USE OF COMPOSITE VARIABLES: AN EPIDEMIOLOGICAL PERSPECTIVE

David Coggon

MRC Lifecourse Epidemiology Unit

UNIVERSITY OF Southampton School of Medicine

EPIDEMIOLOGY

- Investigates the distribution of health and its determinants
- Descriptive studies describe the distribution of health-related variables in defined populations at defined times • Analytical studies explore the interrelationship of health-related variables (causation or prediction)

SPECIFICATION AND MODELLING OF

VARIABLES

- Both types of study require specification of the variables to be analysed
- May be measured directly (e.g. weight, peak expiratory flow, blood lead)
- Often are composites of two or measures (e.g. many case definitions, incidence, BMI, forced expiratory ratio)
- Analytical studies may also entail statistical modelling

MRC Lifecourse Epidemiology Unit

UNIVERSITY OF Southampton School of Medicine

STATISTICAL MODELLING

• Makes assumptions about nature of relationships between variables • Assumed relationships should have conceptual validity

• Validity can be explored empirically • In some cases, the use of composite variables can be considered a form of modelling

REASONS FOR USING COMPOSITE VARIABLES

- Simple scaling (e.g. incidence, prevalence)
- Best practicable proxy for a gold standard (e.g. case definition for myocardial infarction)
- No gold standard available, but composite has conceptual and possibly predictive validity (e.g. scales for depression, somatising tendency, ERI, job strain)

PROS AND CONS OF COMPOSITE VARIABLES

- Composite variables can simplify analysis, and are of proven utility
- But combining primary measures in composite variables may lose useful information (e.g. BMI, ERI)
- Use of ratios has been a particular concern

PROBLEMS WITH RATIOS

- Pearson (1897) referred to “spurious correlation” that can occur between ratios even if all of the component variables of the ratios are uncorrelated

- Neyman (1952) constructed fictitious example in which number of storks per 10,000 women and number of babies per 10,000 women was highly correlated across a sample of counties, although within counties having same numbers of women number of storks was unrelated to number of babies (counties with most women tended to have fewer storks/10,000 women and fewer babies/10,000 women)

RECOMMENDATIONS OF KRONMAL (1993)

- “The use of ratios in regression analyses should be avoided”

- “Ratios should only be used in the context of a full linear model in which the variables that make up that ratio are included and the intercept term is also present”

SOME THOUGHTS ON KRONMAL

- Problems identified reflect sub-optimal specification of models and are not limited to ratios.
- They could apply to products (X*Y = X/[1/Y]), and to compound variables more generally
- Any variable can be expressed as a ratio of two other variables (X = [X/Y] / [1/Y])
- Including a term for each variable and intercept may still not be optimal (e.g. power functions might be relevant)
- Potential disadvantages of over-elaborate models,

especially if data-driven

DECISIONS IN DESIGN AND ANALYSIS

- Data to be collected
- Specification of variables from those data (may be directly measured or a function of what is measured)

- Specification of statistical models

HOW SHOULD DECISIONS BE MADE?

• No universal rules, only guiding principles • Prime consideration is study question and conceptual understanding

• Collection of data should be practicable and efficient •

Variables analysed should have validity (conceptual, repeatability, predictive, against a gold standard) • Consistency with other research

• Statistical models should be conceptually valid but parsimonious

• Where optimal choices are uncertain, can explore alternatives, but data-driven models may in part reflect random sampling variation

Ultimately, what matters is practical utility

FACTORS FAVOURING USE OF COMPLEX VARIABLES

- Strong conceptual grounding, ideally with empirical evidence of validity

• Parsimony without undue loss of validity • Compatiblility with other research (meta analysis)