using SAS Software*, SAS Global Forum 336-2011, SAS Institute Inc, Cary, NC http://support.sas.com/resources/papers/proceedings11/336-2011.pdf
- **Rao, JNK.** *Small Area Estimation.* NJ, Wiley, 2003
- **Office of National Statistics (2008)** First Release, Labour Market Statistics, May 2008. Tech. Rep. 14 May 2008, Office of National Statistics.
- **Scholes, S., Pickering, K. and Rayat, P. (2007)** Healthy lifestyle behaviours: model based estimates for Middle Layer Super Output Areas and Local Authorities in England, 2003–2005. User guide, National Centre for Social Research and NHS-IS.