Family Practice, 2020, 719–722
doi:10.1093/fampra/cmaa017
Advance Access publication 27 February 2020

CASFM Methods Briefs

An introduction to clustered data and multilevel analyses

Jessalyn K Holodinsky[a], Peter C Austin[b,c,d] and Tyler S Williamson[e,f,g,*]

[a]Evaluative Clinical Sciences Platform, Sunnybrook Research Institute, Toronto, ON, Canada, [b]ICES, Toronto, ON, Canada, [c]Institute of Health Management, Policy, and Evaluation, University of Toronto, Toronto, ON, Canada, [d]Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, ON, Canada, [e]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, [f]O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada and [g]Alberta Children's Hospital Research Institute for Child and Maternal Health, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

*Correspondence to Tyler S Williamson, Department of Community Health Sciences, Centre for Health Informatics, Cumming School of Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada. Email: tyler.williamson@ucalgary.ca

What are clustered data?

Clustered data arise when the subjects are physically grouped into different groups (or clusters), with at least some of the groups containing multiple subjects (this grouping can be due to things like geography or through a shared relationship, such as with a family doctor). This gives the data a multilevel structure in which subjects are nested within these clusters or groups. An example being: patients clustered within family physicians. The multilevel structure can have more than two levels: patients are clustered within family physicians who in turn are clustered within a clinic (Fig. 1). This can extend for many levels, as clinics may be further clustered within a city, health system or state. For this article, we will restrict our discussion to multilevel structures with two levels—patients clustered within family practices; however, the concepts and methods extend to those with more than two levels.

Consequences of clustered data

The presence of clustering induces additional complexity, which must be accounted for in data analysis. Outcomes for two observations in the same cluster are often more alike than are outcomes for two observations from different clusters, even after accounting for patient characteristics. This within-cluster homogeneity in outcomes violates the assumption of most regression models that the observations are independent. Multilevel analyses allow for the appropriate analysis of data with multilevel structure where there is no longer independence among observations (1).

Using a traditional regression method, when the assumption of independence is violated, the estimation of regression coefficients and their associated standard errors can be biased (2). Treating group-level variables as though they are measured at the individual level can lead to standard errors being underestimated, which can lead to erroneously significant results and artificially narrow confidence intervals (3). This is partly due to sample size inflation problems resulting from failure to account for multilevel data structure (4). Falsely treating individuals as independent erroneously increases the precision of estimates made because of erroneously increasing the degrees of freedom in the analysis. Ignoring the nested data structure may result in relationships being found to be significant when they truly are not—this is known as *misestimated precision* (5).

Issues related to ignoring multilevel structure can occur *even if the group level factors are not a part of your research question* (6). If your data arose from a multilevel structure, it is important to take this into account in your analysis regardless of the research question at hand.

The intraclass correlation coefficient

A key concept in multilevel analysis is the intraclass correlation coefficient (ICC). The ICC quantifies the proportion of the variation in the outcome that can be attributed to systematic differences in the outcome between clusters (6). The ICC tells you the degree of similarity between individuals belonging to the same group. From an epidemiologic perspective, this allows one to ascertain how much of the variability in outcome can be attributed to the clustering unit (1). For example, how much variability in patient outcomes can be attributed to the hospital to which they were admitted. If there is no systematic between-hospital variation in patient outcomes, we can say that there is an absence of 'hospital effect' on patient outcomes.

The ICC is calculated by dividing the between-cluster variation in the outcome by the total variation in the outcome—similar to the process of comparing the between and within group variances

in analysis of variance. The ICC is equal to the correlation between two individuals drawn from the same group, and it can range from 0 to 1. If it is 0, there is no evidence of clustering effects in the data. If the ICC is 1, then the grouping accounts for all the variation in the data, meaning all individuals within the same group have identical responses on the outcome variable. The ICC is rarely ever 0 or 1. In our experience, values are commonly <0.10, although plausible or realistic values for the ICC will vary according to the measure and to the type of clustering that is present.

Multilevel analysis

Multilevel analysis allows for more than just accurate estimation of regression coefficients and standard errors due to non-independence and quantification of between-cluster variation (the ICC). As variables can be measured at different levels of the hierarchy, it allows for correct inferences about cluster-level variables to be made. Additionally, the magnitude of the association between variables and the outcome can be allowed to vary between clusters, which is something that cannot easily be handled by traditional regression techniques (6).

Consider a study examining patient appointment lengths for their first visit to a family physician, which varies both within and across family practices. The unit of analysis is the patient visit, and the outcome (appointment length) is measured at the level of the patient or patient visit. The study included 100 patient visits from each of 10 different physicians. Using a traditional regression model, we would conclude that older patients have, on average, longer appointments (as in Fig. 2—top panel). However, traditional regression analysis ignores the fact that patients seen at the same family practice are not independent of one another—they have the same physician, which would almost certainly impact appointment length. Using a multilevel analysis would allow for the relationship between patient age and appointment length to be shown in the context of the individual physicians. Allowing for the exploration of between-physician differences in the baseline appointment lengths allows for richer conclusions to be drawn from the data. Perhaps physicians who have been practicing for several years have different baseline appointment lengths than their less experienced counterparts, but overall older patients tend to have longer appointments (as in Fig. 2—middle panel). Models that allow intercepts to vary but retain common slopes across groups are known as random intercepts models (4).

Considering a potentially more complex relationship, suppose more experienced physicians take on more complex patients. As such, their baseline appointment length is longer than less experienced physicians, but their appointment length varies very little with patient age. The less experienced physicians, which have a shorter baseline appointment length, exhibit a strong relationship between appointment length and patient age (as in Fig. 2—bottom panel). Multilevel analyses can simultaneously account for the difference in baseline appointment length as well as the different relationship between patient age and appointment length across physicians. Models that allow both the intercepts and slopes to vary across groups are known as random slopes models (4).

The above example begs the question, 'could I have achieved these results using simple linear regression adjusting for "physician" using a dummy variable?' This would be analogous to fitting separate regression lines characterizing the relationship between appointment length and age for each physician. There are a few difficulties with this approach. First, the use of dummy variables is statistically inefficient, especially if the number of groups is large. Second, the use of dummy variables at the group-level negates the ability to analyze group-level factors such as the years of experience of the physician, size of practice, sex of the physician etc. as these factors would be compressed into the physician identifier.

We have provided an example with a continuous outcome; however, there are multilevel analogues for binary outcomes, counts, multinomial outcomes and survival data. There are also other ways in which data can have multilevel structure: longitudinal data (repeated observations nested within each patient), cross-classified data (patients are nested in physicians and patients are nested in neighbourhoods, but physicians are not necessarily nested within the same neighbourhood as the patient) and cluster randomized trial data. More in-depth discussions of multilevel modelling and applying these techniques to healthcare data are found in several other texts (7–12).

We have focused this discussion on multilevel analyses in which the level of analysis is the lowest level of the cluster. Another method for analyzing clustered data is population-average models estimated using generalized estimating equation methods, which are primarily concerned with population-average effect of a covariate rather than the

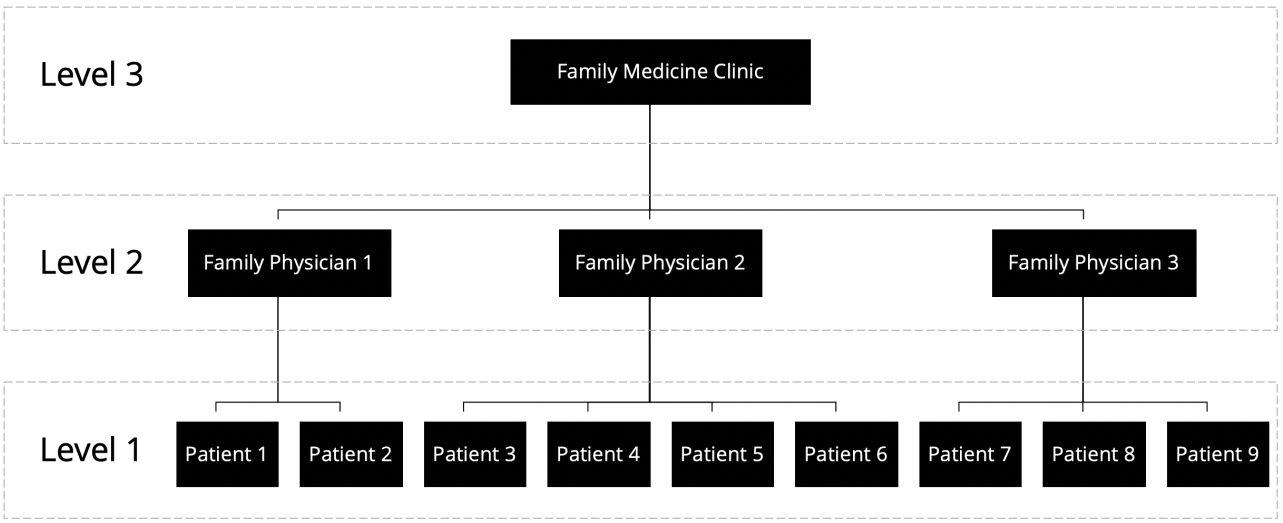

Figure 1. Graphic displaying a multilevel data structure with three levels where patients (level 1) are clustered within family physicians (level 2), which are clustered within a family medicine clinic (level 3).

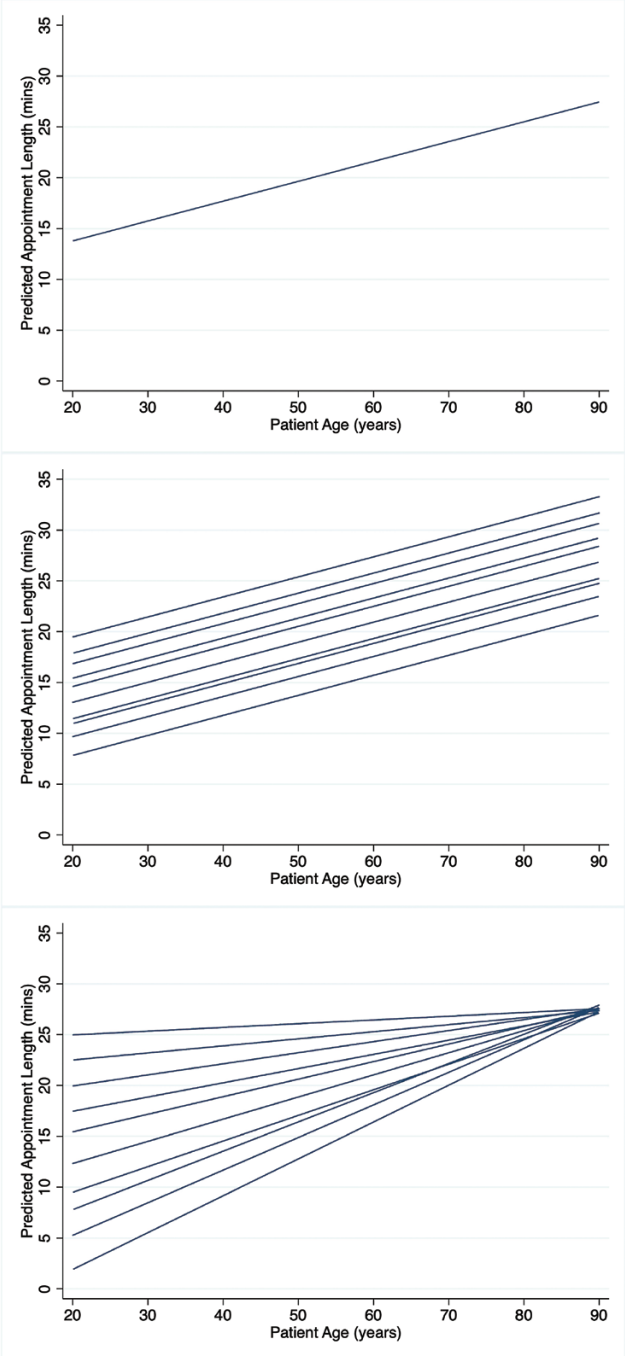

Figure 2. Results of three different analyses of first appointment length (minutes) as a function of patient age (years). (Top) Simple linear regression. As patient age increases, appointment length also increases in a linear fashion. (Middle) Multilevel linear regression allowing for random intercepts for each physician. Each line represents an individual physician. Some physicians have appointment lengths that are longer overall than others; however, the slope of the line (the relationship between patient age and appointment length) is the same across all physicians. (Bottom) Multilevel linear regression allowing for both random intercepts and slopes for each physician. Each line represents an individual physician. Some physicians have appointment lengths, which are longer overall than others, and the slope of the line (the relationship between patient age and appointment length) is also different across all physicians.

conditional or cluster-specific effect. There are fundamental differences between population-average models and multilevel models that make them better suited to different situations. The primary advantages of the multilevel model are as follows: (i) it allows the researcher to explicitly quantify the magnitude of between-cluster variation in the outcome and (ii) it can easily handle multiple levels of clustering. Population-average models, on the other hand, make fewer parametric distributional assumptions (and furthermore are robust to misspecification of the assumed correlation structure), but do not readily allow for more than one level of clustering. For a full discussion of the advantages and disadvantages of population-average models and their comparison to multilevel models, see Hubbard (13) and Gardiner (14). Neither multilevel models nor population-average models should be used if the unit of analysis is the cluster itself (e.g. an ecological study).

For further reading, we highlight here a few examples of studies using multilevel analyses in the family medicine literature. These examples show the diversity in questions and types of multilevel structure that can be handled by multilevel analyses. Dawidowicz *et al.* used multilevel analyses to identify factors associated with non-participation in cancer screening programmes among middle-aged women (15). In this study, patients were clustered within geographic regions. Along with patient-level factors, geographic factors including socioeconomic status (at the regional level), number of family physicians/100 000 people and the local presence of various specialists (gynecologist, gastroenterologist, radiologist and/or midwife) were analyzed. The ICC for the model predicting non-participation in screening programs was reported to be 0.0111, indicating that 1.11% of the variation in non-participation can be attributed to differences between the geographic clusters. It was found that several patient- and regional-level factors were associated with non-participation in screening programs. Example interpretations for some of the factors reported are as follows: self-employed workers (*patient-level factor*) have 3.76 times the odds (95% confidence interval [CI]: 3.23, 4.38) of non-participation compared with salaried workers after adjusting for other patient- and geographic-level factors, having a local radiologist in area (*geographic-level factor*) was associated with 0.94 times the odds (95% CI = 0.85, 1.02) of non-participation compared with no local radiologist in area after adjusting for other patient- and geographic-level factors (15). Guthrie used multilevel analyses in a study of personal continuity of care (seeing the same doctor) (16). In this study, a three-level model was used (patients clustered within family physicians, clustered within practices) to assess the relationship between practice-, physician- and patient-level characteristics on the outcome of patients seeing their usual physician or not. Overall, 61.6% of patients saw their usual family doctor; however, this varied significantly between practices and between physicians. This variation is illustrated using the ICC that was 0.347 for patients within practices and 0.284 for patients within physicians. The study reports the full results of the multilevel analysis in its Table 1, which includes several different practice-, physician- and patient-level factors related to the outcome of seeing one's usual physician. Finally, in a study by van Dijk *et al.*, adherence to national prescription formularies over 5 years was analyzed (17). This study had a complex multilevel data structure involving practice, patient and time. Here, it was found that the between-practice variation in formulary adherence varied by patient diagnosis, with this inter-practice variation remaining constant over the 5 years. In the reporting of this study, while inter-practice is discussed in several different contexts, the authors do not report an ICC for their model.

Summary

Multilevel data structure occurs often in primary care data, and it is important for researchers to understand the nuances of working with this data structure. In short, when analyzing data, it is important to provide consideration to the data structure. If variation in

an outcome may be impacted by variables at multiple levels *even if these effects do not pertain to the primary research question*, multilevel analysis is indicated and should be used.

Declaration

Funding: departmental resources. Dr Austin was supported by a Mid-Career Investigator award from the Heart and Stroke Foundation.
Ethical approval: none.
Conflicts of interest: The authors have no conflicts of interest to report.

References

1. Snijders T, Bosker R. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London, UK: Sage Publications.
2. Rice N, Leyland A. Multilevel models: applications to health data. *J Health Serv Res Policy* 1996; **1**: 154–64.
3. Hox JJ, Kreft IG. Multilevel analysis methods. *Sociol Methods Res* 1994;**22**(3):283–99.
4. Hox JJ. *Multilevel Analysis*. New York, NY: Routledge, 2010.
5. Leyland AH, Groenewegen PP. Multilevel modelling and public health policy. *Scand J Public Health* 2003; **31**: 267–74.
6. Robson K, Pevalin D. *Multilevel Modelling in Plain Language*. Thousand Oaks, CA: SAGE Publications Inc, 2016.
7. Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. *Can J Public Health* 2001; **92**: 150–4.
8. Merlo J, Chaix B, Yang M, Lynch J, Råstam L. A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *J Epidemiol Community Health* 2005;**59**(6):443–9.
9. Merlo J, Yang M, Chaix B, Lynch J, Råstam L. A brief conceptual tutorial on multilevel analysis in social epidemiology: investigating contextual phenomena in different groups of people. *J Epidemiol Community Health* 2005;**59**(9):729–36.
10. Merlo J, Chaix B, Yang M, Lynch J, Råstam L. A brief conceptual tutorial on multilevel analysis in social epidemiology: interpreting neighbourhood differences and the effect of neighbourhood characteristics on individual health. *J Epidemiol Community Health* 2005;**59**(12):1022–8.
11. Merlo J, Chaix B, Ohlsson H *et al*. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *J Epidemiol Community Health* 2006; **60**: 290–7.
12. Diez-Roux AV. Multilevel analysis in public health research. *Annu Rev Public Health* 2000; **21**: 171–92.
13. Hubbard AE, Ahern J, Fleischer NL *et al*. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 2010; **21**: 467–74.
14. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? *Stat Med* 2009; **28**: 221–39.
15. Dawidowicz S, Le Breton J, Moscova L, *et al*. Predictive factors for non-participation or partial participation in breast, cervical and colorectal cancer screening programmes. *Fam Pract* 2020; 37: 15–24.
16. Guthrie B. Continuity in UK general practice: a multilevel model of patient, doctor and practice factors associated with patients seeing their usual doctor. *Fam Pract* 2002; **19**: 496–9.
17. van Dijk L, de Jong JD, Westert GP, de Bakker DH. Variation in formulary adherence in general practice over time (2003–2007). *Fam Pract* 2011; **28**(6): 624–31.