Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Division of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, https://he02.tci-thaijo.org/index.php/OSIR

The Grammar of Science: Do Clusters Really Matter?

Jaranit Kaewkungwal*

Mahidol University, Thailand

*Corresponding author email: jaranitk@biophics.org

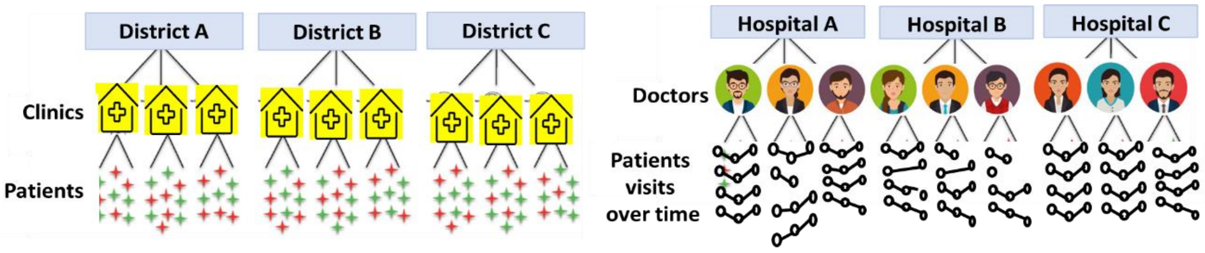
Received: 27 Aug 2025; Revised: 22 Sep 2025; Accepted: 26 Sep 2025
https://doi.org/10.59096/osir.v18i3.277904

Clustering of observations is a frequent occurrence in epidemiological and clinical research.[1–3] When planning a study, it can be beneficial to consider whether clusters exist in the population and whether the sampling approach takes them into account. For instance, in two-stage sampling, where clusters (e.g., villages) are selected first and units (e.g., households) are sampled within them, clustering within such hierarchical structure may substantially impact statistical analysis results.[3] Individuals within the same group in a population may not be independent—for example, those who share health-related environments or affect each other's behaviors and exposures in a cohort study.[1] In complex surveys, where participants are drawn from the same setting (e.g., students in a classroom or family members in a household), recruitment may be planned to examine group membership effects.[2] In cluster-randomized trials where randomization occurs at the cluster level rather than the individual level, clustering can affect study conclusions, particularly when treatment effects vary across clusters.[3]

Even when clustering is present in the designs described, it is often not considered in statistical analyses. In this paper, we explore how clustering affects the analysis of clustered data using logistic regression.

Clustered Data: Concept and Implications

Several terms are commonly used to describe clustered data, including "clustering," "nesting," "grouping," and "hierarchies," which are often used interchangeably. All of these terms refer to the concept that observations that can be organized into several distinct groups at a lower, micro level within one or more higher-level, macro units. Each macro unit represents a "level," and datasets can include multiple levels (multilevel), such as in two-level or three-level sampling designs.[4,5] As shown in Figure 1, clustered data can occur at multiple levels. Patients are nested within doctors, and doctors are, in turn, grouped within hospitals. Similarly, repeated measurements on the same individual can be viewed as nested data, where the observations are nested within the person. The impact of clustering may differ across hierarchical levels. For example, patients (level-1 units) within a single hospital or community (level-2 units) may show minimal variation in background characteristics. By contrast, differences in infrastructure, preparedness, and patient backgrounds across hospitals or communities lead to greater heterogeneity among these units.[6]



Level 1: Patients are clustered within a clinic.
Level 2: Clinics are clustered within a district.

Level 1: Patient visits are repeated measures.
Level 2: Patients are clustered within a doctor.
Level 3: Doctors are clustered within a hospital.

Figure 1. Multilevel structure of clustered data

Clustered data are common and could be problematic. A fundamental assumption underlying most standard statistical methods is that individual observations are independent, meaning that the value of one observation does not affect another.[1,7,8] Clustering often causes observations within the same cluster to be more alike than measurements from different clusters.[9] When individuals are clustered, they are not fully independent of each other. Similarities, or homogeneity, between subjects in clusters reduces the variability of their responses, compared with that expected from a random sample.

Failing to account for clustering can lead to substantial increases in Type I error rates and reduce the statistical power to detect differences between groups.[10,11] Generally, analyzing clustered data requires a larger sample size than independent data to achieve comparable person-level power. Incorporating more clusters and allowing for varying cluster sizes can improve estimate accuracy and enhances the ability to detect differences between clusters.[4] It is important to note that, when the statistical model is correctly specified and the degree of clustering is moderate, coefficient estimates typically remain unbiased. However, if the correlation among observations within clusters is ignored, the estimation of standard errors can be substantially biased—either underestimated or overestimated—leading to incorrect inferences and potentially misleading conclusions.[7] Some researchers even suggest that, in cases where statistical analysis taking into account clustering effect may not be strictly necessary, applying it can still yield approximately correct standard errors.[3]

Clustering can influence statistical inference in regression analyses, especially when the outcome variable remains clustered even after accounting for all measured predictors.[2] It also matters when both residuals and predictor variables are correlated within clusters.[3] Ignoring clustering may lead to biased estimates or inaccurate standard errors in regression model. When observations vary more between clusters than within clusters, standard regression models tend to overestimate the precision of predictor effects. Conversely, when observations are less clustered, the precision may be slightly underestimated.[2]

Logistic Regression vs. Multilevel Mixed Logistic Regression

Logistic regression is a widely used statistical method, especially in epidemiology. In particular, binary logistic regression describes the relationship between one or more predictor variables (X) and a binary outcome (Y), where Y takes one of two possible values: 0 (no event) or 1 (event occurs).

Unlike linear regression, which assumes a continuous and normally distributed outcome, logistic regression applies a logit (log-odds) transformation to the outcome. The log-odds is the natural logarithm of the odds, where odds represent the ratio of the probability of the event occurring to the probability of it not occurring, ln(Py/1-Py) or ln(Py=1/Py=0). As illustrated in Figure 2(a), in a simple logistic regression model, a one-unit increase in predictor X changes the log-odds of the outcome by an amount equal to the coefficient β_1. When this coefficient (β_1) is exponentiated, it produces the odds ratio (OR), which represents the multiplicative change in odds associated with a one-unit increase in the predictor.[12,13]

The logistic function, also called the sigmoid function, produces an S-shaped curve (Figure 2(b)) that maps log-odds to probabilities. The resulting probability, P(Y=1), is referred to as a conditional probability because it is calculated given specific values of the predictors (X). In other words, the probability of the event occurring depends on the values of the variables included in the model—P(Y=1 | X).[14]

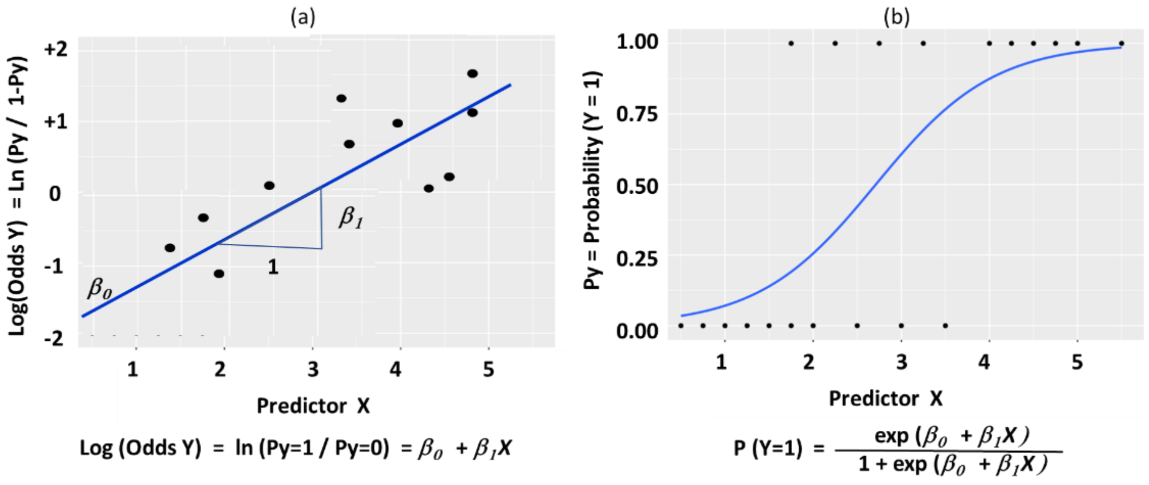


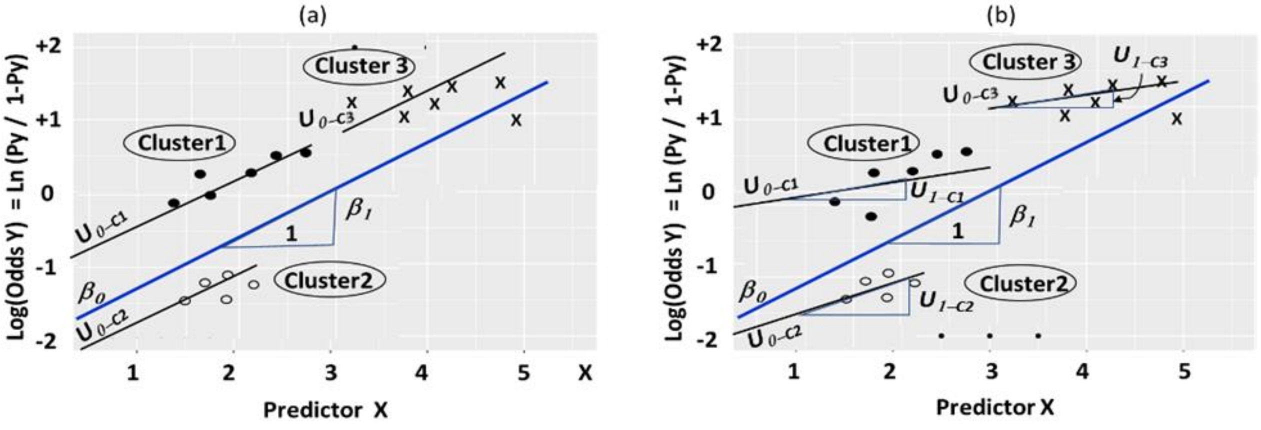
Figure 2. Logistic regression model

Multilevel mixed logistic regression extends the traditional logistic regression framework by incorporating the dependency among observations caused by clustering.[6,15] This method is frequently used in health research to properly handle clustered data when estimating the influence of both individual-level and group-level predictors on binary outcomes.[6] Ignoring clustering in a standard logistic regression, rather than addressing it with a multilevel mixed model, can result in substantial problems due to variability within clusters.[16] When a multilevel approach is not applied, an alternative is to include cluster indicators as dummy variables—though this is only practical when the number of clusters is relatively small. Nevertheless, this approach is often inefficient and less parsimonious.[17]

Notably, multilevel mixed models are extensions of the three most common regression approaches: linear, logistic, and Poisson. These models, also known as mixed-effects, hierarchical, or multilevel models, provide a statistical framework for analyzing data organized into multiple levels.[18] They are designed for situations where observations are clustered, enabling researchers to estimate both overall effects and cluster-specific differences.[19] Derived from the general linear mixed-effects model framework, they are termed "mixed" because they combine fixed effects—parameters that remain the same across clusters—with random effects, which vary between clusters.[4] Fixed effects capture consistent influences across all units, whereas random effects represent variability among them. When the units are individuals, random effects reveal individual-level differences. Common types include random intercepts, which account for differences in cluster means, and random slopes, which reflect variations in how predictors affect outcomes across clusters.[18]

The random intercept model includes a cluster-specific intercept that is estimated separately for each cluster. Its fixed component consists of the overall intercept (β_0) and slope (β_1), which apply to all observations, while the random component represents the unique intercept (U_0) for each cluster (Figure 3(a)). By incorporating random intercepts, the model accounts for unobserved group-level heterogeneity in the outcome, allowing baseline differences between groups to be properly reflected in the analysis.[20] This is especially useful for evaluating how much of the outcome's variation exists between clusters compared to within clusters. When clusters differ substantially, their intercepts deviate from the fixed component, resulting in a larger standard deviation of cluster-specific intercepts. Conversely, when observations within clusters are very similar, their outcomes tend to align closely with the fixed component.[1]

The random slope model allows the relationship between a predictor (or predictors) and the outcome to vary across clusters.[2,21] In contrast, the random intercept model allows intercepts to differ by group but assumes the slope (β_1) is the same across all groups. The random slope model relaxes this assumption by letting slopes vary randomly between groups.[1,20] As a result, both slopes and intercepts are treated as random effects, giving each cluster its own intercept (U_0) and slope (U_1). The model equation is adjusted accordingly to account for variability in both intercepts and slopes across clusters (Figure 3(b)). This flexibility is particularly useful for understanding how regression patterns differ across various group-level contexts.[20,22]



ln (Py=1 / Py=0) = $\beta_0 + U_{0\text{-}Cj} + \beta_1 X$

Random intercept model

ln (Py=1 / Py=0) = $\beta_0 + U_{0\text{-}Cj} + \beta_1 X + U_{1\text{-}Cj} X$

Random intercept and random slope model

Figure 3. Multilevel mixed logistic regression models

In theory, the random slope model is appealing because it allows the relationship between predictor(s) and the outcome to vary across clusters. When the lower-level unit is at the individual level, it is even more plausible that individual characteristics will influence the outcome differently. There is considerable debate on this issue. If there are strong a priori reasons to believe that a fixed effect should vary across individuals or clusters, random slopes should be included, provided the data can support such a model.[23,24] This is particularly important when examining cross-level interactions, where the effect of a variable at one level (e.g., individual-level) on the outcome is influenced by a variable at a higher level (e.g., cluster-level). In such cases, the literature recommends using a random slope model. Ignoring these interactions can lead to seriously biased and anti-conservative inferences.[25]

Selecting a random slope model, however, comes with several challenges. Most studies using multilevel mixed models prefer a random intercept model, as it is simpler to assume that the relationship between predictors and the outcome is consistent across all groups. The rationale for including random slopes is less straightforward and should be guided by subject-matter knowledge. It is generally recommended to first identify variables for which a group-dependent effect (random slope) is plausible.[26] If the model converges without warnings, random slopes can generally be retained in the model. Inclusion decisions are usually driven by theoretical considerations rather than statistical significance in a particular sample. Nevertheless, likelihood ratio tests can be applied, and slopes that do not improve model fit should be removed to maintain model parsimony.[23]

There are several drawbacks associated with random slope models. They can sometimes encounter singular fits, either because the correlation between slopes and intercepts is estimated near ±1, or because the variance of the random slopes is estimated near zero. In the first case, a model without the correlation can be fitted; in the latter, the random slopes are typically removed.[23] In practice, including random slopes often leads to overfitting. Moreover, mixed models assume that random effects are multivariate normal—a condition that may not hold, particularly when random slopes are included.[24]

When choosing between a random intercept and a random slope model, researchers can fit both models and compare model fit metrics. Prioritize variables expected to have the strongest effects, then estimate the model including the selected fixed and random effects. Note that data generally contain less information about random effects than fixed effects, so including many random slopes can slow estimation or even prevent convergence. Importantly, not all predictors need random slopes; only those for which a group-dependent effect is theoretically justified should be considered. Evaluate the significance of random slopes and remove those that are not significant. Similarly, assess regression coefficients and exclude non-significant predictors and consider whether to include interaction effects between predictors in level-one variables. Random slopes for interaction terms are generally discouraged, as they are often difficult to interpret.[26]

Goodness of Fit of the Model

The concept of goodness of fit refers to how effectively a statistical model captures the patterns in observed data. It assesses the agreement between predicted results and actual outcomes, providing an indication of the model fit. Selecting an appropriate model often involves a trade-off between accurately explaining the data and avoiding overfitting or unnecessary complexity. Several metrics are available to guide this decision, with the most widely used being the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria help balance model fit and simplicity, leading to models that explain the data well without adding excessive complexity.[27–30]

AIC evaluates models by considering both fit and complexity. It measures how well the model explains the data while applying a penalty for the inclusion of additional parameters to prevent overfitting. The formula is: $\text{AIC} = -2 \ln(\text{Likelihood}) + 2k$ where "Likelihood" reflects the model's fit to the data and k is the number of parameters. In essence, AIC combines the log-likelihood with a complexity penalty, ensuring a balance between model fit and parsimony. For example, in logistic regression, adding extra predictors will only improve AIC if they substantially enhance the model's fit, thereby reducing the risk of overfitting. In practice, goodness of fit is closely tied to model selection, especially in deciding how many significant predictors should be included in the model.

BIC is similar to AIC but imposes a stronger penalty for complexity, particularly in large datasets. Its formula is: $\text{BIC} = -2 \ln(\text{Likelihood}) + k \ln(n)$ where n represents the sample size. BIC is based on Bayesian probability principles and tends to favor simpler models when the evidence for added complexity is weak. Consequently, BIC is particularly useful in large-sample contexts where the risk of overfitting is high.

Both AIC and BIC estimates how much information is lost when a candidate model is used to approximate reality. Lower values indicate better models, but these metrics are meaningful only when comparing models estimated on the same dataset. In general, AIC tends to favor more complex models relative to BIC, making it a preferred criterion for smaller datasets where over-penalizing complexity could eliminate relevant predictors. Conversely, BIC is often preferred in large datasets because of its stricter penalty, which helps prevent overfitting. In practice, neither AIC nor BIC provides an absolute measure of model quality; rather, they are comparative tools that aid in selecting the most appropriate model among competing alternatives.[27–30]

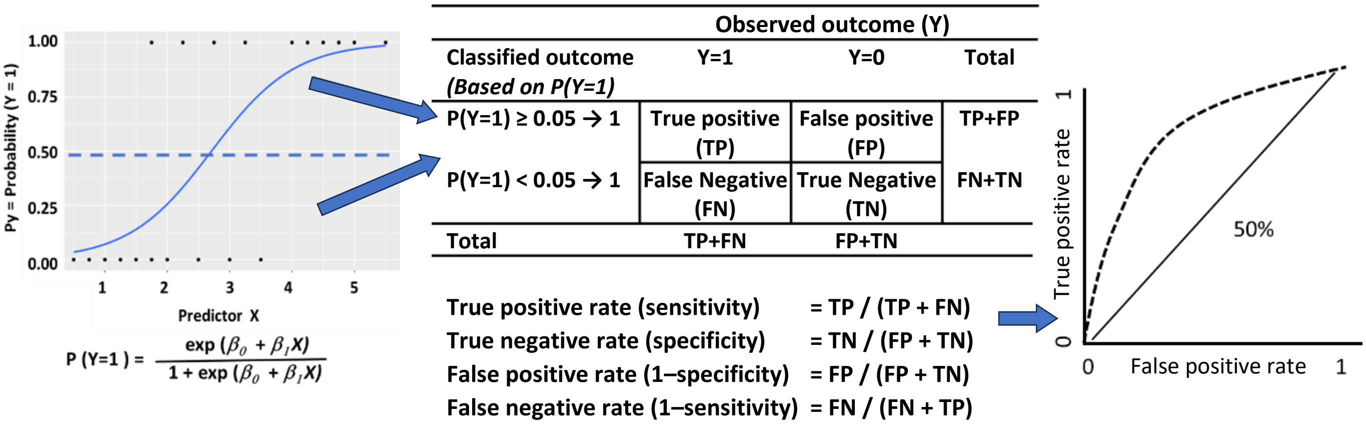
Model Accuracy in Outcome Classification

Logistic regression is one of the most widely used algorithms for classification purposes. Its predictive performance is typically evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC).[17,30,31]

Initially developed in signal detection theory, ROC curves have become a widely used method for evaluating classification performance. Classification involves predicting the category an observation belongs to based on given features. To illustrate this classification concept, let us consider a diagnostic test evaluated against a 'gold standard' that determines the true disease status.[32] If the test predicts positive and the true condition is positive, it is a True Positive (TP). If the prediction is positive but the condition is negative, it is a False Positive (FP). Similarly, a negative prediction that matches a negative condition is a True Negative (TN), and a negative prediction for a positive condition is a False Negative (FN). From these, Sensitivity (or True Positive Rate, TPR) is calculated as TP/(TP+FN), representing the proportion of correctly identified positives. Specificity (or True Negative Rate, TNR) is TN/(TN+FP). The False Positive Rate (FPR) is FP/(FP+TN), which equals (1−Specificity), and the False Negative Rate (FNR) is FN/(FN+TP), or (1−Sensitivity). While some diagnostic tests produce binary results (positive or negative), others provide continuous scores. For such cases, a cutoff threshold is applied to determine the predicted class. Adjusting this threshold impacts sensitivity and specificity—improving one often reduces the other. ROC curves illustrate this trade-off by plotting FPR on the x-axis against TPR on the y-axis across various threshold values. Lower values on the x-axis correspond to fewer false positives, while higher values on the y-axis indicate more true positives. This visualization provides a comprehensive view of a classifier's performance under different threshold settings.[30,33,34]

So, how does logistic regression perform classification? The process starts by fitting a model and computing predicted conditional probabilities P(Y) for each observation. A threshold—commonly 0.5—is then used to assign class labels: predictions above 0.5 are classified as 1 (positive), and those below as 0 (negative). ROC analysis is then applied to assess the model's ability to discriminate between actual outcomes (Y = 0/1) across different thresholds using P(Y). Here, TPR is the proportion of actual positives correctly classified as positive, while FPR is the proportion of actual negatives incorrectly classified as positive.[17,30,31]

The Area Under the Curve (AUC) summarizes the ROC curve into a single value that reflects a model's overall capability to distinguish between positive and negative outcomes. It represents the likelihood that a randomly chosen positive case and a negative case are correctly ranked by the model. AUC values range from 0 to 1, where 0.5 indicates no discrimination (equivalent to random guessing), and 1 represents perfect classification performance (Figure 4).[17,30,31]



$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

	Observed outcome (Y)		
Classified outcome *(Based on P(Y=1)*	Y=1	Y=0	Total
P(Y=1) ≥ 0.05 → 1	True positive (TP)	False positive (FP)	TP+FP
P(Y=1) < 0.05 → 1	False Negative (FN)	True Negative (TN)	FN+TN
Total	TP+FN	FP+TN	

True positive rate (sensitivity) = TP / (TP + FN)
True negative rate (specificity) = TN / (FP + TN)
False positive rate (1–specificity) = FP / (FP + TN)
False negative rate (1–sensitivity) = FN / (FN + TP)

Figure 4. ROC curve & AUC for logistic regression classification

Relevancy of Clustering in the Model

One major challenge is that the degree of correlation among observations within a cluster can significantly impact study results. Even when this correlation is small or statistically insignificant, it can still affect the validity of the analysis.[16] Ignoring such correlation may lead to inaccurate *p*-values, overly narrow confidence intervals, and biased parameter estimates, ultimately resulting in misleading interpretations.[5]

Several metrics help quantify and interpret between-cluster heterogeneity and the influence of cluster-level variables. Examples include the median odds ratio (MOR), the 80% interval odds ratio (IOR-80), and the sorting out index (SOI).[6] Among these, the most commonly used measure is the intra-cluster correlation coefficient (ICC). The ICC, denoted by the Greek letter ρ (rho), indicates the similarity or relatedness of observations within the same cluster. It reflects the proportion of outcome variance explained by differences between clusters.[1,11] (ICC can also serve other purposes, such as evaluating measurement reliability/stability by assessing the correlation between two observations from the same group).[9]

There are multiple ways to compute the ICC, but the basic approach defines it as the ratio of variance between clusters to the total variance in the data. Like other correlation measures, ICC ranges from 0 to 1 and can be interpreted in both positive and negative directions. Its magnitude represents the degree of similarity within clusters: a higher ICC implies stronger clustering effects.[16] When all clusters have unique values, the ICC approaches 1; when clusters are identical, it approaches 0. In practical terms, an ICC near 0 suggests minimal contribution of clustering to the model, whereas an ICC close to 1 indicates strong clustering and significant relevance of clusters (Figure 5).[3,19]

$$ICC\ (\rho) = \frac{\sigma^2_{\text{between cluster}}}{\sigma^2_{\text{between cluster}} + \sigma^2_{\text{within cluster (residual)}}}$$

cluster	Y
1	10
1	10
1	10
2	15
2	15
2	15
3	21
3	21
3	21
4	34
4	34
4	34
5	45
5	45
5	45

Unique value in each cluster

ICC=1

cluster	Y
1	10
1	34
1	45
2	10
2	34
2	45
3	10
3	34
3	45
4	10
4	34
4	45
5	10
5	34
5	45

Similar values in all clusters

ICC=0

ICC: intra-cluster correlation coefficient

Figure 5. Intra-cluster correlation coefficients

Case Study

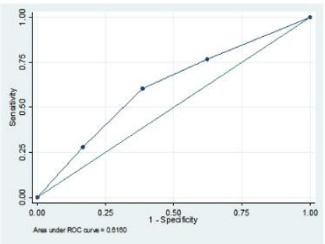
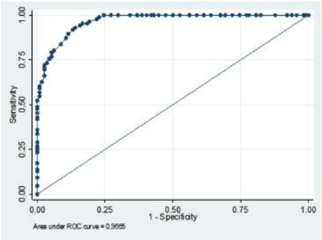
To illustrate the impact of clustering, consider two simulated datasets, each containing 200 observations divided into 20 clusters (10 observations per cluster). Both datasets include a binary outcome variable (Y=0/1) and two predictors (X1, X2). The primary distinction between them is the degree of clustering in the outcome: one dataset demonstrates a strong clustering effect, while the other shows a weak effect (Figure 6).

	Cluster																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	*High clustering effect*																			
0	10	4	10	9	10	1	3		7	10		3	10	10	5	6	5	1		10
1		6		1		9	7	10	3		10	7			5	4	5	9	10	
Y	*Low clustering effect*																			
0	4	5	5	5	5	7	4	4	6	4	8	3	5	3	2	5	7	2	3	9
1	6	5	5	5	5	3	6	6	4	6	2	7	5	7	8	5	3	8	7	1

Figure 6. Clustering effect in two hypothetical datasets

To evaluate model performance, two approaches were applied: standard logistic regression (which ignores clustering) and multilevel mixed-effects logistic regression with random intercepts (which accounts for clustering).

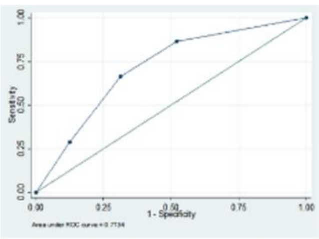
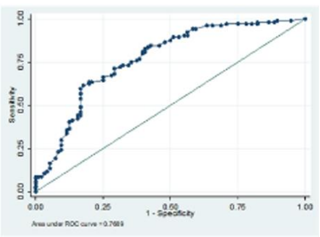
In the high-clustering dataset (Figure 7), the mixed-effects model substantially outperformed standard logistic regression in terms of fit. Classification accuracy showed a marked difference: the AUC for logistic regression was approximately 61%, compared to 97% for the mixed model. Here, the ICC was 0.89, underscoring the critical importance of accounting for clustering.

				Logistic regression		Mixed model–Logistic regression (with cluster effect)	
		Y=0	Y=1	Odds ratio (95% CI)	*p*-value	Odds ratio (95% CI)	*p*-value
X1	1	44	52	2.41 (1.36–4.29)	0.003	9.05 (2.74–29.81)	<0.001
	0	70	34	1		1	
X2	1	46	38	1.12 (0.63–2.00)	0.699	2.81 (0.93–8.47)	0.066
	0	68	48	1		1	
Model goodness of fit							
AIC				269.71		158.41	
BIC				279.61		171.61	
Relevancy of clustering							
ICC (95% CI)				-		0.89 (0.62–0.98)	
Accuracy of model classification							
AUC (95% CI)				0.61 (0.54–0.69)		0.97 (0.95–0.99)	

CI: confidence interval. AIC: Akaike information criterion. BIC: Bayesian information criterion. AUC: area under the curve. ICC: intra-cluster correlation coefficient.

Figure 7. Performance of models accounting for clustering versus ignoring clustering in a high-ICC dataset

In contrast, in the low-clustering scenario (Figure 8), both models produced similar fit statistics (AIC, BIC). Classification performance was also comparable: the AUC for logistic regression was about 71%, while the mixed model achieved 76%. The intraclass correlation coefficient (ICC) for the mixed model was 0.07, indicating that adjusting for clustering offered little advantage.

				Logistic regression		Mixed model–Logistic regression (with cluster effect)	
		Y=0	Y=1	Odds ratio (95% CI)	*p*-value	Odds ratio (95% CI)	*p*-value
X1	1	30	69	4.47 (2.44–8.18)	<0.001	4.90 (2.55–9.42)	<0.001
	0	66	35	1		1	
X2	1	32	51	2.05 (1.11–3.79)	0.022	2.12 (1.11–4.04)	0.022
	0	64	53	1		1	
Model goodness of fit							
AIC				252.44		252.70	
BIC				262.34		265.90	
Relevancy of clustering							
ICC (95% CI)				-		0.07 (0.01–0.37)	
Accuracy of model classification							
AUC (95% CI)				0.71 (0.64–0.78)		0.76 (0.70–0.83)	

CI: confidence interval. AIC: Akaike information criterion. BIC: Bayesian information criterion. AUC: area under the curve. ICC: intra-cluster correlation coefficient.

Figure 8. Performance of models accounting for clustering versus ignoring clustering in a low-ICC dataset

These examples show that the importance of clustering largely depends on the level of ICC. When ICC is low, using either a standard logistic regression or a mixed-effects model makes little difference. However, when ICC is high, ignoring clustering can result in poorer model fit, biased estimates, and reduced predictive accuracy.

Key Takeaways: Do Clusters Really Matter?

Clustering matters most when ICC is high—ignoring it can affect your results. When ICC is low, simpler models work fine, but with high ICC, mixed-effects models are recommended for more accurate and reliable predictions.

Acknowledgements

An AI tool, ChatGPT (OpenAI, 2025), was used to generate language suggestions during the preparation of this manuscript, and all outputs were reviewed for accuracy and appropriateness.[35] The author reviewed, edited, and take responsibility for the final content.

Suggested Citation

Kaewkungwal J. The grammar of science: do clusters really matter? OSIR. 2025 Sep;18(3):183–91. doi:10.59096/osir.v18i3.277904.

References

1. Columbia University Mailman School of Public Health. Multi-Level Modeling [Internet]. New York: Columbia University Mailman School of Public Health; [cited 2025 Aug 15]. <https://www.publichealth.columbia.edu/research/population-health-methods/multi-level-modeling>
2. Ntani G, Inskip H, Osmond C, Coggon D. Consequences of ignoring clustering in linear regression. BMC Med Res Methodol. 2021 Jul 7;21(1):139. doi: 10.1186/s12874-021-01333-7.
3. Bellemare MF. Metrics Monday: when (not) to cluster? [Internet]. Saint Paul: Marc F. Bellemare; [updated 2017 Nov 13; cited 2025 Aug 15]. <https://marcfbellemare.com/wordpress/12712>
4. Hoffman L. Introduction to multilevel models (MLMs) for clustered data [Internet]. Iowa City: Lesa Hoffman; [cited 2025 Aug 15]. 21 p. <https://www.lesahoffman.com/PSQF7375_Clustered/PSQF7375_Clustered_Lecture1_Intro_MLM.pdf>
5. Zyzanski SJ, Flocke SA, Dickinson LM. On the nature and analysis of clustered data. Ann Fam Med. 2004 May-Jun;2(3):199–200. doi:10.1370/afm.197.
6. Adam NS, Twabi HS, Manda SOM. A simulation study for evaluating the performance of clustering measures in multilevel logistic regression. BMC Med Res Methodol. 2021 Nov 13;21(1):245. doi:10.1186/s12874-021-01417-4.
7. McNeish DM. Analyzing clustered data with OLS regression: the effect of a hierarchical data structure. Multiple Linear Regression Viewpoints. 2014;40(1):1–16.
8. Austina PC, Merlod J. Intermediate and advanced topics in multilevel logistic regression analysis. Stat Med. 2017 Sep 10;36(20):3257-3277. doi:10.1002/sim.7336.
9. King J. Clustered data: data analysis for psychology in R 3 [Internet]. Edinburgh: Department of Psychology, University of Edinburgh; [cited 2025 Aug 15]. 34 p. <https://uoepsy.github.io/dapr3/2324/lectures/dapr3_2324_01b_clusters.html#1>
10. Miles J. Methods for dealing with clustered data [Internet]. Southampton: National Centre for Research Methods Social Sciences. [cited 2025 Aug 15]. 57 p. <https://eprints.ncrm.ac.uk/id/eprint/4725/1/Methods%20for%20Dealing%20with%20Clustered%20Data.pdf>
11. Barratt H, Kirwan M, Shantikumar S. Clustered data - effects on sample size and approaches to analysis [Internet]. London: Faculty of Public Health; c2018 [cited 2025 Aug 15]. <https://www.healthknowledge.org.uk/public-healthtextbook/research-methods/1a-epidemiology/clustered-data>
12. Hosmer DW. Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: John Wiley & Sons, Inc; 2013. 510 p. doi:10.1002/9781118548387.
13. Agresti A. An Introduction to categorical data analysis. 3rd ed. Hoboken: John Wiley & Sons, Inc; 2006. 372 p. doi:10.1002/0470114754.
14. Taboga M. Logistic classification model (logit or logistic regression) [Internet]. North Charleston: Kindle Direct Publishing; 2021 [cited 2025 Aug 15]. <https://www.statlect.com/fundamentals-of-statistics/logistic-classification-model.>
15. Sommet N, Morselli D. Keep calm and learn multilevel logistic modeling: a simplified three-step procedure using Stata, R, Mplus, and SPSS. International Review of Social Psychology. 2017;30(1):203–18. doi:10.5334/irsp.90.

16. Galbraith S, Daniel JA, Vissel B. A study of clustered data and approaches to its analysis. J Neurosci. 2010 Aug 11;30(32):10601–8. doi:10.1523/JNEUROSCI.0362-10.2010.

17. Lee S. ROC & AUC in logistic regression: a primer [Internet]. New York: Number Analytics LLC; [updated 2025 May 16; cited: 2025 Aug 15]. <https://www.numberanalytics.com/blog/roc-auc-logistic-regression-primer>

18. Oberauer K. The Importance of Random Slopes in Mixed Models for Bayesian Hypothesis Testing. Psychol Sci. 2022 Apr;33(4):648–65. doi:10.1177/09567976211046884.

19. Al Amin M, Qin Y. Multilevel analysis in Stata: a step-by-step guide [Internet]. Princeton: Princeton University Library; [updated 2024 Aug 14; cited 2025 Aug 15]. <https://libguides.princeton.edu/multilevel>

20. Sparks CS. DEM 7473 - week 3: basic hierarchical models - random intercepts and slopes [Internet]. Boston: RStudio; 2018 Sep 17 [cited: 15 Aug 2025]. <https://rpubs.com/corey_sparks/420770>

21. Centre for Multilevel Modelling, University of Bristol. Random slope models [Internet]. Bristol: University of Bristol; [cited 2025 Aug 15]. <https://www.bristol.ac.uk/cmm/learning/videos/random-slopes.html>

22. College of Public Health & Health Professional. University of Florida. Random slope models [Internet]. Gainesville: University of Florida Health; [cited 2025 Aug 15]. <https://users.phhp.ufl.edu/rlp176/Courses/SurveyBiostat/LMM/RSmodels.html>

23. Long R. What are the arguments in favor and against using random slopes? [Internet]. New York: Stack Exchange Inc; 2021 May 17 [cited 2025 Aug 15]. <https://stats.stackexchange.com/questions/524599/what-are-the-arguments-in-favor-and-against-using-random-slopes>

24. Long R. Is it a must to include a random slope in a mixed model? [Internet]. New York: Stack Exchange Inc; 2020 Aug 28 [cited 2025 Aug 15]. <https://stats.stackexchange.com/questions/485048/is-it-a-must-to-include-a-random-slope-in-a-mixed-model>

25. Heisig JP, Schaeffer M. Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. European Sociological Review. 2019;35(2):258–79. doi:10.1093/esr/jcy053.

26. Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modeling. 2nd ed. London: SAGE Publications; 2004. 368 p.

27. Jani Data Diaries. Choosing the best model: f friendly guide to AIC and BIC [Internet]. San Francisco: A Medium Corporation; 2024 Nov 7 [cited 2025 Aug 15]. <https://medium.com/@jshaik2452/choosing-the-best-model-a-friendly-guide-to-aic-and-bic-af220b33255f>

28. Banerjee S. Model magic with AIC & BIC: navigating fit and elegance [Internet]. San Francisco: A Medium Corporation; 2023 Oct 16 [cited 2025 Aug 15]. <https://shekhar-banerjee96.medium.com/model-magic-aic-bic-mdl-navigating-fit-and-elegance-726c784edf9b>

29. Kumar A. AIC in logistic regression: formula, example [Internet]. New York: Analytics Yogi; 2023 Nov 30 [cited 2025 Aug 15]. <https://vitalflux.com/aic-in-logistic-regression-formula-example/>

30. Faculty of Medicine and Health Sciences. Goodness of fit in logistic regression [Internet]. Montreal: McGill University; [cited 2025 Aug 15]. 17 p. <https://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logfit.pdf>

31. Arya N. Classification metrics walkthrough: logistic regression with accuracy, precision, recall, and ROC [Internet]. San Juan: KDnuggets; 2022 Oct 13 [cited 2025 Aug 15]. <https://www.kdnuggets.com/2022/10/classification-metrics-walkthrough-logistic-regression-accuracy-precision-recall-roc.html>

32. LaMorte WW. Screening for disease: test validity [Internet]. Boston: School of Public Health, Boston University; [cited 2025 Aug 15]. <https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_screening/EP713_Screening3.html>

33. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin Biochem Rev. 2008;29 Suppl 1(Suppl 1):S83–7.

34. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. CJEM. 2006;8(1):19–20. doi:10.1017/s1481803500013336.

35. OpenAI. ChatGPT [Internet]. 2025 [cited 2025 May 6] <https://chat.openai.com>