SSM - Population Health 3 (2017) 600–608

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

Article

Model-based and design-based inference goals frame how to account for neighborhood clustering in studies of health in overlapping context types

Gina S. Lovasi[a,*], David S. Fink[b], Stephen J. Mooney[c], Bruce G. Link[d]

[a] Drexel University, 3600 Market Street, Office 751, Philadelphia, PA 19104, United States
[b] Columbia University, 722 West 168th Street, Room 724, New York, NY 10032, United States
[c] Harborview Injury Prevention and Research Center, 401 Broadway, 4th floor, Seattle, WA 98122, United States
[d] University of California Riverside, U4649 9th Street, Riverside, CA 92501, United States

ARTICLE INFO

Keywords:
Multilevel analysis
Epidemiologic measurement
Health surveys

ABSTRACT

Accounting for non-independence in health research often warrants attention. Particularly, the availability of geographic information systems data has increased the ease with which studies can add measures of the local "neighborhood" even if participant recruitment was through other contexts, such as schools or clinics. We highlight a tension between two perspectives that is often present, but particularly salient when more than one type of potentially health-relevant context is indexed (e.g., both neighborhood and school). On the one hand, a model-based perspective emphasizes the processes producing outcome variation, and observed data are used to make inference about that process. On the other hand, a design-based perspective emphasizes inference to a well-defined finite population, and is commonly invoked by those using complex survey samples or those with responsibility for the health of local residents. These two perspectives have divergent implications when deciding whether clustering must be accounted for analytically and how to select among candidate cluster definitions, though the perspectives are by no means monolithic. There are tensions within each perspective as well as between perspectives. We aim to provide insight into these perspectives and their implications for population health researchers. We focus on the crucial step of deciding which cluster definition or definitions to use at the analysis stage, as this has consequences for all subsequent analytic and interpretational challenges with potentially clustered data.

1. Background

Human experience takes place in multiple overlapping contexts, including geographic contexts such as neighborhoods and cities, organizational contexts such as schools and clinics, and social contexts such as families and friendship networks. Though the variability of health-relevant exposures and outcomes within and between these contexts has long been a focus of study (Mooney, Knox, & Morabia, 2014; Morabia, 2014; Pincus & Stern, 1937), in recent years, research teams have increasingly had opportunities to link measures from more than one type of context within the same study population (Box 1).

The integration of multiple context types into our research reflects the multiplicity of overlapping contexts that shape our social experience and related health risks. Health-relevant sorting into neighborhoods (Bischoff & Reardon, 2013), schools (Reardon & Owens, 2014), clinics (Sarrazin, Campbell, Richardson, & Rosenthal, 2009), and workplaces (Goh, Pfeffer, & Zenios, 2015) has been well-documented in the literature, complicating our ability to study the implication of changing such contexts for our health. Beyond physical contexts there are social networks and affinity groups that affect the health of individuals. The numerous overlapping contexts in which individuals are embedded result in correlations within "clusters" (a term that we will use for brevity to indicate the spatial units, institutional settings, or other macro-units to which individuals in a study population are indexed via a cluster identifier). The availability of repeated measures over time in longitudinal studies bring further complexity as well as value (Leckie, 2009). One or more of the clusters may take on particular salience because of the study design or context characteristics available for linkage (Fig. 1). Doing so makes salient the often implicit tensions between two inferential perspectives labeled as model-based and design-based.

This paper identifies two common perspectives and their implications when considering a clustering-based analytic approach (e.g. by using random effects or cluster robust standard errors) for studies linking context to health. As such analytic approaches have become easier to implement in standard statistical software (Diez Roux, 2000;

* Corresponding author.
E-mail addresses: gsl45@drexel.edu (G.S. Lovasi), dsf2130@columbia.edu (D.S. Fink), sjm2186@u.washington.edu (S.J. Mooney), bruce.link@ucr.edu (B.G. Link).

http://dx.doi.org/10.1016/j.ssmph.2017.07.005
Received 1 February 2017; Received in revised form 17 July 2017; Accepted 18 July 2017
2352-8273/ © 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Box 1
The Role of the Intra-Class Correlation in Selecting a Cluster Definition.

Either perspective might turn to tools such as the intra-class correlation (ICC) to quantify how distinct clusters are with respect to the health outcome of interest (Merlo, Chaix, Yang, Lynch, & Rastam, 2005). An ICC that is distinct from zero (or one with a 95% confidence interval that excludes zero) might be used to justify a particular cluster definition by some with a model-based perspective (Snijders & Bosker, 2012b, 2012c, 2012d). The ICC may also be used by those with either perspective to point to areas of potential interest for substantive investigation. Because the ICC correlation within clusters, any decision that relies on ICC values may be reversed when the same study data are used to investigate a different outcome. Similarly, a pilot study and a full scale study might reach different conclusions about the need to account for clustering, even though each used the same sampling strategy to study the same outcome. Thus, the design effect will depend not only on the sampling strategy but also on the population sampled and the outcome itself. For example, among students there may be stronger clustering of standardized test scores by classroom due to influence of the teacher and of processes by which students of similar abilities are assigned to the same classroom, whereas school-level clustering may be stronger for physical activity outcomes due to shared physical fitness facilities and physical education policies. To visualize this phenomenon, it may help to consider a sparse sample (Fig. 2b), where the social contexts will often be unique or shared by few participants. Low power would result in wide confidence bounds around the ICC, and we would be unlikely to exclude zero (however, even when power is low to detect whether the ICC for the outcome is distinguishable from zero, there may be sufficient power to detect an association with one of the measured cluster characteristics). By contrast, in a dense sample (Fig. 2c) there is greater statistical power to distinguish the ICC from zero, and more opportunity to investigate the contributions of both measured and unmeasured characteristics of shared environments.

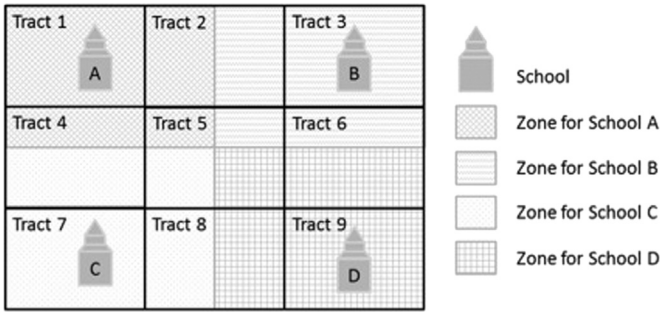

Fig. 1. A schematic diagram of overlapping sources of clustering. Subjects recruited from schools A and D are both clustered in schools and in an overlapping subset of census tracts. Which, if any, of these clustering sources does an analyst need to account for? Notes: This study recruited students from schools A and D, then measured neighborhood conditions in census tracts referring to students in those tracts (1, 2, 4, 5, 6, 8, and 9). Does the analyst need to account for clustering on tracts, on schools, or both? How should we decide, noting that the clusters are overlapping and not hierarchical? A design-based perspective would emphasize the recruitment setting, indicating that inference about students in general must account for clustering of students within schools. A model-based perspective would emphasize whether clustering is important to approximating the probability model generating the observed data.

Singer, 1998), how specifically to analyze clustered data, and whether hierarchical or cross-classified techniques are truly necessary, should be considered carefully (Mitchell, 2001). Attention to what have been called "model-based" and "design-based" inference goals (Snijders & Bosker, 2012c; Sterba, 2009), and the tensions between perspectives and within each perspective, can elucidate how we decide on which cluster definition (or definitions) to account for, a decision that in turn affects all subsequent analytic and inferential steps. We aim to provide insight into these perspectives and their implications for an applied population health research audience. We first discuss distinguishing features of each perspective, and then turn to how they offer divergent guidance under the increasingly common circumstance of having more than one type of context available to account for non-independence (Fig. 1).

Consider, for example, an investigation of swimming skills (Hulteen et al., 2015) among children in a given city, with relevance to both physical activity (Fisher et al., 2005) and drowning risk (Brenner et al., 2009). The investigative team systematically samples schools within the city, and then children within those schools, such that sampling probabilities are known. Suppose also that residents of some neighborhoods have received frequent marketing of private swimming lessons at their local swimming pool (for the sake of illustration, we suppose this is unmeasured, as would often be the case for local social norms or other behaviorally-relevant characteristics of context). Empirically, it might be that residual clustering in the outcome is greater based on neighborhood than by school. Exposures of interest addressed by the investigative team across several empirical manuscripts are defined at the individual (e.g., gender), school (e.g., physical education hours/week), and neighborhood level (e.g., area-based socioeconomic indicators). A team that adopts a design-based perspective would be attentive to sampling weights and inference to the city population, but might not require adding a random effect to account for within-neighborhood clustering because that clustering is a reflection of the clustering truly present in the city (rather than being investigator-imposed). By contrast, the model-based perspective would primarily be focused on specification of the model, accounting for neighborhood clustering if the processes shaping the skills of two children within the same neighborhood are not considered independent; the model-based team might consider an unweighted analysis using adjustment as a possibly more efficient alternative to a weighted analysis. Both perspectives are flexible, and ideally the advantages of each will be considered, but we posit that being able to name and distinguish them will help to avoid confusion.

2. Distinguishing features of, and selected tensions within, a model-based perspective

A model-based perspective emphasizes the processes producing outcome variation, and observed data are used to characterize that data-generating process. This is the majority perspective in statistical textbooks, including those focused specifically on multi-level modeling (Snijders & Bosker, 2012d). Attention is paid to minimizing bias and maximizing efficiency, and if weighting is used it is often for these purposes. Crucially for the topic at hand, a model-based inference perspective primarily considers independence of observations with respect to residual correlations in the observed data. Measured cluster-level characteristics may be of interest to explain such residual correlations, in which case a model structure is specified and the parameters estimated accordingly, or there may simply be an interest to account for the variance structure because the assumption about observations' independence does not hold.

Even before we consider the contrast with a perspective focused on design-based inference, it is worth mentioning two tensions among those seeking to make model-based inference. First, in decisions on whether to condition on clusters, some may prefer a strategy specified *a priori*, while others may look to the data to guide the structure of the

Download English Version:

https://daneshyari.com/en/article/5123292

Download Persian Version:

https://daneshyari.com/article/5123292

Daneshyari.com