A Goodness-of-Fit Test for Mixed-Effects Logistic Regression

Ariel Linden

University of California, San Francisco

Department of Medicine

Division of Clinical Informatics & Digital Transformation (DoC-IT)

ariel.linden@ucsf.edu

Abstract

Mixed-effects logistic regression is widely used for analyzing binary outcomes in hierarchical data, yet formal goodness-of-fit tests remain limited to random intercept models and do not address sparse cluster settings. We extend a grouping-based Wald test to mixed-effects logistic models with random slopes. The procedure groups observations by their predicted probabilities within clusters, adds pooled group indicators to the model, and tests their joint significance using a Wald statistic. To accommodate small clusters, we introduce a data-driven rule for selecting the number of groups, $G = \min(10, n_{\min})$, where $n_{\min}$ is the smallest cluster size. This prevents estimation failure when clusters contain fewer than ten observations. Simulation studies across 24 null scenarios show that the test maintains nominal Type I error in three-level random slope models, including at smaller sample sizes than previously studied. The test demonstrates increasing power to detect fixed-effects misspecification: power against omitted nonlinearity increased from 0.07 to 1.00 across effect sizes, and power against omitted interactions reached 0.87. As expected, the test has no power to detect an omitted clustering level, reflecting its

focus on residual patterns in predicted probabilities. In sparse balanced designs, fixing $G = 10$ led to complete test failure, whereas the data-driven rule performed reliably. The method is implemented in the community-contributed Stata program `mlm_gof` after fitting a mixed-effects logistic regression model.

Keywords: goodness-of-fit; mixed-effects logistic regression; multilevel models; random slopes; simulation study

1 Introduction

Multilevel data structures are common in health services, epidemiologic, and social science research. Patients are nested within doctors, hospitals within regions, and repeated measurements within individuals. When the outcome is binary (e.g., mortality, disease onset, or treatment success) the appropriate analytic tool is the mixed-effects logistic regression model, which accounts for clustering and partitions variation across levels. The terms *mixed-effects*, *multilevel*, and *hierarchical* are used interchangeably in the literature; here we use *mixed-effects*. Similarly, we use *random slopes* to refer to models in which regression coefficients vary across clusters. The use of these models has increased substantially in recent decades across medicine, public health, and the social sciences.[1]

A fitted model is only useful to the extent that it adequately describes the data. Before drawing inferences about covariate effects or variance components, it is essential to assess model fit. For single-level logistic regression, several goodness-of-fit tests are well established. The Hosmer–Lemeshow test[2,3] groups observations by predicted probabilities and compares observed and expected frequencies. Lipsitz et al.[4] extended this approach by adding group indicator variables to the model and testing their joint significance via a Wald statistic, making the method adaptable to correlated data. Further extensions have addressed survey-weighted models,[5] complex sampling designs,[6] and multinomial outcomes.[7]

In contrast, goodness-of-fit testing for mixed-effects logistic regression has been limited. Cool et al.[8] highlighted the absence of suitable methods for higher-level multilevel binary models. Sturdivant and Hosmer[9] proposed residual-based statistics but did not evaluate power. Perera et al.[10] developed the first formal test for two-level models by adapting grouping-based methods to clustered data. Fernando and Sooriyarachchi[11] extended this framework to three-level models using limited-information approaches. Although these contributions represent important advances, they are restricted to random intercept models and do not address the more general case in which both intercepts and slopes vary randomly across clusters, which is the appropriate specification whenever the effect of a covariate may differ across clusters.

A second limitation of existing tests is the fixed use of ten groups in the grouping procedure, following Hosmer and Lemeshow.[2] In mixed-effects settings, clusters may contain fewer than ten observations, making this choice infeasible: empty cells arise, indicator variables become degenerate, and the augmented model cannot be estimated. Existing methods do not address this issue or provide guidance for sparse cluster settings.

This paper makes two contributions. First, we extend the grouping-based Wald test framework[10,11] to mixed-effects logistic regression models with random slopes. Second, we introduce a data-driven rule for selecting the number of groups, $G = \min(10, n_{\min})$, where $n_{\min}$ is the minimum cluster size. Simulation results show that this rule is necessary to avoid test failure when clusters are small. We evaluate the proposed method using an extensive simulation study covering Type I error, power under three types of misspecification, and a direct comparison of the data-driven grouping rule with the conventional choice of $G = 10$. The method is implemented in the community-contributed Stata program `mlm_gof`.[12]

The remainder of the paper is organized as follows. Section 2 describes the model, test procedure, and simulation design. Section 3 reports simulation results. Section 4 presents an applied example. Section 5 discusses results and Section 6 concludes.

2 Methods

2.1 The Mixed-Effects Logistic Regression Model with Random Slopes

Let y_{ij} denote a binary outcome for observation i in cluster j ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$). A two-level mixed-effects logistic regression model with a random intercept and a random slope on covariate x_{2ij} is

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{1j}\, x_{2ij}, \tag{1}$$

where $p_{ij} = \Pr(y_{ij} = 1 \mid \mathbf{u}_j)$, $\beta_0, \beta_1, \beta_2$ are fixed effects, and $(u_{0j}, u_{1j})^\top \sim N(\mathbf{0}, \boldsymbol{\Omega})$ are cluster-specific random effects. The random intercept u_{0j} captures between-cluster variation in baseline log-odds, while the random slope u_{1j} allows the effect of x_2 to vary across clusters. Conditional predicted probabilities are obtained as

$$\hat{p}_{ij} = \text{logit}^{-1}\big(\hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + \hat{\beta}_2 x_{2ij} + \tilde{u}_{0j} + \tilde{u}_{1j}\, x_{2ij}\big), \tag{2}$$

where $\tilde{u}_{0j}$ and $\tilde{u}_{1j}$ are empirical Bayes estimates. A three-level extension adds random effects at an intermediate level (e.g., subjects within families). This formulation generalizes prior work, which considered random intercept models only.[10,11]

2.2 Derivation of the Goodness-of-Fit Test Statistic

The test follows the model-based approach of Lipsitz et al.,[4] extended to clustered data by Perera et al.[10] for the two-level case and by Fernando and Sooriyarachchi[11] for the three-level case. The null hypothesis is that the fitted model adequately describes the data. Under the null, the predicted probabilities $\hat{p}_{ij}$ should capture the observed outcomes without systematic residual structure.

Let G denote the number of groups formed from the distribution of $\hat{p}_{ij}$ within each cluster. Define $I_{ij}^g = 1$ if observation i in cluster j belongs to group g ($g = 2, \ldots, G$). The augmented model is

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \gamma_2 I_{ij}^2 + \cdots + \gamma_G I_{ij}^G + u_{0j} + u_{1j}\, x_{2ij}. \tag{3}$$

The null hypothesis is $H_0 \colon \gamma_2 = \cdots = \gamma_G = 0$. The joint Wald statistic is

$$W = \hat{\boldsymbol{\gamma}}^\top \big[\widehat{\text{Var}}(\hat{\boldsymbol{\gamma}})\big]^{-1} \hat{\boldsymbol{\gamma}}, \tag{4}$$

which is approximately χ^2 with $G - 1$ degrees of freedom.[4] Large values of W indicate lack

of fit.

2.3 The Step-by-Step Procedure Implemented in `mlm_gof`

Step 1. Fit the baseline model. Estimate the mixed-effects logistic model using maximum likelihood and obtain predicted probabilities $\hat{p}_{ij}$.

Step 2. Select the number of groups. Set $G = \min(10, n_{\min})$, where $n_{\min}$ is the minimum number of observations in any level-2 cluster.

Step 3. Group observations. Within each level-2 cluster, sort $\hat{p}_{ij}$ and assign observations to G approximately equal-sized groups using the asymptotic ranking approach of Rosner et al.,[13] which preserves the overall ranking across clusters while enabling within-cluster grouping.

Step 4. Create indicators. For each observation, define $G - 1$ binary indicator variables $I_{ij}^{g} = 1$ if observation i in cluster j belongs to group g ($g = 2, \ldots, G$), with group 1 as the reference. Indicators for the same group g across clusters are pooled into a single variable for the full dataset.[10]

Step 5. Refit the model. Estimate the augmented model (3) including the indicators and the original random-effects structure.

Step 6. Compute the test statistic. Compute W from equation (4) and compare to a χ^2_{G-1} distribution to obtain the p-value.

Step 7. Report results. Report W, degrees of freedom, p-value, and the value of G used.

This implementation differs from prior work[10,11] in that estimation uses full maximum likelihood and the original random-effects structure is retained in the augmented model.

2.4 The Data-Driven Group Selection Rule

The conventional choice $G = 10$ is not always feasible in mixed-effects settings where clusters may be small. When $n_{\min} < 10$, dividing clusters into ten groups produces empty cells and prevents estimation. We therefore use $G = \min(10, n_{\min})$, which ensures that each group contains at least one observation per cluster. When clusters are sufficiently large, this reduces to the standard choice. Simulation results (Section 3.3) show that this rule is necessary for valid inference in sparse cluster settings.

2.5 Simulation Study Design

We conducted three simulation studies to evaluate the proposed test. The data-generating process was common to all components; design parameters are summarized in Tables 1–3. The data-generating process and null model are three-level throughout. The three-level specification is the more demanding case, as it involves a more complex clustering structure and additional variance components relative to a two-level model. Results under three-level specifications therefore provide evidence that the test will also perform adequately in the simpler two-level setting. 1000 replications were generated for each of the following 44 scenarios (total of 44,000 replications).

Data-generating process. The true model is

$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2ijk} + v_j + u_{kj} + w_{kj}\, x_{2ijk}, \tag{5}$$

where $x_{1ijk} \sim U(-3, 3)$, $x_{2ijk} \sim$ Bernoulli(0.5), $v_j \sim N(0, \sigma_v^2)$ is the family-level random intercept, $u_{kj} \sim N(0, \sigma_u^2)$ is the subject-level random intercept, and $w_{kj} \sim N(0, \sigma_w^2)$ is the subject-level random slope on x_2, all mutually independent. Fixed parameters: $\beta_0 = -1.0$, $\beta_1 = 0.5$, $\beta_2 = 0.3$. ICC $= \sigma^2/(\sigma^2 + \pi^2/3)$.

Part 1: Type I error. Data are generated from the correctly specified model and fitted by Stata's `melogit` command. Factors varied: $J \in \{15, 30, 50\}$, $K \in \{5, 10\}$, $n = 20$, ICC

$\in \{0.10, 0.30\}$, accounting for 24 scenarios. See Table 1.

Part 2: Power. Fixed conditions: $J = 30$, $K = 5$, $n = 20$, ICC = 0.20. Three misspecification types: (a) omitted quadratic term ($\beta_3 x_1^2$, $\beta_3 \in \{0.02, 0.05, 0.10, 0.15\}$); (b) omitted interaction ($\beta_3(x_1 \times x_2)$, $\beta_3 \in \{0.3, 0.6, 0.9\}$); (c) omitted level (true three-level model fitted as two-level, $\sigma_{\text{extra}} \in \{0.5, 1.0, 1.5\}$), for a total of 10 scenarios. See Table 2.

Part 3: Group selection sensitivity. $G = \min(10, n_{\min})$ vs. $G = 10$ at $n_{\text{small}} \in \{3, 5, 6, 8, 10\}$. Unbalanced design: 10 small clusters (n_{small} obs) and 40 large clusters ($n = 20$). Balanced design: all 50 clusters have n_{small} observations. Both designs together account for a total of 10 scenarios. See Table 3.

This design differs from prior work[10,11] in three respects: estimation uses full maximal likelihood rather than penalized or marginal quasi-likelihood methods (PQL/MQL); the addition of random slopes in the model; and smaller sample sizes (N = 1,500–7,500) than the minimum N = 4,500 used by Fernando and Sooriyarachchi[11], allowing for validation at more modest sample sizes than previously studied. All analyses were conducted using Stata version 19 (StataCorp LLC, College Station, TX).

3 Results

3.1 Type I Error (Part 1)

Figures 1 and 2 present empirical rejection rates of `mlm_gof` at $\alpha = 0.05$ across all 24 null scenarios. Rejection rates ranged from 0.035 to 0.060, with all values within the Monte Carlo bounds of $[0.036, 0.064]$.[14] No systematic pattern of inflation or conservatism was observed. Performance remained stable at the smallest design ($J = 15$, $K = 5$, $n = 20$; N = 1,500), indicating that the test maintains nominal Type I error even at sample sizes smaller than those considered in prior studies.[10,11]

3.2 Power (Part 2)

Table 4 summarizes empirical power across the three misspecification scenarios.

Omitted quadratic term. Power was low at the smallest effect ($\beta_3 = 0.02$, power $= 0.074$), consistent with negligible misspecification, and increased rapidly, exceeding 0.90 at $\beta_3 = 0.10$, indicating strong sensitivity to nonlinear functional form misspecification.

Omitted interaction. Power increased more gradually, reaching 0.54 at $\beta_3 = 0.6$ and 0.87 at $\beta_3 = 0.9$, consistent with moderate-to-strong sensitivity at larger effect sizes.

Omitted clustering level. Power remained low across all values of σ_{extra}, indicating that the test does not detect misspecification arising from omission of a clustering level.

3.3 Group Selection Sensitivity (Part 3)

Results are shown in Table 5.

Unbalanced design. Both the data-driven rule and $G = 10$ maintained acceptable Type I error across all values of n_{small}. Differences were minimal, reflecting the dominance of larger clusters in the test statistic.

Balanced design. The data-driven rule produced valid results across all scenarios within Monte Carlo bounds. In contrast, fixing $G = 10$ failed in all replications when $n_{\text{small}} < 10$, yielding no valid test results. When $n_{\text{small}} = 10$, both approaches produced identical results.

4 Applied Example

4.1 Study Context and Data

We illustrate the implementation of `mlm_gof` using an artificial dataset reflecting a realistic disease management study. Disease management programs target individuals with chronic

conditions or elevated risk, through structured behavioral and clinical interventions.[15] Pre-diabetes is a common target given the effectiveness of lifestyle interventions in preventing progression to type 2 diabetes.[16] The dataset represents a family-based intervention study in which patients with pre-diabetes are enrolled along with family members, yielding a three-level structure: repeated clinic visits (level 1) nested within subjects (level 2) nested within families (level 3). The binary outcome is whether pre-diabetes was reversed at each follow-up visit. The dataset includes approximately 900 observations across 30 families, with 5–7 subjects per family and 5 visits per subject. Covariates include intervention assignment (1 = intervention, 0 = control), body mass index centered at 30 kg/m^2 (bmi_c), and visit number (1–5). The data-generating model includes a family-level random intercept, a family-level random slope on visit, and a subject-level random intercept.

4.2 Model Fitting and Goodness-of-Fit Assessment

We fit the model

$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1 \,\text{intervention} + \beta_2 \,\text{bmi\_c} + \beta_3 \,\text{visit} + v_j + w_j \,\text{visit} + u_{kj}, \tag{6}$$

where v_j is the family-level random intercept, w_j is the family-level random slope for visit, and u_{kj} is the subject-level random intercept. Table 6 reports the results. Because each subject contributes five observations, $n_{\min} = 5$ and the data-driven rule sets $G = 5$.

The goodness-of-fit test yields $W = 2.92$ with $df = 4$ ($p = 0.404$), providing no evidence against adequate fit, as expected since the fitted model corresponds to the data-generating process. Using $G = 10$ would lead to estimation failure in this setting (Section 3.3).

5 Discussion

We developed and validated a goodness-of-fit test for mixed-effects logistic regression models. The test extends the grouping-based Wald framework of Perera et al.[10] and Fernando and

Sooriyarachchi[11] to accommodate both random slopes and sparse cluster settings. It is implemented in the Stata program `mlm_gof`, allowing direct application after fitting a mixed-effects logistic regression model.

The simulation results yield three principal findings. First, the test maintains nominal Type I error across a range of three-level random slope specifications. This holds at the smallest design ($N = 1{,}500$), considerably smaller than the minimum $N = 4{,}500$ at which Fernando and Sooriyarachchi[11] reported instability, and smaller than the settings where Perera et al.[10] observed distortion (rejection rate 0.013). Second, the test has good power against fixed-effects misspecification, including omitted nonlinear terms and omitted interactions. Both prior studies evaluated power using a single $\log X^2$ alternative that yields near-perfect power in large samples but does not characterize the sensitivity profile across effect sizes; the graduated grids used here are more informative. Third, the test has no power to detect omission of a clustering level, regardless of the magnitude of the omitted variance component. This finding reflects a structural limitation of the method rather than a deficiency of the simulation design. The grouping procedure detects systematic patterns in the conditional predicted probabilities within clusters; omission of a random intercept does not induce such patterns. As a result, goodness-of-fit tests based on grouping cannot detect this form of misspecification. Researchers concerned about the adequacy of the random-effects structure should instead rely on likelihood-based model comparison or information criteria.

The data-driven group selection rule $G = \min(10, n_{\min})$ is the second major contribution of this work. The simulation results demonstrate that this rule is necessary, not merely convenient. In balanced cluster designs with n_min<10, fixing $G = 10$ led to complete failure of the test in every replication, producing no valid results. In contrast, the data-driven rule yielded valid inference across all scenarios while recovering the conventional choice when clusters were sufficiently large. Neither Perera et al.[10] nor Fernando and Sooriyarachchi[11] address this issue, instead they adopt the standard choice of $G = 10$ from Hosmer and Lemeshow.[2] The present findings show that this convention does not generalize to mixed-

effects settings with small clusters.

Several limitations should be noted. The Wald statistic relies on asymptotic approximations in the number of clusters; when the number of clusters is small (e.g., $J < 15$), the χ^2 approximation may be unreliable, and bootstrap or permutation-based approaches may be preferable. The test is also sensitive to the choice of the number of groups G, as both the degrees of freedom and the grouping structure influence power; sensitivity analysis across alternative values of G is straightforward using the `groups()` option in `mlm_gof`. Finally, the method relies on full maximum likelihood estimation via adaptive Gauss–Hermite quadrature. Prior implementations of grouping-based goodness-of-fit tests for mixed-effects logistic regression[10,11] used penalized or marginal quasi-likelihood (PQL/MQL) methods, which are known to produce biased estimates in binary mixed-effects models, particularly in small samples or with high intraclass correlation.[17,18] The present approach avoids this limitation, though its performance under PQL/MQL estimation has not been evaluated.

From an applied perspective, `mlm_gof` addresses a gap in the diagnostic toolkit for mixed-effects logistic regression. A non-significant result indicates that grouping of predicted probabilities does not reveal systematic lack of fit, providing a useful complement to other diagnostics such as residual analysis and likelihood-based model comparison. In combination, these tools support a more comprehensive assessment of model adequacy. More broadly, the evaluation of goodness-of-fit is an essential component of rigorous statistical analysis. As Linden and Roberts[19] emphasize, the validity of conclusions drawn from fitted models depends on demonstrating that those models adequately represent the data. In mixed-effects logistic regression, where model structures are more complex and assumptions less transparent than in single-level analyses, the availability of a formal goodness-of-fit test is particularly valuable. The proposed method facilitates routine assessment and reporting of model fit in applied research using mixed-effects logistic models.

6 Conclusions

We have presented a goodness-of-fit test for mixed-effects logistic regression models that include random slopes, implemented in the Stata program `mlm_gof`. The test extends the grouping-based Wald framework designed only for models with random intercepts and additionally introduces a data-driven group selection rule that prevents failure in sparse cluster contexts. It is validated across a broad range of multilevel designs. Simulation results demonstrate excellent Type I error control, good power to detect fixed-effects misspecification, and the necessity of the data-driven rule when cluster sizes fall below ten. The test is sensitive to functional form misspecification but not to omitted random effects, a distinction that should guide its use alongside likelihood-based model comparison tools. Together, these properties make `mlm_gof` a practical and principled tool for assessing model adequacy in mixed-effects logistic regression.

References

[1] Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. *Stat Med.* 2017;36(20):3257–3277.

[2] Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Commun Stat Theory Methods.* 1980;9(10):1043–1069.

[3] Hosmer DW, Lemeshow S. *Applied Logistic Regression.* 2nd ed. New York, NY: John Wiley & Sons; 2000.

[4] Lipsitz SR, Fitzmaurice GM, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. *J R Stat Soc Ser C Appl Stat.* 1996;45(2):175–190.

[5] Archer KJ, Lemeshow S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata J.* 2006;6(1):97–105.

[6] Archer KJ, Lemeshow S, Hosmer DW. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Comput Stat Data Anal.* 2007;51:4450–4464.

[7] Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. *Stat Med.* 2008;27:4238–4253.

[8] Cool G, Lebel A, Sadiq R, Rodriguez MJ. Modelling the regional variability of the probability of high trihalomethane occurrence in municipal drinking water. *Environ Monit Assess.* 2015;187(12):746.

[9] Sturdivant RX, Hosmer DW. Smoothed residual based goodness-of-fit statistics for logistic hierarchical regression models. *Comput Stat Data Anal.* 2007;51(8):3898–3912.

[10] Perera AAPNM, Sooriyarachchi MR, Wickramasuriya SL. A goodness of fit test for the multilevel logistic model. *Commun Stat Simul Comput.* 2016;45(2):643–659.

[11] Fernando G, Sooriyarachchi R. The development of a goodness-of-fit test for high level binary multilevel models. *Commun Stat Simul Comput.* 2022;51(5):2710–2730.

[12] Linden A. MLM_GOF: Stata module for computing the goodness-of-fit test after mixed-effects logistic regression. Statistical Software Components S459670. Boston College Department of Economics; 2026.

[13] Rosner B, Glynn RJ, Lee M-LT. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics.* 2003;59(4):1089–1098.

[14] Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York, NY: John Wiley & Sons; 1981.

[15] Linden A, Adler-Milstein J. Medicare disease management in policy context. *Health Care Financ Rev.* 2008;29(3):1–11.

[16] Biuso TJ, Butterworth S, Linden A. A conceptual framework for targeting prediabetes with lifestyle, clinical and behavioral management interventions. *Dis Manag.* 2007;10(1):6–15.

[17] Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. *J R Stat Soc Ser A Stat Soc.* 1996;159(3):505–513.

[18] Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal.* 2006;1(3):473–514.

[19] Linden A, Roberts N. A user's guide to the disease management literature: recommendations for reporting and assessing program outcomes. *Am J Manag Care.* 2005;11(2):113–120.

Table 1: Type I Error Simulation Design

Factor	Values
Number of clusters (J)	15, 30, 50
Subjects per cluster (K)	5, 10
Observations per subject (n)	20
Intraclass correlation (ICC)	0.10, 0.30
Total scenarios	24
Replications per scenario	1000*
Significance level	0.05

* Monte Carlo bounds for 1000 replications at $\alpha = 0.05$: $[0.036, 0.064]$.

Table 2: Power Simulation Design

Misspecification	DGP	Fitted model	Parameter	Values
Omitted quadratic	Includes $\beta_3 x_1^2$	Omits $\beta_3 x_1^2$	β_3	0.02, 0.05, 0.10, 0.15
Omitted interaction	Includes $\beta_3(x_1 \times x_2)$	Omits $\beta_3(x_1 \times x_2)$	β_3	0.3, 0.6, 0.9
Omitted level	3-level model	2-level model	σ_{extra}	0.5, 1.0, 1.5

Fixed conditions: $J = 30$, $K = 5$, $n = 20$, ICC= 0.20; 1000 replications per scenario.

Table 3: Group Selection Simulation Design

Cluster designs		
Design	Clusters	Cluster sizes
Unbalanced	10 small; 40 large	Small: $n_{\text{small}} \in \{3, 5, 6, 8, 10\}$; Large: $n = 20$
Balanced	50 clusters	All: $n_{\text{small}} \in \{3, 5, 6, 8, 10\}$
Evaluation settings		
Grouping rules compared		$G = \min(10, n_{\min})$ vs. $G = 10$
Outcome measures		Type I error; failure rate
Replications per scenario		1000
Failure definition		Missing p-value

Table 4: Empirical power of the goodness-of-fit test against three types of model misspecification

Misspecification	Parameter	Effect size	Power
Omitted quadratic	β_3	0.02	0.074
		0.05	0.278
		0.10	0.923
		0.15	0.998
Omitted interaction	β_3	0.3	0.143
		0.6	0.544
		0.9	0.874
Omitted level	σ_{extra}	0.5	0.043
		1.0	0.060
		1.5	0.044

Fixed conditions: $J = 30$, $K = 5$, $n = 20$, ICC$= 0.20$. Power for omitted level scenarios is indistinguishable from the nominal $\alpha = 0.05$.

Table 5: Type I error rates: data-driven $G = \min(10, n_{\min})$ versus $G = 10$, by minimum cluster cell size

		Unbalanced design		Balanced design	
$n_{\min}$	G	Rej. default	Rej. $G = 10$	Rej. default	Rej. $G = 10$
3	3	0.040	0.034	0.028	—
5	5	0.044	0.050	0.034	—
6	6	0.062	0.044	0.050	—
8	8	0.032	0.040	0.038	—
10	10	0.048	0.048	0.030	0.030

Unbalanced: $J_{\text{small}} = 10$ clusters with $n_{\min}$ obs and $J_{\text{large}} = 40$ clusters with $n = 20$. Balanced: all $J = 50$ clusters have $n_{\min}$ obs. "—" = $G = 10$ produced no valid result for all replications. Monte Carlo bounds: $[0.036, 0.064]$.

Table 6: Mixed-effects logistic regression results: pre-diabetes reversal study ($N = 905$)

Parameter	OR	SE	95% CI	p
Fixed effects				
Intervention	5.786	1.293	(3.734, 8.966)	<0.001
BMI centered (per unit above 30 kg/m^2)	0.901	0.032	(0.840, 0.967)	0.004
Visit number	1.313	0.136	(1.071, 1.609)	0.009
Intercept (baseline odds)[a]	0.147	0.042	(0.084, 0.256)	<0.001
Random effects (SD)				
Family-level: intercept	0.741	0.298	(0.336, 1.631)	—
Family-level: slope (visit)	0.454	0.102	(0.293, 0.704)	—
Family-level: intercept–slope corr.	−0.435	0.303	(−0.833, 0.261)	0.152
Subject-level: intercept	0.569	0.174	(0.313, 1.036)	—
Goodness-of-fit (`mlm_gof`)				
Groups used (G)	5			
Wald χ^2 ($df = 4$)	2.922			0.404

[a] Baseline odds conditional on zero random effects. OR = odds ratio; SD = standard deviation; 95% CIs for random effects are likelihood-ratio-based. $G = \min(10, n_{\min}) = \min(10, 5) = 5$.

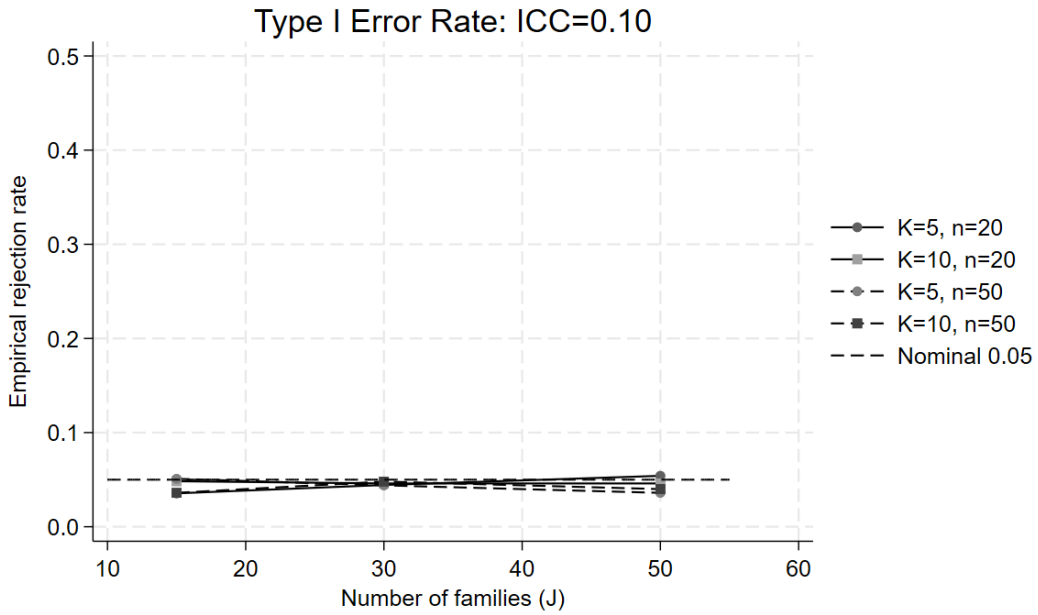

Figure 1: Empirical Type I error rate by number of families (J), subjects per family (K), and observations per subject (n). ICC = 0.10. Dashed line = nominal 0.05. All values within Monte Carlo bounds $[0.036, 0.064]$.

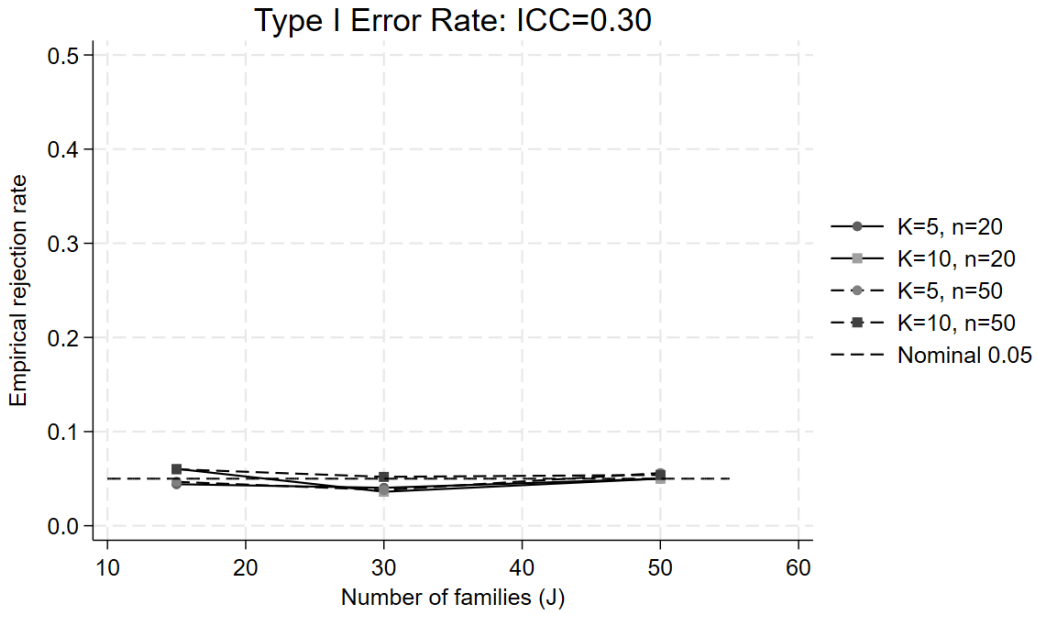

Figure 2: Empirical Type I error rate by number of families (J), subjects per family (K), and observations per subject (n). ICC = 0.30. Dashed line = nominal 0.05. All values within Monte Carlo bounds $[0.036, 0.064]$.