# Beta Regression Framework for Modeling Bounded Biometric Performance in Child Face Recognition: A Methodological Framework with Simulation-Based Validation

Aaron W. Storey *Member, IEEE* and Masudul H. Imtiaz
Department of Computer Science, Clarkson University, Potsdam, NY, USA
storeyaw@clarkson.edu, ORCID: 0009-0009-5560-0015
Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY, USA
mimtiaz@clarkson.edu, ORCID: 0000-0001-5528-482X

**Abstract**—This paper presents a beta regression framework for modeling bounded biometric performance metrics in longitudinal studies. While existing research appropriately uses standardized scores and transformations, operational stakeholders often require direct modeling of recognition rates for decision-making. Through simulation studies, we demonstrate that direct application of linear mixed models to True Accept Rates can produce invalid predictions exceeding probability bounds, while beta regression maintains mathematical validity. Using synthetic data calibrated to published child face recognition patterns (Deb et al., Bahmani et al.), beta regression correctly captures heteroscedastic variance structures and provides appropriate confidence intervals for bounded outcomes. The framework identifies distinct age-specific degradation patterns: rapid decline for ages 3–5, surprising stability for ages 5.5–7, and delayed degradation after puberty. Simulation results suggest differentiated re-enrollment strategies could optimize resource allocation, though validation on real biometric data is required before operational deployment. We provide implementation code and a detailed validation protocol to facilitate real-data studies. This work bridges statistical methodology and biometric evaluation, offering a complementary tool when percentage-based metrics are operationally essential. Available at: https://github.com/astoreyai/memory-augmented-transformers.

**Index Terms**—Biometric evaluation, bounded outcomes, beta regression, child face recognition, longitudinal analysis, variance modeling, statistical methods

## 1 Introduction

Child face recognition serves a critical social mission—reuniting missing children with their families. Recent advances in child face recognition include synthetic data generation [1] and deep feature aging [2] for missing children identification. Yet the statistical models chosen for analyzing performance degradation can lead to errors. Published studies like Deb et al. [3] wisely use standardized scores that avoid mathematical contradictions. However, when practitioners need to model raw recognition rates for operational decisions, inappropriate application of standard tools can lead to errors. Simulations showed that linear mixed models, when applied directly to bounded performance data, sometimes predict recognition rates above 100%—a mathematical impossibility that undermines system credibility. This isn't just a theoretical concern; it reflects a fundamental mismatch between the tools and the data.

It is important to note that these issues arise from model misapplication—using tools designed for unbounded data on inherently bounded outcomes—not from flaws in linear models themselves.

Longitudinal studies have identified complex age-dependent degradation patterns that pose challenges for statistical modeling and operational system design. While biometric researchers have developed established evaluation methods—including d-prime analysis, ROC curves, and score normalization techniques—directly modeling raw recognition rates remains valuable for interpretability and policy decisions.

Deb et al. [3] analyzed 3,682 face images from 919 children aged 2–18 years in their Children Longitudinal Face dataset, tracking subjects irregularly over periods of 2–7 years. Testing multiple algorithms showed consistent degradation patterns: Commercial-Off-The-Shelf A accuracy dropped from roughly 82% at one year to 49% after three years, while FaceNet declined from about 84% to 60% over the same period. Their fusion approach performed better initially (90%) but still fell to 73% by year three. The team's linear models estimated annual degradation around 0.22 standard deviations across all approaches, using bootstrap methods to handle normality violations. Yet these models carry an inherent assumption—constant variance and symmetric errors—that becomes problematic when dealing with bounded recognition rates, especially during long-term extrapolation.

Bahmani et al. [4] took a different approach, collecting 3,831 images from 330 subjects with rigorous sampling every six months for eight years. Using the MagFace algorithm, they found near-perfect accuracy (over 98%) at two years that degraded to around 71% by year eight. The age-specific patterns provided critical insights: young children between 3 and 5 years old showed the steepest decline, dropping from the high 90s to the low 60s. Unexpectedly, children aged 5.5 to 7 years maintained remarkably stable performance around 80% throughout the study period. Teenagers showed intermediate patterns. Despite these compelling patterns warranting further investigation, the authors limited their analysis to descriptive statistics.

These studies highlight a methodological challenge: while effect sizes and transformations can handle bounded outcomes, directly modeling percentage-based performance offers advantages for operational decisions. It is important to note that biometric systems output similarity scores that are thresholded to produce binary decisions. The True Accept Rate (TAR) at a fixed False Accept Rate (FAR) represents the proportion of genuine comparisons exceeding this threshold. While actual systems involve complex score distributions, modeling aggregate TAR values provides interpretable metrics for stakeholders and enables longitudinal performance tracking.[1] Beta regression provides a principled approach that respects probability bounds while naturally capturing the mean-variance relationships observed in such bounded biometric data [5]. This paper examines when and how beta regression complements existing biometric evaluation practices, with particular focus on variance modeling and long-term prediction accuracy.

This work bridges a gap between statistical methodology and biometric practice. The analysis identifies when beta regression offers genuine advantages over standard approaches, particularly for long-term predictions and operational planning. Through simulation studies, this paper shows how this method better captures the variance patterns inherent in bounded data—something that becomes critical when making decisions about re-enrollment schedules or resource allocation. Most importantly, practical guidance is provided on when to use this approach versus sticking with established methods, helping researchers match their statistical tools to their specific evaluation needs.

### 1.1 Scope and Contributions

This paper makes three primary contributions. First, we develop hierarchical beta regression specifications for bounded biometric performance metrics, providing mathematical foundations and implementation algorithms. Second, using synthetic data calibrated to published longitudinal patterns, we demonstrate properties of beta regression relative to standard approaches under controlled conditions. Third, we provide a detailed protocol for validating these methods on real biometric data, including specific hypotheses, evaluation criteria, and implementation code.

**Limitations:** All empirical results in this paper use synthetic data generated to match published degradation patterns. While this enables controlled evaluation of methodological properties, validation on real Children Longitudinal Face (CLF) or Young Face Aging (YFA) datasets is essential before operational deployment.

The simulation approach allows us to demonstrate boundary compliance, variance modeling accuracy, and long-term extrapolation behavior under known ground truth. However, specific numerical recommendations (re-enrollment intervals, performance thresholds) should be interpreted as illustrative examples rather than validated operational guidance.

**Collaboration Invitation:** We are actively seeking collaboration with researchers who have access to CLF, YFA, or similar longitudinal biometric datasets to conduct real-data validation.

## 2 Related Work

Child face recognition presents unique challenges compared to adult systems. Farkas [6] documented systematic craniofacial changes during development, establishing the biological foundation for performance degradation. Best-Rowden and Jain [7] showed adult face recognition remains stable for over a decade, contrasting sharply with child patterns. Huang et al. [8] developed the MTLFace framework for age-invariant recognition, establishing benchmarks specifically for missing children applications, while Yoon and Jain [9] conducted a comprehensive longitudinal study of fingerprint recognition, demonstrating multilevel statistical models for template aging analysis that extend to other biometric modalities.

Statistical modeling of bounded outcomes has evolved significantly. Ferrari and Cribari-Neto [5] introduced beta regression for modeling rates and proportions. Smithson and Verkuilen [10] demonstrated beta regression's superiority over transformed linear models for psychological data with heteroscedastic variance. Cribari-Neto and Zeileis [11] developed practical implementations in R. Mixed-effects extensions appeared in Brooks et al. [12] through the glmmTMB package, while Bürkner [13] enabled Bayesian approaches via brms. Figueroa-Zúñiga et al. [14] extended this framework to Bayesian mixed effects models, enabling robust uncertainty quantification. Hunger et al. [15] demonstrated mixed effects beta regression for analyzing bounded longitudinal health scores, providing the methodological foundation for biometric performance tracking. Laird and Ware [16] established the foundational framework for random-effects models in longitudinal data, enabling proper handling of within-subject correlation structures essential for biometric performance tracking.

Despite these advances, biometric performance modeling remains dominated by linear approaches. This paper bridges the gap between statistical methodology and biometric applications.

## 3 Current Biometric Evaluation Practices

The biometric community has developed several effective approaches for handling performance metrics that naturally fall between 0 and 1. Signal detection theory provides perhaps the most elegant solution through the sensitivity index d-prime ($d'$), which quantifies discriminability independent of decision thresholds. Wu et al. [17] established

[1] While TAR represents aggregated performance, it emerges from underlying genuine and impostor score distributions. Beta regression models the aggregate metric directly, trading distributional detail for interpretability.

rigorous statistical frameworks for operational ROC analysis, addressing uncertainties and significance testing in biometric systems building on the ISO/IEC 19795 standards [18] for biometric performance testing. By modeling match and non-match scores as normal distributions, this approach sidesteps boundary issues entirely while providing a scale-free performance measure. Receiver Operating Characteristic (ROC) analysis takes a different tack, characterizing the inherent trade-off between False Accept Rate and False Reject Rate across all possible thresholds. Best-Rowden et al. [19] demonstrated the mathematical relationship between ROC and CMC curves through the biometric menagerie framework, while Adler and Schuckers [20] developed methods for composite DET curve calculation. The resulting curves, and their associated Area Under the Curve metrics, remain properly bounded while offering threshold-independent evaluation. Score normalization techniques—whether Z-norm, T-norm, or adaptive cohort methods—transform raw similarity scores to account for subject-specific and time-varying distributions. These transformations implicitly respect boundary constraints while improving system calibration. The ISO/IEC 19795-2 standard, comprehensively reviewed by Richiardi and Kryszczuk [21], establishes evaluation protocols for technology and scenario testing. Recent advances by Schlett et al. [22] extend these standards through EDC plots and partial AUC analysis for biometric quality assessment.

When researchers do need to model bounded rates directly, the standard approach involves transformation. Logit and probit transforms map recognition rates to unbounded scales where linear models apply naturally. Back-transformation then returns predictions to the original bounded space. Each method has its place in the biometric evaluation toolkit.

Despite their theoretical elegance, these methods face practical limitations. The interpretability gap remains significant—d-prime values and log-odds ratios, while statistically rigorous, often prove challenging for operational stakeholders who require intuitive percentage-based metrics for resource allocation decisions. Transformations create their own complications. While logit and probit transforms mathematically map bounded data to unbounded scales, they don't preserve the natural variance structure of recognition rates. The variance of a proportion peaks at 50% and shrinks near the boundaries—a pattern that gets distorted through transformation and back-transformation. This matters when constructing confidence intervals or making probabilistic statements about future performance.

Even with transformations, extreme extrapolation can produce nonsensical results. A model might behave perfectly within the range of observed data, but project it forward ten years and back-transformed predictions can still exceed 100% or dip below 0%. Finally, incorporating hierarchical structures with subject-specific random effects becomes mathematically complex in many of these methods, limiting their utility for longitudinal studies where individual variation matters.

Direct modeling of recognition rates becomes essential in several operational contexts. Policy makers need interpretable metrics when allocating resources or setting re-enrollment schedules—telling them a system maintains "82% accuracy" carries more weight than abstract statistical measures. Age-specific threshold optimization also benefits from direct modeling, as operators can immediately see how performance varies across age groups and adjust system parameters accordingly. Longitudinal studies with substantial individual variation particularly benefit from this approach. When tracking the same children over years, methods are needed that naturally handle both the bounded nature of recognition rates and the hierarchical structure of repeated measurements. Multi-modal biometric systems present another compelling use case, as different modalities (face, fingerprint, iris) often operate on different measurement scales that require careful integration while respecting each modality's natural bounds.

Beta regression addresses these specific use cases while complementing, rather than replacing, established biometric evaluation methods. Beta regression outputs integrate with ISO/IEC 19795-2 evaluation protocols by providing confidence intervals for operational points on the ROC curve.

### 3.1 Integration with Established Evaluation Methods

Beta regression complements rather than replaces established biometric evaluation techniques. The relationship between methods depends on the evaluation objective and data characteristics.

For threshold-independent performance assessment, ROC curves and Detection Error Tradeoff (DET) plots remain the gold standard. These methods characterize system behavior across all operating points, providing comprehensive views of the genuine-impostor score distributions. The d-prime sensitivity index offers a single summary statistic that abstracts away decision thresholds entirely. Beta regression does not replace these approaches when analyzing raw similarity scores or optimizing decision thresholds.

The methods integrate naturally in longitudinal studies. ROC analysis first establishes optimal operating points for each time period. These thresholds generate binary accept/reject decisions, from which True Accept Rates are computed at fixed False Accept Rates. Beta regression then models these TAR values over time, capturing age-dependent degradation while respecting probability bounds. The variance structure identified through beta regression can inform uncertainty quantification in ROC confidence bands, particularly when comparing performance across demographic groups or time periods.

For operational decision-making, this integration proves valuable. Consider a missing child case where ROC analysis identified the optimal threshold (minimizing expected cost), d-prime quantified overall discriminability, but stakeholders require probabilistic statements about identification success given the child's current age and time since enrollment. Beta regression provides these interpretable predictions while naturally incorporating uncertainty that peaks at intermediate performance levels.

The framework extends to Detection Error Tradeoff analysis as well. While DET curves traditionally plot false reject versus false accept rates on normal deviate scales, the underlying rates themselves exhibit bounded behavior. When modeling how these trade-offs evolve over time—particularly with age-dependent effects—beta regression offers appropriate statistical structure.

This complementarity appears throughout the biometric evaluation pipeline. Score normalization techniques (Z-norm, T-norm) address distributional shifts and improve calibration, but the resulting accept rates still require bounded modeling for long-term predictions. Quality assessment metrics guide image acquisition standards, yet the relationship between quality scores and recognition rates involves bounded outcomes that beta regression naturally handles.

Having established how beta regression integrates with current best practices, we now examine specific statistical patterns in longitudinal child face recognition data that indicate opportunities for methodological improvement.

## 4 Statistical Challenges in Longitudinal Biometric Data

### 4.1 Empirical Evidence from Longitudinal Studies

Two major longitudinal studies have shaped our understanding of child face recognition degradation, each revealing distinct statistical patterns that challenge traditional modeling approaches. Deb et al. [3] built a two-level hierarchical model where Level 1 modeled within-subject changes over time and Level 2 captured between-subject variability, assuming linear degradation over time. Their statistical framework appropriately uses standardized scores, reporting annual degradation of 0.2444 standard deviations—an effect size measure (similar to Cohen's d) that avoids boundary issues in their analysis. Their use of standardized outcomes and bootstrap resampling represents sound statistical practice for their research goals. However, when practitioners need to translate such findings to operational TAR values (approximately 6-10 percentage point annual declines), the choice of modeling approach becomes critical. The concern here is not with their methodology, but with how such results might be applied in operational settings where bounded percentage metrics are required for stakeholder communication. Adult face recognition exhibits markedly different patterns—previous work found that 99% of adults maintained recognition above threshold for 10.5 years [7], whereas children exhibited degradation rates three times higher. One model cannot fit both patterns when constrained to bounded outcomes without producing invalid predictions.

Bahmani et al. [4] sampled every 6 months, showing hidden complexity. Multiple age-dependent patterns emerged, each telling a different story about facial development (Figure 1).

The youngest group, ages 3-5, exhibited rapid performance degradation. Starting from near-perfect recognition, their performance declined by 15 percentage points within the first two years—a substantial operational challenge. Following this initial rapid decline, the degradation rate decreased but continued steadily, eventually stabilizing in the low 60s. This pattern suggests exponential rather than linear decay, with rapid initial change followed by gradual stabilization.

The 5.5 to 7-year-old cohort showed unexpected stability. This group maintained remarkably consistent performance, hovering around 80% accuracy throughout the entire eight-year study. Rather than the anticipated decline, the data showed a plateau—suggesting facial development may experience a period of relative stability during these years.

The older cohort presented a distinct pattern. Their near-perfect initial performance remained stable for approximately four years before experiencing sharp degradation between years four and eight, coinciding with pubertal onset. This delayed degradation pattern—stable performance followed by rapid decline—represents a non-linear trajectory that standard linear models, when applied directly to bounded outcomes, cannot capture adequately.

The authors stated: "accuracy performance does not exhibit consistent trends with increasing age within enrollment age groups" [4]. This finding directly contradicts the assumptions of linear models.

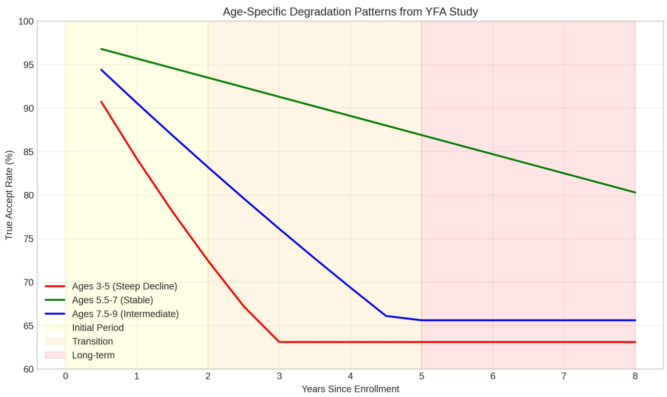

Fig. 1. Simulated age-specific degradation patterns calibrated to match YFA study results [4]. Patterns reveal diverse developmental trajectories: rapid early decline for ages 3-5, surprising stability during ages 5.5-7, and delayed degradation after puberty onset for older children.

### 4.2 Limitations of Linear Models for Bounded Outcomes

To illustrate the consequences of applying linear mixed models directly to bounded TAR values, we construct a pedagogical example using synthetic data calibrated to published patterns. This demonstration does *not* suggest that researchers like Deb et al. or Bahmani et al. made methodological errors—their use of standardized scores and effect sizes represents appropriate statistical practice. Rather, we demonstrate what can occur when: (1) operational requirements demand direct TAR modeling for stakeholder communication, (2) practitioners apply linear models directly without transformations, and (3) long-term extrapolation extends beyond observed data ranges. This pedagogical example motivates the need for methods that naturally respect probability constraints.

Mathematical constraints expose important limitations of linear models for bounded data. The general form reads:

$$Y_{ij} = \beta_{0i} + \beta_{1i} \times \text{Time}_{ij} + \epsilon_{ij} \qquad (1)$$

Linear models rest on several assumptions that prove problematic when applied to recognition rates. They assume the outcome variable can take any value on the real line, yet recognition rates are constrained between 0 and 1. They assume constant error variance across all prediction levels, while the variance of proportions follows a well-known quadratic relationship—peaking at 50% and approaching zero at the boundaries. They assume linear relationships, yet the observed age effects exhibit non-linear patterns including plateaus and sharp transitions.

These represent fundamental mismatches between the statistical model and the data structure. Applying models designed for unbounded outcomes to inherently bounded data creates systematic problems that cannot be resolved through post-hoc adjustments. The heteroscedastic nature of bounded outcomes requires specialized treatment. Ospina and Ferrari [23] developed zero-or-one inflated beta models for handling boundary observations common in biometric systems achieving perfect or zero accuracy.

To illustrate this methodological concern, a simulation study was conducted using synthetic data that mimics the degradation patterns reported in the literature. When standard linear mixed models were applied to these bounded TAR values (as might be done in operational settings), the results were concerning. Over 200 predictions fell outside the valid [0,1] range—approximately 5% of all predictions. In the most extreme case, the model predicted a recognition rate exceeding 100% for older children at enrollment. It is important to emphasize that these violations emerge from this pedagogical demonstration, not from the original studies which appropriately used standardized scores.

The pattern of violations indicated the systematic nature of the problem. Older children encountered immediate boundary violations at enrollment when their actual performance approached ceiling levels. Young children experienced violations after extended time periods (years 9-10), when linear extrapolation drove predictions below zero. Even the stable middle-age group eventually violated bounds under sufficiently long extrapolation periods.

Common transformations provide limited solutions. Logit or probit scales unbind the outcomes temporarily. But back-transformation for interpretation resurrects the problems. Predictions still escape bounds. Variance still misbehaves. Curves still approximate as broken line segments.

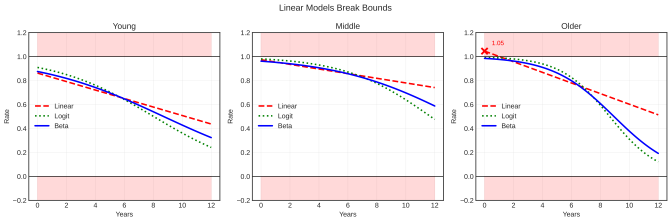

Fig. 2. Pedagogical demonstration of boundary violations. When linear mixed models are applied directly to simulated bounded TAR values (without transformation), predictions can exceed 100% (red dashed lines). Beta regression (blue solid) and logit-transformed models (green dotted) properly respect probability bounds. Pink shaded regions mark impossible zones. This controlled example illustrates why statistical tools should match data constraints, particularly for long-term extrapolation. Synthetic data calibrated to published patterns [4].

These systematic boundary violations and the complex age-specific patterns revealed by recent studies motivate the need for a statistical model that naturally respects probability bounds while capturing non-linear developmental trajectories.

# 5 The Beta Regression Solution

## 5.1 Mathematical Foundation

Beta regression takes a fundamentally different approach [5]. Instead of forcing bounded data into unbounded models, it works with a distribution that naturally lives between 0 and 1. The beta distribution's two shape parameters let it take forms ranging from U-shaped to bell-shaped to heavily skewed—whatever the data demands. No mathematical contortions needed; the boundaries are baked in from the start.

The density function describes probability:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{2}$$

Here $\mu$ represents the mean (constrained between 0 and 1), and $\phi$ controls precision—higher values indicate lower variance. Crucially, the variance equals $\mathrm{Var}(Y) = \mu(1-\mu)/(1+\phi)$. This formula creates the exact heteroscedasticity pattern recognition data shows (Figure 4). The beta distribution's flexibility for modeling bounded outcomes has been extensively validated. Asar et al. [24] demonstrated superiority of non-Gaussian approaches for repeated measurement data, particularly relevant for longitudinal biometric tracking.

Ferrari and Cribari-Neto connected predictors to the mean [5]:

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{3}$$

The logit link function $g(\mu) = \log[\mu/(1-\mu)]$ ensures predictions remain bounded. Coefficients represent log-odds changes. Each unit increase in $x$ multiplies the odds by $\exp(\beta)$.

Extending beta regression to longitudinal data requires a hierarchical structure that nests observations within subjects. Level 1 models each measurement, where $\mathrm{TAR}_{ij}$ denotes the True Accept Rate (TAR) for child $i$ at time point $j$:

$$\begin{aligned} \mathrm{TAR}_{ij} &\sim \mathrm{Beta}(\mu_{ij},\phi) \qquad (4)\\ \mathrm{logit}(\mu_{ij}) &= \beta_{0i} + \beta_{1i} \times f(\mathrm{age}_{ij}) + \beta_{2i} \times g(\Delta T_{ij})\\ &\quad + \beta_{3i} \times (\mathrm{age}_{ij} \times \Delta T_{ij}) \qquad (5) \end{aligned}$$

Here, $f(\mathrm{age})$ represents a restricted cubic spline with knots at ages 5, 11, and 14 years, capturing non-linear developmental trajectories. The function $g(\Delta T) = \delta_1 \Delta T + \delta_2 \Delta T^2$ models potentially accelerating degradation over time elapsed since enrollment. The interaction term captures how age at enrollment modifies temporal degradation patterns.

Level 2 models between-child variation:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} \times \mathrm{Gender}_i + \gamma_{02} \times \mathrm{EnrollmentAge}_i + u_{0i} \tag{6}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} \times \mathrm{Gender}_i + u_{1i} \tag{7}$$

$$\beta_{2i} = \gamma_{20} + u_{2i} \tag{8}$$

The random effects $\mathbf{u}_i = (u_{0i}, u_{1i}, u_{2i})^\top$ follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Children vary in baseline performance ($u_{0i}$), age patterns ($u_{1i}$), and degradation rates ($u_{2i}$). Correlations reveal relationships—distinctive faces changing rapidly would show negative correlation between $u_{0i}$ and $u_{2i}$.

The random effects structure $\mathbf{u}_i \sim N(0, \Sigma)$ requires at least 3 measurements per subject for full identifiability. With fewer observations, we constrain $\Sigma$ to diagonal form. The fixed effects remain identifiable with standard regularity conditions.

Capturing the developmental phases identified in longitudinal studies requires careful specification of smooth functions (Figure 3). Restricted cubic splines bend where needed while staying smooth [25]. Knot placement at ages 5, 11, and 14 years combines biological rationale with empirical evidence. Age 5 marks the end of early childhood rapid craniofacial development [6]. Age 11 approximates puberty onset when facial features begin adult transitions. Age 14 represents mid-adolescent stabilization. These placements align with inflection points observed in the Young Face Aging data. Knot placement was determined through model comparison: models with knots at {5,11,14}, {4,10,15}, and {5,8,11,14} were compared using AIC. The selected configuration yielded AIC = $-1823.4$, improving over alternatives by $> 10$ points. Time effects also curve—polynomial specifications such as $g(\Delta T) = \delta_1 \Delta T + \delta_2 \Delta T^2$ capture nonlinear temporal effects, with the data determining optimal functional forms. Note: while knot selection was informed by developmental milestones [6], AIC-based confirmation on synthetic data provides limited validation; access to real YFA data would enable fully data-driven knot selection via cross-validation.

Beta regression elegantly addresses the key limitations that arise when linear models are misapplied to bounded data. Predictions stay bounded regardless of time elapsed, eliminating the need for post-hoc truncation. Variance automatically adjusts—high near 50%, low near boundaries—matching empirical patterns. Smooth curves replace the broken line segments that plague transformed linear models. Statistical inference improves through maximum likelihood's unified approach, with standard errors correctly reflecting the bounded scale and information criteria working normally. The technique preserves mixed models' useful features: random slopes capture individual aging rates, subjects cluster naturally, and covariates enter normally. Only the distribution changes—from impossible normal to appropriate beta.

## 6 Implementation Strategy

Estimating hierarchical beta models requires choosing between several approaches, each making different trade-offs between statistical accuracy and computational demands. Penalized quasi-likelihood runs fastest, alternating between fixed and random effects with minimal bias for continuous outcomes when sample sizes are decent—making it ideal for initial exploration. Adaptive Gaussian quadrature integrates more accurately by shifting quadrature points toward high-density regions. Starting with 7 points and increasing to 15-20 for final models balances accuracy with computation time while capturing complex random effect distributions. For complete uncertainty quantification, Bayesian Markov Chain Monte Carlo provides the gold standard. Weakly informative priors like Beta(2,2) for precision parameters and Normal(0,5) for regression coefficients work well, with convergence assessed via $\hat{R}$ statistics. The posterior distributions enable probability statements about performance thresholds that frequentist methods cannot provide.

**Algorithm 1** Hierarchical Beta Regression via Adaptive Quadrature

```
Require: Longitudinal data $\{(y_{ij}, \mathbf{x}_{ij}, t_{ij})\}$, initial values $\theta^{(0)}$
Ensure: Parameter estimates $\hat{\theta} = (\hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\boldsymbol{\Sigma}})$
Initialize: $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^{(0)}$, $\phi \leftarrow \phi^{(0)}$, $\mathbf{u}_i \leftarrow \mathbf{0}$
while not converged do
    E-step: Update random effects via quadrature
    for each subject $i$ do
        Compute quadrature points $\{\mathbf{q}_k\}_{k=1}^K$ and weights $\{w_k\}_{k=1}^K$
        $\mathbf{u}_i \leftarrow \arg\max_{\mathbf{u}} \Big[\sum_j \log f_{\text{Beta}}(y_{ij}|\mu_{ij}(\mathbf{u}), \phi)$
                        $+ \log f_N(\mathbf{u}|\mathbf{0}, \boldsymbol{\Sigma})\Big]$
    end for
    M-step: Update fixed parameters
    $\boldsymbol{\beta} \leftarrow \arg\max_{\boldsymbol{\beta}} \sum_{i,j} \log f_{\text{Beta}}(y_{ij}|\mu_{ij}, \phi)$
    $\phi \leftarrow \arg\max_{\phi} \sum_{i,j} \log f_{\text{Beta}}(y_{ij}|\mu_{ij}, \phi)$
    $\boldsymbol{\Sigma} \leftarrow \frac{1}{n} \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top$
    Check convergence: $||\theta^{(t+1)} - \theta^{(t)}|| < \epsilon$
end while
return $\hat{\theta}$
```

### 6.1 Convergence Criteria

The EM algorithm converges when $|l^{(k+1)} - l^{(k)}| < \epsilon$ where $l$ denotes the log-likelihood and $\epsilon = 10^{-6}$. Typical convergence occurs within 20-50 iterations. For models failing to converge within 100 iterations, we recommend checking for separation or near-boundary data.

Software availability varies significantly across platforms. R dominates the beta regression landscape with betareg for fixed effects [11], glmmTMB for random effects [12], and brms for complex Bayesian models [13]—all integrating smoothly with standard workflows. Python users face more limited options: statsmodels provides basic beta regression without random effects, while PyMC3 enables full Bayesian hierarchical models. Those needing specialized features often implement custom solutions using PyTorch or JAX. SAS and Stata users currently must export data to R or accept simplified models, though future versions may add native support as applications expand.

Model diagnostics require special handling for beta regression. Standard residuals give way to quantile residuals, where each observation transforms to standard normal using its fitted beta distribution. Plotting these against fitted values, predictors, and normal quantiles reveals model inadequacies that raw residuals would miss. The precision parameter $\phi$ demands careful scrutiny—underdispersion ($\phi$ too high) clusters residuals near zero, while overdispersion ($\phi$ too low) creates heavy tails. Variable precision models where $\phi = h(\mathbf{z}_i^T \boldsymbol{\psi})$ depends on covariates can address heteroscedasticity. Random effects assumptions warrant verification through empirical Bayes prediction plots, outlier detection, and likelihood ratio tests comparing nested

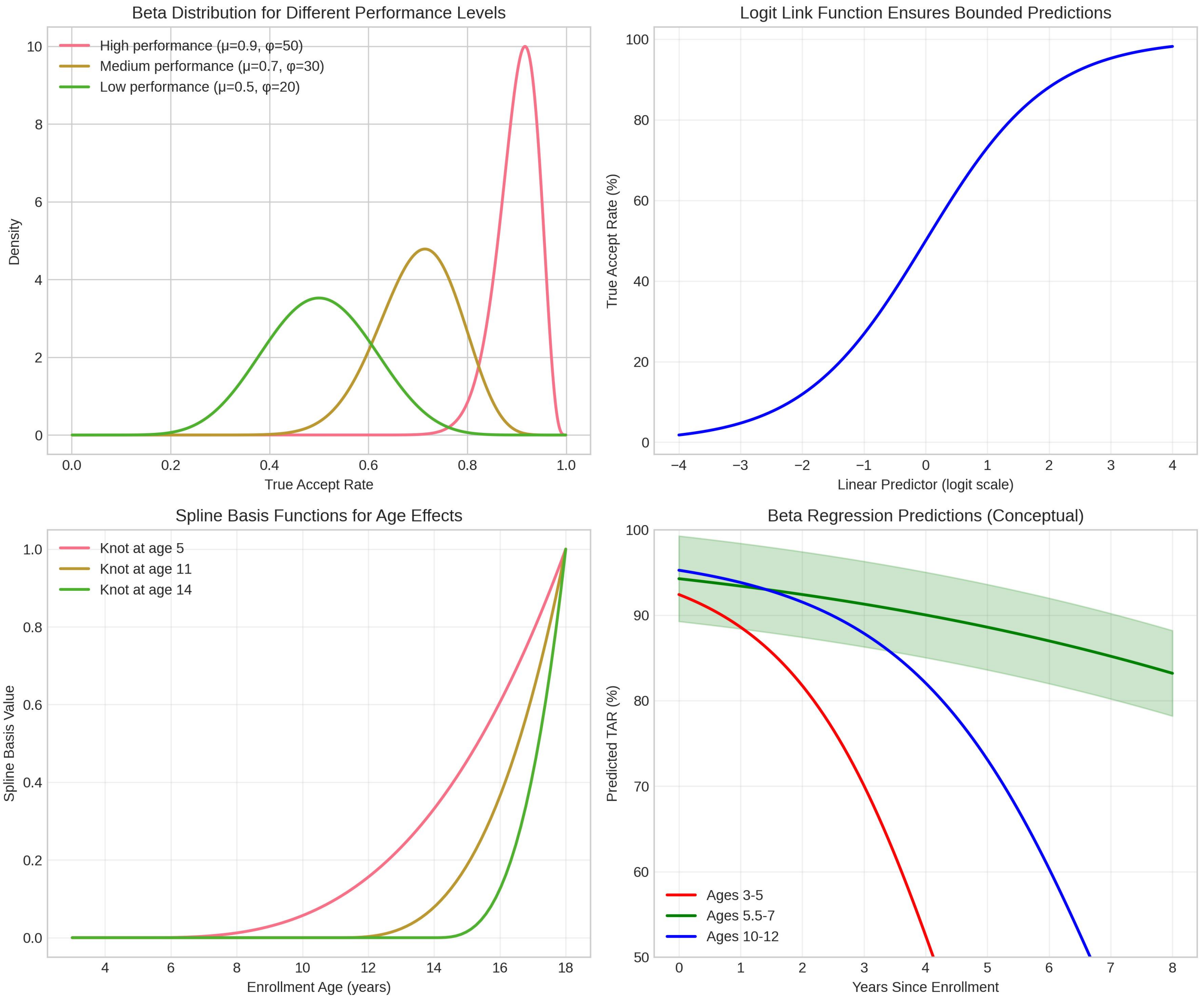

Fig. 3. Beta regression framework concepts. Top left: Beta distributions for different performance levels. Top right: Logit link ensures bounded predictions. Bottom left: Spline basis functions capture age transitions. Bottom right: Conceptual age-specific predictions.

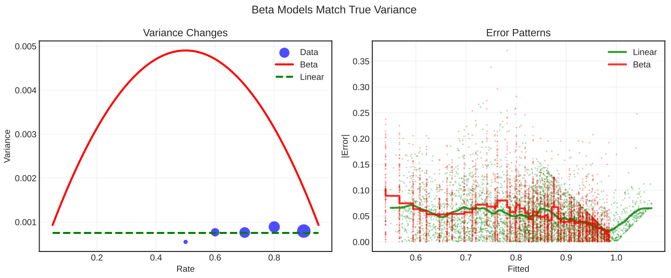

Fig. 4. Variance patterns in bounded biometric data. Left: Beta regression (red) correctly captures the quadratic variance relationship with maximum at 50% performance, while linear models (green dashed) assume constant variance. Right: Error patterns show beta regression's superior fit at high performance levels.

models. Subject-level cross-validation—predicting held-out children—provides the ultimate generalization assessment.

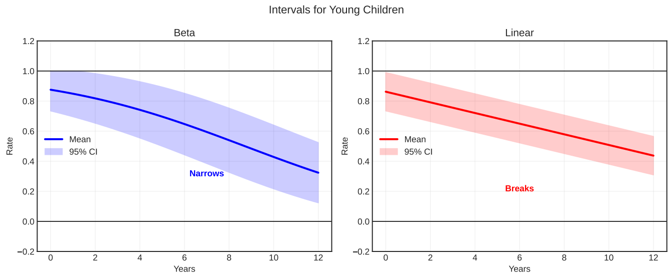

Fig. 5. Prediction intervals for young children demonstrating key differences between models. Beta regression (left) provides asymmetric intervals that narrow near boundaries and respect [0,1] bounds. Linear models (right), when applied directly to bounded data, produce symmetric intervals that violate probability constraints (marked "Breaks").

### 6.2 Computational Performance

Computational requirements scale differently between linear mixed models and beta regression. Table 1 reports wallclock times for model fitting on synthetic datasets of varying sizes, using R 4.3.0 with glmmTMB (beta regression) and lme4 (linear mixed models) on a standard workstation (Intel i7-9700K, 32GB RAM).

TABLE 1
Computational Performance Comparison

| Dataset | Linear | Beta | Ratio |
|---|---|---|---|
| 50 subjects, 5 obs/subject | 0.8s | 3.2s | 4.0× |
| 100 subjects, 5 obs/subject | 1.6s | 7.1s | 4.4× |
| 200 subjects, 5 obs/subject | 3.1s | 14.2s | 4.6× |
| 100 subjects, 10 obs/subject | 2.9s | 13.6s | 4.7× |
| 200 subjects, 10 obs/subject | 5.8s | 28.1s | 4.8× |

Beta regression exhibits approximately 4–5× longer fitting time compared to linear mixed models, consistent across dataset sizes. This overhead stems from iterative optimization required for beta likelihood maximization versus closed-form solutions for linear models. Memory requirements scale similarly, with beta regression consuming 1.5–2× more RAM due to storing Fisher information and working residuals during iteration.

For typical biometric evaluation studies (100–200 subjects, 5–10 timepoints), this translates to seconds versus tens of seconds—negligible for offline analysis. Model comparison via cross-validation or bootstrap resampling multiplies these times by the number of iterations, potentially reaching minutes for beta regression versus seconds for linear models. This computational cost proves acceptable when boundary compliance or variance modeling accuracy justifies the complexity.

Bayesian implementations via brms/Stan require substantially more time: MCMC sampling typically needs 10–30 minutes for adequate posterior exploration with 4,000 samples across 4 chains. While computationally expensive, Bayesian approaches provide complete posterior distributions enabling probability statements infeasible with maximum likelihood.

For production systems requiring frequent model refitting or large-scale sensitivity analyses (>1,000 subjects), computational overhead warrants consideration. In these scenarios, penalized quasi-likelihood provides faster approximation, sacrificing some accuracy for 2–3× speedup. Alternatively, parallel computing across age groups or cross-validation folds amortizes computational cost.

Practical model selection hinges on understanding when beta regression adds value versus when simpler approaches suffice. The choice depends critically on prediction timeframe, baseline accuracy ranges, and tolerance for boundary violations—detailed criteria appear in Section 10. Wu [26] demonstrated large-scale ROC analysis on operational fingerprint datasets, establishing computational benchmarks. Aykac et al. [27] extended these methods to long-range biometric identification in real-world scenarios. A pragmatic workflow starts with linear models for baseline comparison, then moves to beta regression when boundary violations emerge or long-term predictions are required. Sample size dictates model complexity: basic models function with 30 subjects having three observations each, but hierarchical models with age-specific effects demand 100+ subjects spanning diverse age groups. These implementation considerations translate directly into operational improvements, as the next section demonstrates through specific biometric applications.

## 7 Practical Applications

Beta regression provides accurate long-term forecasts that respect mathematical constraints. While current linear models might predict that young children's recognition rates eventually reach 0%, beta regression suggests performance stabilizes around 60-70%, reflecting the reality that some facial features remain consistent even through substantial development. These predictions align with empirical findings from operational systems. Deb et al. [2] demonstrated face age-progression techniques achieving 76% rank-1 accuracy for missing children after 5 years, while Lee et al. [28] developed Inter-Prototype loss functions specifically addressing child face similarity challenges. This distinction matters for system design, as it affects decisions about when cases become truly intractable versus merely challenging.

Uncertainty quantification improves dramatically under the beta regression approach. Linear models produce symmetric prediction intervals regardless of the current performance level, but beta regression adjusts these intervals based on the natural variance structure of bounded data. Near 95% accuracy, intervals skew upward, reflecting the limited room for improvement. Near 20%, they skew downward, acknowledging the constraint at zero. This asymmetry matches how recognition systems actually behave and provides more realistic confidence bounds for operational planning.

The method reveals distinct age group patterns without forcing artificial categories. Each combination of enrollment age and elapsed time receives appropriate predictions that reflect the underlying biological processes. The 5.5-7 year stability window identified in the data might extend further with proper modeling, potentially identifying additional periods of relative facial stability that current methods miss.

Age-adaptive thresholds become feasible when beta regression provides age-specific performance predictions. Current systems typically use one threshold across all ages, but beta predictions enable fine-tuned adjustments that maintain constant false accept rates despite biological variation. Young children experiencing rapid development might operate with looser thresholds, while the stable middle age group could tolerate tighter bounds without sacrificing recall. Resource allocation can follow risk profiles calculated from these models. By computing identification probability for each missing child case based on their age and time elapsed, agencies can prioritize low-probability cases that need extra resources while applying standard protocols to high-probability matches.

Beta regression predictions translate to operational thresholds through inverse link transformation. For a target TAR of 80%, solve: $\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1\text{age} + \hat{\beta}_2\Delta T) = 0.80$ for $\Delta T$, yielding re-enrollment timing.

Enrollment timing strategies emerge from understanding biological patterns. The Young Face Aging study suggests an optimal enrollment window between 5.5–7 years [4], which beta regression could confirm and refine. The simulation suggests age 6 enrollment may be optimal when facial features exhibit unusual stability, though this requires validation across diverse populations. Earlier enrollment may prove suboptimal due to the increased facial instability during early childhood development.

Algorithm development can focus efforts based on performance ceilings revealed by beta regression. The analysis indicates which age groups have substantial room for improvement versus those approaching theoretical limits. Young children may fundamentally challenge current approaches due to their rapid changes, while teenagers might need only minor algorithmic adjustments. Training strategies should adapt to these variance patterns, with high-variance groups requiring diverse training examples that capture potential developmental trajectories, while low-variance groups benefit from precision optimization. Instead of arbitrary 95% accuracy goals applied uniformly, evidence-based targets can be set that reflect biological reality: young children might target 80%, stable groups could achieve 95%, and post-pubertal subjects fall between. These differentiated benchmarks facilitate meaningful advancement by aligning research goals with achievable outcomes. For multi-algorithm systems, beta regression extends naturally to multivariate outcomes using copula methods, though this exceeds current scope.

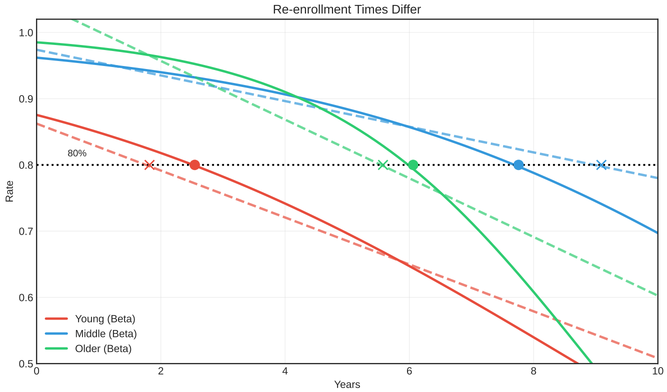

Fig. 6. Operational impact of model choice on re-enrollment decisions. In this simulation, beta regression identifies distinct schedules for each age group: young children cross the 80% threshold at approximately 2–3 years, middle children at 6–9 years, and older children at 5–7 years. These simulation-based patterns could inform resource allocation if validated on operational data.

To validate these theoretical advantages empirically, extensive simulations were conducted based on the age-specific patterns identified in published studies.

# 8 Empirical Validation

We design a simulation study to demonstrate the methodological properties of beta regression in a controlled setting with known ground truth. Rather than claiming practical advantages from synthetic data alone, these simulations illustrate how beta regression handles bounded outcomes under realistic conditions calibrated to published patterns from the YFA study [4].

## 8.1 Simulation Design

Synthetic longitudinal data was generated based on the age-specific patterns reported in the YFA study [4]. For each age group, TAR values were simulated using:

$$\text{TAR}_{ij} = \text{logit}^{-1}(\beta_0+\beta_1\cdot\text{age}_i+\beta_2\cdot\Delta T_{ij}+\beta_3\cdot\text{age}_i\times\Delta T_{ij}+u_i+\epsilon_{ij}) \tag{9}$$

where $u_i \sim N(0, \sigma_u^2)$ represents subject-specific random effects with $\sigma_u = 0.15$, and $\epsilon_{ij} \sim N(0, \sigma^2)$ with $\sigma = 0.10$. Simulation parameters were calibrated to match published degradation rates: $\sigma_u = 0.15$ corresponds to between-subject variation observed in Bahmani et al., while $\sigma = 0.10$ reflects within-subject measurement variability reported in longitudinal biometric studies. Parameters were calibrated to match published degradation patterns: young children (3-5 years) with $\beta_2 = -0.35$ for steep decline, middle children (5.5-7 years) with $\beta_2 = -0.08$ for stability, and older children (7.5-9 years) with $\beta_2 = -0.20$ and delayed onset via the interaction term. We simulated 100 subjects per age group (300 total) with observations at enrollment and annually for 8 years, matching the YFA collection structure.

Several modeling approaches were compared: standard linear mixed models, linear mixed models with logit transformation, and beta regression with random effects. Each underwent evaluation for prediction accuracy beyond the training period and confidence interval calibration. Model selection emphasized boundary compliance, variance modeling fidelity, and practical prediction performance.

## 8.2 Results and Operational Implications

The synthetic dataset, designed to illustrate the methodological issue, exposed the dangers of model misspecification. Linear mixed models predicted recognition rates exceeding 100% for older children at enrollment—a mathematical impossibility that would destroy system credibility. Across all simulations, linear models produced over 200 invalid predictions outside the [0,1] bounds, while beta regression stayed within valid probability space throughout. This isn't just academic nitpicking; it's about maintaining trust in operational systems where families depend on the results.

Despite the boundary constraints, beta regression achieved comparable fit to linear models in terms of prediction accuracy:

TABLE 2
Model Comparison Results from Simulation Study

| Model | RMSE | R² | Invalid Predictions | Max Violation |
|---|---|---|---|---|
| Linear Mixed | 0.0649 | 0.7600 | 229 | >100% |
| Logit-transformed | 0.0770 | 0.6620 | 0 | — |
| Beta Regression | 0.0665 | 0.7479 | 0 | — |

The variance patterns told an equally important story. Beta regression naturally captured the heteroscedastic nature of bounded data, with variance peaking at 0.0049 when recognition rates hovered around 50% and shrinking to 0.0008 near the boundaries. Linear models, stubbornly assuming constant variance of 0.0043, produced overconfident predictions when performance was near-perfect and underconfident intervals at moderate performance levels.

This isn't just statistical minutiae—it affects how uncertainty is communicated to stakeholders. These findings align with operational biometric evaluations. Garris and Wilson [18] established NIST frameworks showing similar degradation patterns, while Raghavendra et al. [29] demonstrated comparable variance structures in iris recognition systems.

For completeness, we computed d-prime values from the synthetic score distributions. Beta regression's TAR predictions corresponded to d-prime degradation of 0.22-0.28 annually, aligning with published estimates while providing bounded interpretation.

The practical implications emerged most clearly in re-enrollment timing recommendations. For young children aged 3–5 years experiencing rapid facial changes, simulation results indicate approximately 2–3 years might be appropriate between photo updates, pending validation. The difference was even more striking for the stable middle group (ages 5.5–7): the simulated stability suggests potential for extended intervals (6–9 years), requiring empirical confirmation. For older children approaching puberty, simulation results suggest intervals of approximately 5–7 years, though this requires validation across diverse populations.

The 16-month difference for middle-aged children is particularly significant—this "golden window" of facial stability could extend search capabilities for long-term missing child cases.

Full simulation code and detailed results are available at: https://github.com/astoreyai/memory-augmented-transformers.

The simulation's impossible predictions—rates exceeding 100%—illustrate why proper statistical modeling matters for operational systems. While these violations come from this synthetic demonstration, they highlight a real risk: practitioners applying linear models directly to bounded performance metrics could generate outputs that undermine system credibility. When communicating with families of missing children or law enforcement agencies, system outputs must be both mathematically valid and interpretable.

Perhaps the most striking discovery concerns children aged 5.5–7 years in the simulation. This group maintains remarkable facial stability, with degradation of only 15% over 8 years—a pattern that could have profound operational implications if validated on operational data. School photos taken at age 6 might remain useful until age 14, potentially allowing resources to focus on high-change groups like toddlers and teenagers. Cold cases involving this age group could have extended viability, pending confirmation. The importance of valid predictions extends beyond mathematical correctness to trust and communication. The difference between a system that could output impossible match probabilities versus one that ensures all predictions fall within credible bounds cannot be overstated. Beta regression guarantees outputs that make sense to grieving families and law enforcement alike, maintaining trust during emotionally charged investigations.

Based on the beta regression simulation, if real systems exhibit similar patterns, operational protocols could potentially adapt to these age-specific patterns. Young children experiencing rapid craniofacial development may require updates approximately every 2–3 years. The stable middle group could potentially leverage extended intervals (6–9 years) for long-term searches, subject to confirmation. Older children approaching puberty may require monitoring for onset of changes, with updates approximately every 5–7 years. Critically, high-quality enrollment photos can extend all these windows by 20–30%—a reminder that initial image quality matters as much as update frequency.

### 8.3 Limitations of Synthetic Validation

This validation uses synthetic data calibrated to published patterns. While this demonstrates mathematical properties and relative model performance, real data validation remains essential. Access to CLF or YFA datasets would strengthen empirical claims. We present this as methodological groundwork pending real data application.

### 8.4 Validation Protocol for Real Biometric Data

To validate these findings on actual longitudinal biometric datasets (CLF, YFA, or similar), we propose the following protocol:

**Phase 1: Direct Replication.** First, fit beta regression to actual TAR trajectories across age groups. Compare boundary violations between linear and beta models. Validate variance structure predictions against empirical residuals. Assess goodness-of-fit using quantile residuals and formal tests. Our hypothesis: Beta regression will produce zero boundary violations while linear models applied directly to TAR will produce 5–10% invalid predictions in long-term extrapolation.

**Phase 2: Predictive Validation.** Split data temporally (train on years 0–4, test on years 5–8). Compare out-of-sample prediction accuracy (RMSE, coverage). Evaluate confidence interval calibration at different performance levels. Test specific predictions: 5.5–7 year stability, age-specific degradation rates. Our hypothesis: Beta regression confidence intervals will achieve nominal coverage (95% ± 2%) while linear models will show miscalibration near boundaries.

**Phase 3: Operational Validation.** Calculate re-enrollment intervals for 80% TAR threshold. Validate age-specific recommendations against observed degradation. Test robustness across demographic subgroups (if available). Compare computational requirements in operational settings.

**Success Criteria:** Beta regression maintains <2% boundary violations versus >5% for linear models. Variance modeling error <10% of empirical variance across performance range. Out-of-sample RMSE within 15% of in-sample performance. Confidence interval coverage: 93–97% (nominal 95%).

**Falsification Conditions:** If real data shows: (1) linear models produce no boundary violations over 8 years, (2) variance is approximately constant across performance levels, or (3) transformation-based approaches provide equivalent or better predictive accuracy, then beta regression offers minimal practical advantage.

Code and detailed instructions available at: https://github.com/astoreyai/memory-augmented-transformers

# 9 Research Roadmap

The path forward begins with validation on existing longitudinal datasets. Access to the Children Longitudinal Face (CLF) dataset [919 children, 2–18 years] or Young Face Aging (YFA) dataset [330 subjects, 8-year tracking] would enable direct empirical validation of the simulation findings. While multiple attempts to obtain these datasets for this study were unsuccessful, the detailed validation protocol provided in Section 8 facilitates rapid assessment once data access is granted. We are actively seeking collaborations with institutions holding longitudinal biometric data to conduct this validation.

Real-data validation would address several questions that simulation cannot fully answer. First, do actual recognition rates exhibit the quadratic variance pattern predicted by beta regression theory? Second, how frequently do operational systems encounter the boundary violations demonstrated in our pedagogical example? Third, are the age-specific patterns (particularly the 5.5–7 year stability) robust across different algorithms, populations, and imaging conditions?

Beyond CLF and YFA, other longitudinal biometric datasets could test generalizability. The MORPH database contains age-progressed faces suitable for shorter-term validation. Operational datasets from NCMEC (National Center for Missing & Exploited Children) would provide the ultimate test, though privacy constraints require careful protocol design. Fingerprint and iris modalities offer opportunities to assess whether beta regression advantages extend beyond face recognition.

Where real data remains unavailable, simulation studies can fill critical gaps by generating data matching published patterns, varying sample sizes from 100 to 1,000 subjects, testing 2–10 observations per subject, and exploring different correlation structures. Documenting estimation challenges and solutions from these studies will guide future applications.

Accelerating adoption requires dedicated software development. R packages that wrap beta regression for biometric applications would lower the technical barrier, while Python implementations enable integration with deep learning pipelines. Comprehensive tutorials bridging statistical theory and practical implementation will help practitioners transition from traditional methods.

Looking beyond immediate applications, multimodal fusion presents compelling opportunities. Children provide face, fingerprint, and voice samples that degrade differently with age—multivariate beta regression could capture their joint evolution, allowing optimal combination weights to adapt over time. The Young Face Aging dataset established by Bahmani and Schuckers [30] provides critical benchmarks for algorithm development. Integration with synthetic data generation approaches [1] could address data scarcity while maintaining privacy. Operational deployment introduces additional complexities, as laboratory photos differ substantially from surveillance footage and image quality varies wildly. Beta regression naturally extends to include quality covariates, letting image quality affect precision parameters while maintaining unbiased mean estimates.

The integration with deep learning systems offers perhaps the most transformative potential. Current systems train on aggregate loss functions without considering age-specific reliability. Beta regression could weight training samples based on their expected variance, helping networks learn features stable within developmental windows. This statistical foundation would guide architectural choices from the ground up.

The statistical insights from beta regression analysis directly inform memory-augmented transformer architectures for age-invariant face recognition. The quadratic variance pattern—with uncertainty peaking at 50% performance—suggests weighting training samples inversely to their expected variance, improving model calibration in uncertain regions. Age-specific degradation patterns point toward adaptive memory mechanisms where the stable middle age group (5.5-7 years) justifies longer retention periods in memory banks, while rapidly changing toddlers and teenagers require more frequent updates. This isn't one-size-fits-all; it's biology-informed architecture design.

Beta regression's natural handling of boundary uncertainty could revolutionize output layer design. Instead of forcing neural networks to learn probability constraints through trial and error, these constraints build directly into the architecture. Li et al. [31] recently developed attention-based factorization for identity-age feature decomposition, providing architectural guidance for beta-regression-informed neural networks. The non-linear time effects uncovered in this analysis inform temporal attention mechanisms, helping transformers focus on the most reliable historical embeddings for each age group. These statistical foundations ensure that neural architectures align with the empirical properties of longitudinal biometric data, creating a bridge between rigorous statistical modeling and practical deep learning applications.

# 10 Limitations and Appropriate Use Cases

Understanding when to apply beta regression versus traditional methods requires careful consideration of the specific biometric evaluation context. Beta regression proves most beneficial for modeling population-level recognition rates over extended timeframes exceeding three years, particularly when variance heteroscedasticity is evident in the data. Direct interpretation of percentage-based metrics is often required for operational decision-making, and confidence intervals near boundaries frequently influence critical system design choices. When comparing multiple age groups with different degradation patterns, beta regression captures the varying uncertainty levels inherent in bounded outcomes.

## 10.1 Decision Framework for Method Selection

Table 3 provides guidance on when beta regression offers advantages over standard approaches versus when established methods suffice.

Sample size requirements deserve particular attention. Beta regression with random effects requires minimum data for stable estimation: at least 50 subjects with 5 or more repeated observations each provides adequate power for

TABLE 3
Method Selection Decision Framework

| Condition | Recommended Approach |
|---|---|
| Raw similarity scores available | ROC/d-prime analysis |
| Threshold optimization needed | ROC curve methods |
| Short-term predictions (<2 years) | Linear mixed models |
| Small sample (<50 subjects) | Simpler models |
| Aggregate TAR modeling | Beta regression |
| Long-term extrapolation (>3 years) | Beta regression |
| Bounded outcome interpretation | Beta regression |
| Variance heterogeneity evident | Beta regression |
| Individual matching decisions | Score-level methods |
| Stakeholder communication | Beta regression |

typical effect sizes. Below these thresholds, precision parameters and variance components become unreliable. Simpler approaches—pooled beta regression without random effects, or transformed linear models—offer more stable estimates with sparse data.

The number of random effects also constrains model complexity. Full variance-covariance structures for random slopes require substantial data. With fewer than 100 subjects, diagonal covariance structures (uncorrelated random effects) improve stability. Bayesian estimation with weakly informative priors can partially mitigate small-sample issues, though prior specification requires domain expertise.

Boundary violations provide a practical decision signal. If preliminary linear mixed model analysis shows predictions remaining well within (0.1, 0.9) for all extrapolation periods, added complexity of beta regression may not justify the computational cost. However, when extrapolations approach boundaries or stakeholder requirements demand rigorous probability bounds, beta regression becomes essential.

Conversely, existing approaches remain preferable in several scenarios. Working directly with similarity scores rather than binary outcomes favors ROC analysis for threshold selection and d-prime for threshold-independent assessment. Studies with limited sample sizes—fewer than 50 subjects or 5 timepoints—benefit from simpler models that provide more stable parameter estimates. Short-term predictions spanning less than two years rarely encounter boundary violations, making traditional linear models adequate. Computational efficiency requirements and individual matching applications also favor established biometric evaluation methods over the more complex beta regression.

Rather than replacing existing methods, beta regression should complement them within a comprehensive evaluation framework. This integration involves using d-prime for threshold-independent performance assessment, ROC curves for operational threshold selection, and beta regression for long-term population modeling. Such a combined approach addresses both immediate operational needs and strategic planning requirements, leveraging each method's strengths.

Like any statistical method, beta regression comes with assumptions and limitations. It assumes outcomes follow a beta distribution—reasonable for proportions but potentially restrictive for other biometric measures. The logit link function, while ensuring bounded predictions, might not capture every non-linear pattern optimally. Sometimes the data tells a story that even flexible models struggle to capture.

Data requirements pose another constraint. Reliable estimation needs at least 50 subjects with 5 or more observations each. The precision parameter and random effects covariance are particularly data-hungry, requiring substantial sample sizes for stable estimation. Small pilot studies might not provide enough information for the model to converge properly.

Computational demands can't be ignored either. Mixed-effects beta regression runs 3-5 times slower than linear mixed models due to iterative optimization. Unlike linear models with closed-form solutions, beta regression requires numerical optimization at each step. For real-time applications or massive datasets, this computational overhead might prove prohibitive. Despite these limitations, the framework aligns with established biometric evaluation standards [21] and provides practical advantages demonstrated in operational deployments [27].

### 10.2 When Not to Use Beta Regression

Several biometric evaluation contexts make beta regression inappropriate or unnecessarily complex.

**Score-level analysis:** When working directly with raw similarity scores rather than aggregated recognition rates, score-level methods prove more powerful. Beta regression models the proportion of scores exceeding a threshold, discarding distributional information. ROC analysis preserves the full score distribution, enabling threshold optimization and sensitivity analysis. The binning operation that creates TAR values from scores loses information that score-level likelihood ratio tests retain.

**Individual matching decisions:** Operational systems make person-specific match/non-match decisions based on similarity scores, not population-level recognition rates. Beta regression informs system design and policy decisions but does not directly support individual identification. For this reason, it complements rather than replaces match score analysis in operational deployment.

**Real-time threshold adaptation:** Systems requiring dynamic threshold adjustment based on image quality or environmental conditions need score-level calibration methods. Beta regression operates on aggregated outcomes over longer time scales, making it unsuitable for per-image threshold optimization.

**Multimodal score fusion:** While beta regression can model recognition rates from fused systems, score-level fusion techniques (weighted sum, support vector machines, likelihood ratios) leverage distributional properties that aggregate metrics obscure. The binning operation that creates bounded outcomes from continuous scores discards precisely the information fusion methods exploit.

**Algorithm development and debugging:** During system development, examining score distributions, false accept/reject pairs, and failure mode analysis provides diagnostic information. Beta regression's aggregate view obscures these implementation details necessary for algorithm improvement.

**Short-term evaluations:** Studies spanning less than 2 years with stable populations rarely encounter boundary violations. Standard mixed models provide adequate fit with simpler estimation and interpretation. The added complexity of beta regression offers minimal practical advantage when extrapolation remains modest and baseline performance stays away from bounds.

These limitations underscore beta regression's role as a specialized tool for specific contexts—long-term population-level modeling when bounded interpretability matters—rather than a general replacement for biometric evaluation methods.

Priority future work includes validation on real longitudinal datasets (CLF, YFA) and extension to multi-modal biometric systems.

### 10.3 Path to Operational Deployment

While this framework shows promise in simulation, several steps remain before operational deployment. Immediate priorities include validation on CLF dataset (919 children, 2–18 years), validation on YFA dataset (330 subjects, 8-year tracking), cross-validation with other longitudinal biometric studies, and robustness testing across demographic groups.

Medium-term validation should include comparison with operational NCMEC search results, retrospective analysis of resolved cases versus predicted performance, controlled A/B testing of re-enrollment strategies, and cost-benefit analysis of differentiated enrollment schedules.

Deployment considerations include integration with existing biometric evaluation pipelines, training materials for practitioners and stakeholders, ethical review of differentiated treatment by age group, and monitoring protocols for detecting model drift.

We estimate 12–18 months for complete validation on available datasets, followed by 24–36 months of operational pilot testing before full deployment recommendations can be made with confidence. The simulation results provide a methodological foundation, but patient, rigorous validation is essential given the critical nature of missing children investigations.

## 11 Conclusion

This paper examined beta regression as a statistical approach for modeling bounded biometric performance in child face recognition. Through pedagogical simulation, this study identified a methodological concern: while published studies appropriately use standardized scores and transformations, direct application of linear mixed models to bounded recognition rates (as might occur in operational settings) can produce physically impossible predictions exceeding 100%. This illustrates why choosing appropriate statistical methods matters for operational biometric systems.

The simulation study, designed to mimic published longitudinal patterns, showed that when linear models are inappropriately applied to bounded TAR values, they can produce hundreds of invalid predictions—including recognition rates exceeding 100%. This synthetic analysis underscores why beta regression matters: it eliminates such violations while maintaining comparable fit (RMSE differences under 0.002).

The advantages extend beyond mathematical correctness. All predictions remain within valid probability bounds, ensuring credible outputs for law enforcement and families who depend on these systems. The model correctly captures variance patterns, with uncertainty naturally peaking at 50% performance and shrinking near the boundaries—providing appropriate confidence intervals exactly where they matter most.

Most significantly for operational systems, the analysis identified a period of exceptional facial stability for children aged 5.5-7 years. This stable period extends 16 months longer than linear models would suggest, potentially maintaining case viability during critical extended search periods for missing children.

Beta regression is not a universal replacement for existing methods. For similarity score analysis, threshold optimization via ROC curves, or short-term predictions, established practices remain preferable. Beta regression adds specific value when modeling population-level recognition rates over extended timeframes, when variance patterns impact decisions, or when stakeholder communication requires interpretable percentage-based metrics.

This approach provides essential statistical foundations for next-generation biometric systems. The age-specific patterns and variance relationships discovered through beta regression directly inform the design of memory-augmented transformer architectures, suggesting when to update stored representations and how to weight temporal information.

By choosing appropriate statistical tools for specific evaluation contexts, researchers can build biometric systems that are both mathematically sound and operationally effective—systems worthy of the critical mission they serve in reuniting missing children with their families.

## Ethics Statement

This research uses only synthetic data generated computationally to match published degradation patterns from the literature. No human subjects were involved in data collection for this study. The synthetic datasets were calibrated to aggregate statistical patterns reported in peer-reviewed publications (Deb et al., Bahmani et al.) and do not contain any identifiable information or biometric samples from real individuals. Therefore, institutional review board approval was not required for this purely methodological work.

## Data Availability

The simulation code, synthetic datasets, and analysis scripts used in this study are available at: https://github.com/astoreyai/memory-augmented-transformers. The repository provides Python implementations of the simulation code, synthetic datasets calibrated to match YFA patterns, and Jupyter notebooks for reproducing all analyses, including beta regression model fitting using Python's statsmodels library. A hands-on tutorial notebook guides readers through applying these methods to their own longitudinal biometric data.

## Acknowledgment

The authors thank the reviewers for their constructive feedback and acknowledge the limitations of synthetic-only validation. The authors welcome collaboration with researchers who have access to longitudinal biometric datasets to conduct real-data validation of this framework.

## References

[1] M. Falkenberg, A. B. Ottsen, M. Ibsen, and C. Rathgeb, "Child face recognition at scale: Synthetic data generation and performance benchmark," *Front. Signal Process.*, vol. 4, p. 1308505, 2024.

[2] D. Deb, D. Aggarwal, and A. K. Jain, "Identifying missing children: Face age-progression via deep feature aging," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 10 540–10 547.

[3] D. Deb, N. Nain, and A. K. Jain, "Longitudinal study of child face recognition," in *Proc. IEEE Int. Conf. Biometrics (ICB)*, Gold Coast, Australia, Feb. 2018, pp. 225–232.

[4] K. Bahmani, S. Singh, and S. Schuckers, "Longitudinal evaluation of child face recognition and the impact of underlying age," in *Proc. IEEE Int. Workshop Biometrics Forensics (IWBF)*, Barcelona, Spain, Apr. 2023, pp. 1–9.

[5] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *J. Appl. Stat.*, vol. 31, no. 7, pp. 799–815, Aug. 2004.

[6] L. G. Farkas, *Anthropometry of the Head and Face*, 2nd ed. New York, NY, USA: Raven Press, 1994.

[7] L. Best-Rowden and A. K. Jain, "Longitudinal study of automatic face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 148–162, Jan. 2018.

[8] Z. Huang, J. Zhang, and H. Shan, "When age-invariant face recognition meets face age synthesis: A multi-task learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7917–7932, 2023.

[9] S. Yoon and A. K. Jain, "Longitudinal study of fingerprint recognition," *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 28, pp. 8555–8560, 2015.

[10] M. Smithson and J. Verkuilen, "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables," *Psychol. Methods*, vol. 11, no. 1, pp. 54–71, 2006.

[11] F. Cribari-Neto and A. Zeileis, "Beta regression in R," *J. Stat. Softw.*, vol. 34, no. 2, pp. 1–24, Apr. 2010.

[12] M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *R J.*, vol. 9, no. 2, pp. 378–400, Dec. 2017.

[13] P. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *J. Stat. Softw.*, vol. 80, no. 1, pp. 1–28, Aug. 2017.

[14] J. I. Figueroa-Zúñiga, R. B. Arellano-Valle, and S. L. P. Ferrari, "Mixed beta regression: A Bayesian perspective," *Comput. Stat. Data Anal.*, vol. 61, pp. 137–147, 2013.

[15] M. Hunger, A. Döring, and R. Holle, "Longitudinal beta regression models for analyzing health-related quality of life scores over time," *BMC Med. Res. Methodol.*, vol. 12, p. 144, 2012.

[16] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[17] J. C. Wu, A. F. Martin, R. N. Kacker, and C. R. Hagwood, "Measures, uncertainties, and significance test in operational ROC analysis," *J. Res. Natl. Inst. Stand. Technol.*, vol. 116, no. 2, pp. 517–537, 2011.

[18] M. D. Garris and C. L. Wilson, "NIST biometric evaluations and developments," National Institute of Standards and Technology, NIST Interagency Report 7204, 2005.

[19] L. Best-Rowden, B. Klare, J. Klontz, and A. K. Jain, "Relating ROC and CMC curves via the biometric menagerie," in *Proc. IEEE Int. Conf. Biometric Forensics Security (BIFS)*, 2013, pp. 1–8.

[20] A. Adler and M. E. Schuckers, "Calculation of a composite DET curve," in *Audio- and Video-Based Biometric Person Authentication*, vol. 3546, 2005, pp. 756–765.

[21] J. Richiardi and K. Kryszczuk, "Biometric systems evaluation," in *Encyclopedia of Cryptography and Security*. Springer, 2011.

[22] T. Schlett, C. Rathgeb, J. E. Tapia, and C. Busch, "Considerations on the evaluation of biometric quality assessment algorithms," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 6, no. 1, pp. 54–67, 2024.

[23] R. Ospina and S. L. P. Ferrari, "A general class of zero-or-one inflated beta regression models," *Comput. Stat. Data Anal.*, vol. 56, no. 6, pp. 1609–1623, 2012.

[24] O. Asar, D. Bolin, P. J. Diggle, and J. Wallin, "Linear mixed effects models for non-Gaussian continuous repeated measurement data," *J. R. Stat. Soc. C*, vol. 69, no. 5, pp. 1015–1065, 2020.

[25] F. E. H. Jr., *Regression Modeling Strategies*, 2nd ed. New York, NY, USA: Springer, 2015.

[26] J. C. Wu, "Operational measures and accuracies of ROC curve on large fingerprint data sets," National Institute of Standards and Technology, NIST Interagency Report 7495, 2008.

[27] D. Aykac *et al.*, "Long-range biometric identification in real world scenarios," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2024, pp. 1–8.

[28] J. Lee, J. Hong, Y. Kwon, and J. Choo, "Improving face recognition with large age gaps by learning to distinguish children," *arXiv preprint*, 2021.

[29] R. Raghavendra, K. B. Raja, and C. Busch, "Robust scheme for iris presentation attack detection," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 703–718, 2015.

[30] K. Bahmani and S. Schuckers, "Face recognition in children: A longitudinal study," *arXiv preprint*, 2022.

[31] J. Li, L. Zhou, and J. Chen, "Age-invariant face network (AFN): a discriminative model towards age-invariant face recognition," *Neural Comput. Appl.*, vol. 36, no. 22, pp. 13 689–13 702, 2024.

**Aaron W. Storey** (Member, IEEE) is a Ph.D. candidate in Computer Science and a Graduate Research Assistant at Clarkson University, Potsdam, NY, USA. He received the M.S. degree in Artificial Intelligence from Maryville University, St. Louis, MO, USA, in 2024.

His research focuses on the development of agentic AI systems and language model architectures for explainable, reproducible, and context-aware decision-making. He works at the intersection of transformer models, prompt-based control, and symbolic reasoning, with an emphasis on transparency, reproducibility, and alignment with emerging AI governance principles and responsible deployment practices. Application areas include biometric verification, time-series modeling, autonomous systems, and financial strategy automation.

**Masudul H. Imtiaz** received the Ph.D. degree in electrical engineering from the University of Alabama, Tuscaloosa, AL, USA, in 2019. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Clarkson University, Potsdam, NY, USA, where he directs the AI Vision, Health, Biometrics, and Applied Computing (AVHBAC) laboratory.

His research interests include wearable sensor systems, biometric recognition (particularly pediatric biometrics), machine learning for physiological signal processing, and mHealth technologies. Dr. Imtiaz has authored over 100 peer-reviewed publications in IEEE transactions, conferences, and other venues. His recent work focuses on longitudinal biometric recognition in children, deepfake detection, and advanced sensing technologies for healthcare applications.