Simulation Study Protocol

Merlin Urbanski

April 14, 2025

1 Introduction

In many scientific disciplines, especially in biomedical research, missing data is a common and challenging issue. Patient information is often gathered from multiple hospitals, countries, and healthcare providers, each with distinct documentation practices and varying levels of detail. Moreover, not every patient receives the same diagnostic tests or treatments, resulting in datasets where essential measurements may be missing. Despite these challenges, patient data remain valuable for advancing medical knowledge and guiding clinical decisions. To make the most of these incomplete datasets, it is crucial to distinguish between two analytical goals, each requiring its own approach to missing data.

Parameter estimation focuses on accurately capturing relationships between variables, while prediction aims to maximize the accuracy of predicted outcomes. Various strategies address missing data for each goal, allowing researchers to still gain meaningful insights. Traditional methods include complete-case analysis (excluding incomplete cases) and mean-value imputation (filling missing values with averages). Although these approaches have been shown to bias parameter estimates [**?**] and reduce predictive performance [**?**], they remain in use [**?**].

In recent years, imputation methods, such as regression imputation and predictive mean matching, have become widely used in scientific research [**?**]. However, substantive models, especially in biomedical research, are growing increasingly complex, often incorporating techniques like splines and tree-based models that are able to model non-linear relationships. Despite this complexity, many researchers still rely on simpler, default imputation methods that assume linearity. This mismatch can lead to a phenomenon called uncongeniality. Uncongeniality arises when the imputation model fails to account for the assumptions or structure of the substantive model, such as when the imputation assumes linearity while the substantive model incorporates non-linear effects like splines [**?**]. Such mismatches may result in poor model performance, highlighting the need to consider uncongeniality when addressing missing data.

Although Meng defined uncongeniality in 1994 [**?**], it has rarely been the subject of in-depth investigation. Related work on incompatibility, where the imputation and substantive models lack a consistent joint probability distribution, has mainly focused on parameter estimation rather than prediction [**?**, **?**]. To adress this research gap, the aim of this master thesis is to evaluate how uncongeniality affects predictive performance through a simulation study.

2 Methods

2.1 ADEMP

The study adheres to the ADEMP guidelines for the design and reporting of the simulation study [**?**].

2.2 Aim

We aim to determine the effect of congeniality between an imputation model and a substantive prediction model on predictive performance. Under 40 realistic scenarios of univariate missingness ten model combinations will be compared based on their out-of-sample predictive performance.

2.3 Data-Generating Mechanisms

2.3.1 Scenarios

Data with only continous predictors and a continous outcome will be simulated to reflect 40 (5 x 2 x 2 x 2) unique scenarios. This is achieved by varying the following four characteristics of the data: missingness-mechanism (MCAR, weak MAR, strong MAR, weak MNAR, strong MNAR), type of correlation between the one predictor with missingness and the predictors without missingness linear and quadratic, type of correlation between the predictors and the outcome variable linear and quadratic and strength of correlation between the predictors low (0.2) and high (0.8).

Table 1: Summary of factors to be varied in the data simulation. Data will be simulated to reflect 40 unique scenarios (5 x 2 x 2 x 2), by varying the following characteristics.

Factor	**Levels**
Missingness Mechanism	MCAR weak MAR, strong MAR weak MNAR, strong MNAR
Cor.-type between predictors	Linear, Quadratic
Cor.-type between predictors and outcome	Linear, Quadratic
Cor.-strength among predictors	Low (0.2), High (0.8)

All training data sets will have a sample size of 1000 observations while the test sets are 100 times larger and have a sample size of 100,000 observations. Univariate missingness in one of the predictors will occur only in the training data set and will be constant at 30% across all scenarios.

2.3.2 Data-Generating Mechanisms

For all scenarios data of 7 of the 8 predictors will be generated from a multivariate normal distribution.

For all scenarios, the data for the 7 predictors $X_2, \ldots, X_8$ are generated as follows:

$$\mathbf{X}_{2:8} \sim \mathcal{N}_7(\mathbf{0},\, \mathbf{\Sigma}),$$

where the mean vector is given by

$$\mathbf{0} = (0, 0, \ldots, 0)^\top,$$

and the covariance matrix $\mathbf{\Sigma}$ is defined as

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

with $\rho \in \{0.2, 0.8\}$ depending on the simulation scenario.

2.3.3 Generation of X_1 with a Linear Relationship between Predictors

For the linear relationship scenarios, the predictor X_1 is constructed as a weighted sum of the other seven predictors X_2 to X_8 plus an additive noise term. The weights and noise are calibrated to ensure that, in the linear case, X_1 has the same correlation strength (either 0.2 or 0.8) with $X_2, \ldots, X_8$ as the correlations among these predictors.

1. **Linear Combination:**
   X_1 is defined as a scaled sum of the seven predictors:

   $$X_1 = a \cdot \sum_{j=2}^{8} X_j + \varepsilon,$$

   where a is a scaling factor and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$ is the error term.

2. **Low Covariance Case ($cov = 0.2$):**
   In the low covariance scenario, the scaling factor is set to

   $$a = \frac{1}{11},$$

   and the error standard deviation is

   $$\sigma_e = \sqrt{0.6706},$$

   ensuring that the introduction of noise into X_1 results in X_1 correlating with $X_2, \ldots, X_8$ at approximately 0.2, thereby matching the inter-predictor correlations.

3. **High Covariance Case ($cov = 0.8$):**
   In the high covariance scenario, the scaling factor is adjusted to

$$a = \frac{0.8}{5.8},$$

   and the error standard deviation becomes

$$\sigma_e = \sqrt{0.228},$$

   ensuring that X_1 exhibits a correlation of approximately 0.8 with $X_2, \ldots, X_8$, in line with the stronger inter-predictor correlations.

By carefully choosing the scaling factor a and the noise level σ_e, the simulation ensures that the correlation between X_1 and the predictors X_2 to X_8 matches the inter-predictor correlations. As a result, in the high correlation scenario, $\text{Var}(X_1)$ is approximately 1, while in the low correlation scenario it is around 0.8.

2.3.4 Generation of X_1 with a Quadratic Relationship between Predictors

But why are we generating X_1 in such a complicated way when we could have sampled all eight predictors from a multivariate normal distribution and avoiding the issue of differing variances for X_1? The reason is that we want to create a comparable scenario in which the association between the predictors $X_2, \ldots, X_8$ and X_1 is of similar strength as in the linear case, but the relationship is quadratic rather than linear. However, the options for directly comparing a linear to a non-linear relationship are limited. Therefore, we decided to simulate X_1 as a quadratic function of $X_2, \ldots, X_8$ with added noise.

1. **Quadratic Combination:**
   In order to create a quadratic relationship analogous to the linear case, we simply replace each predictor X_j with its square. Thus, X_1 is defined as:

$$X_1 = a \cdot \sum_{j=2}^{8} X_j^2 + \varepsilon,$$

   where a is a scaling factor and $\varepsilon \sim \mathcal{N}(0, \sigma_e^2)$ is the error term.

2. **Low Covariance Case ($cov = 0.2$):**
   For the low covariance scenario, we adopt the same structure as in the linear case, but now applied to the squared predictors. We take the exact same numbers from the linear sceanrio and therefore set

$$a = \frac{1}{11} \quad \text{and} \quad \sigma_e = \sqrt{0.6706},$$

   ensuring that the resulting quadratic effect mirrors the low inter-predictor correlation.

3. **High Covariance Case ($cov = 0.8$):**
   Similarly, for the high covariance scenario, we take the previously calculated values from the linear scenario and set

$$a = \frac{0.8}{5.8} \quad \text{and} \quad \sigma_e = \sqrt{0.228},$$

   .

In essence, we extend the linear model by squaring the predictors to induce a quadratic effect while retaining the overall structure. This approach allows us to compare model performance under both linear and quadratic relationships on a similar scale. However, metrics that evaluate the strength of these relationships—such as R^2 in regression models and the mutual information criterion (MIC)—yield different results when comparing the relationships in this study. Specifically, R^2 suggests that the quadratic (squared) relationship is stronger than the linear one, whereas the MIC indicates that the linear relationship is stronger. Therefore, we believe that our approach allows for a fair comparison. Similarly to the linear scenario, the variance of X_1 is higher in the high-correlation scenario ($\text{Var}(X_1) \approx 1.5$) compared to the low-correlation scenario ($\text{Var}(X_1) \approx 0.8$).

2.4 Generation of the Outcome Variable Y

When generating the outcome variable Y, we adopt a standard approach in which only a subset of the predictors directly affects Y. Out of the 8 predictors, the first 2 are assumed to have a strong effect, the next 2 a weak effect, and the remaining 4 no direct effect on the outcome. Moreover, the effect on Y can be either linear or quadratic.

In our simulation, the outcome is generated as follows:

1. **Linear Effects:**
   When the relationship is linear, the outcome is constructed as a linear combination of the first four predictors with weights of 1.5 for the strong-effect predictors and 0.5 for the weak-effect predictors, plus normally distributed noise:
   $$Y = 1.5\,X_1 + 1.5\,X_2 + 0.5\,X_3 + 0.5\,X_4 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1).$$

2. **Quadratic Effects:**
   When a quadratic relationship is assumed, we extend the linear model by replacing the predictors with their squares. Thus, the outcome is given by:
   $$Y = 1.5\,X_1^2 + 1.5\,X_2^2 + 0.5\,X_3^2 + 0.5\,X_4^2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1).$$

This design allows for a direct comparison of model performance when the true relationship between the predictors and the outcome is linear versus quadratic. The use of different weights for the strong and weak predictors, along with the additive noise, ensures that the generated outcome reflects a realistic and variable response.The generated scenarios are visualized in a simplified manner in Figure 1.

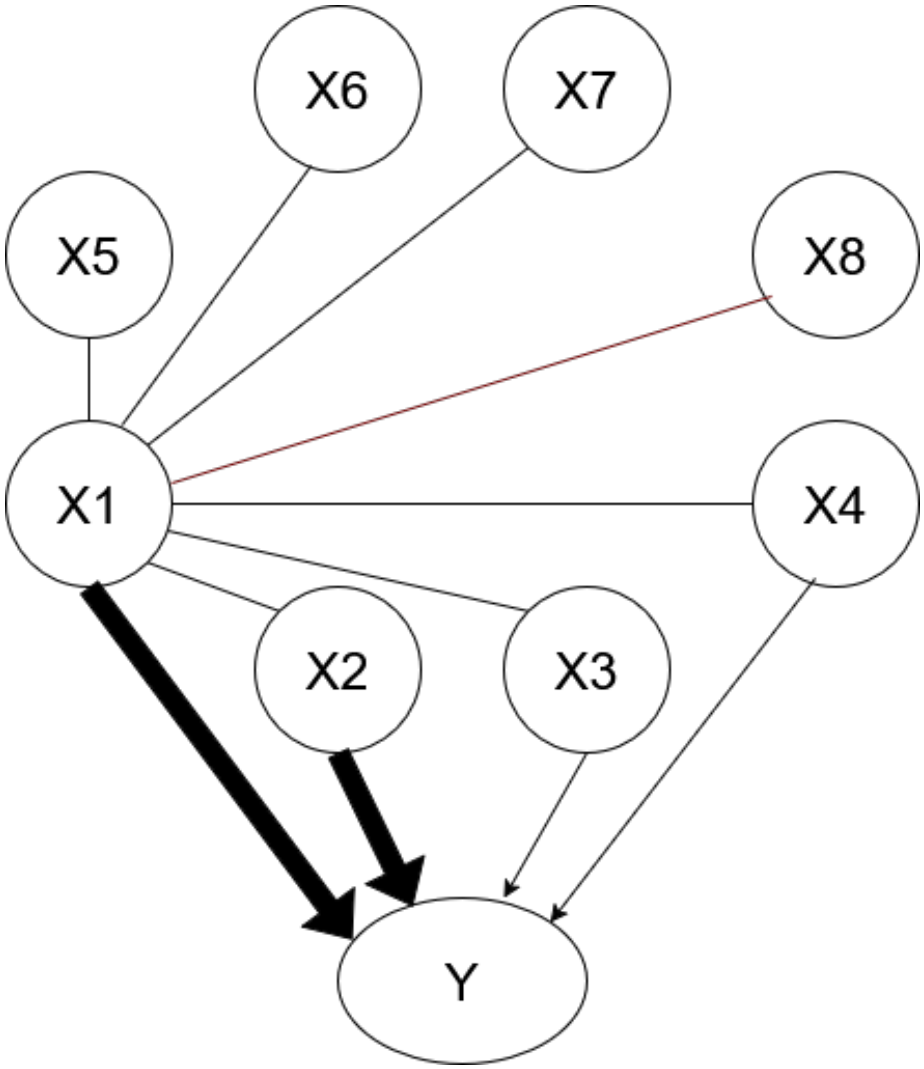

Figure 1: Visualization of the generated data. The lines among X1 and the other predictors are the relationships that can be linear or quadratic. The thick-dark arrows are the strong effects on the outcome variable Y and the thin-light arrows are the weak effects on Y.

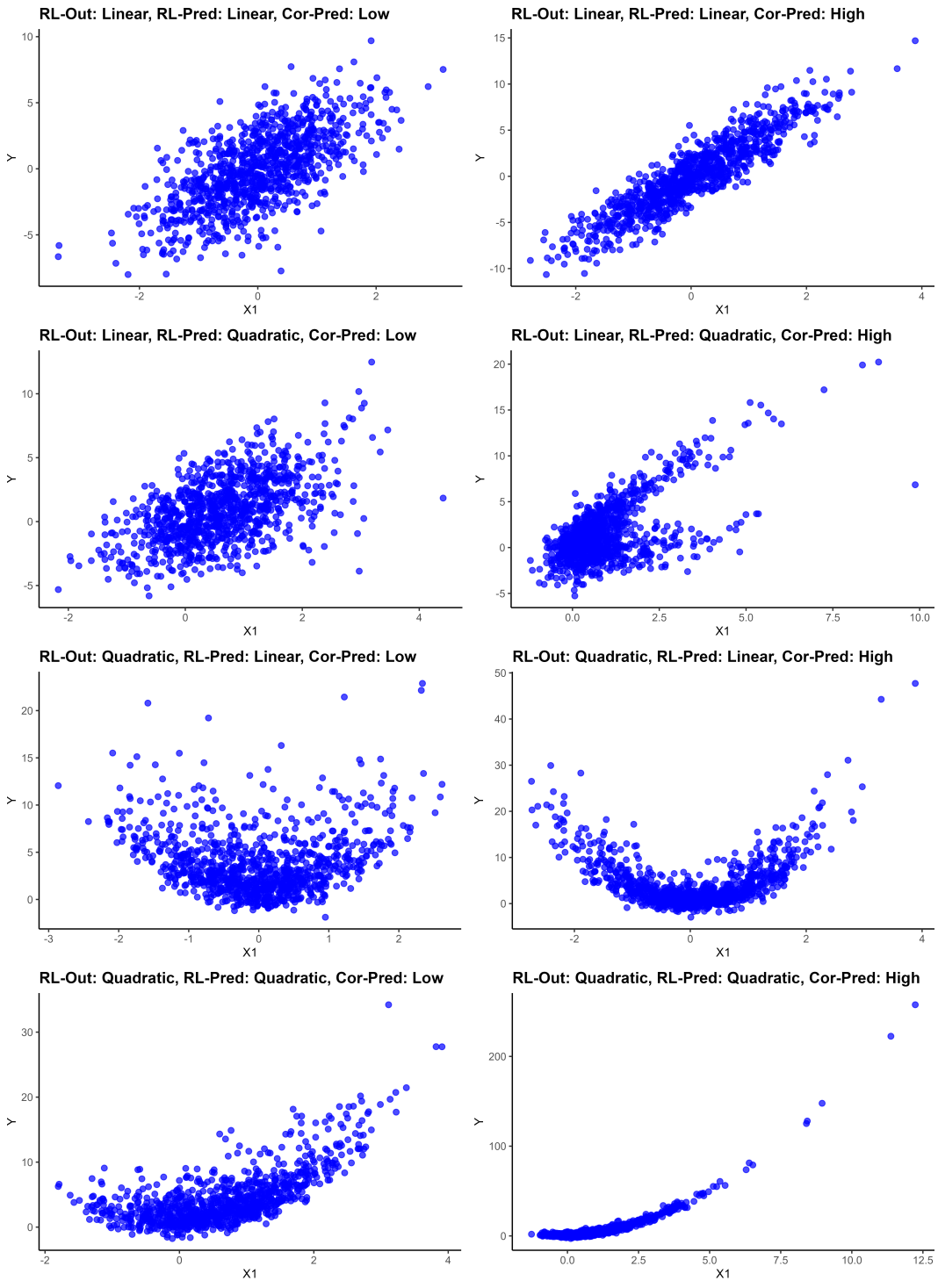

Figure 2: Visualization of the relationship between X1 and the outcome Y in 8 distinct scenarios. In the scenarios things varied: the type relationship between the predictors and outcome (RL-OUT), the type of relationship among predictors (RL-Pred) and strength of the relationship among predictors (Cor-Pred).

2.5 Missing Data Mechanisms

Missing data is imposed on the variable X_1 using five distinct mechanisms. For each mechanism, an indicator M is generated—with $M = 1$ denoting a missing value—by drawing from a Bernoulli distribution with a probability determined by the mechanism in question.

1. **MCAR (Missing Completely At Random):**

   Under the MCAR mechanism, every observation has the same probability of being missing, regardless of any other variable. Specifically, the probability of missingness is fixed at:

   $$P(M = 1) = 0.3.$$

2. **Weak MAR (Missing At Random):**

   In the weak MAR scenario, the probability that an observation is missing depends partly on the covariate X_2 and partly on random variation. First, the rank of X_2 is calculated. Then, the missingness probability is defined as a weighted combination of the rank-based component and a random component. Mathematically, this is expressed as:

   $$P(M = 1) = w \cdot \frac{\text{rank}(X_2)}{n} + (1 - w) \cdot U(0, 1),$$

   where:

   - w is a weight (for example, 0.5) that controls the relative influence of the rank-based term,
   - n is the total number of observations,
   - $U(0, 1)$ denotes a random value drawn from a uniform distribution on the interval $[0, 1]$.

3. **Strong MAR:**

   For the strong MAR mechanism, missingness is driven solely by X_2. In this case, the probability that an observation is missing is given by:

   $$P(M = 1) = \frac{\text{rank}(X_2)}{n},$$

   after which it is scaled appropriately so that the average probability matches the target missing percentage.

4. **Weak MNAR (Missing Not At Random):**

   In the weak MNAR scenario, the mechanism is similar to weak MAR but the dependency is on the variable X_1 itself. The missingness probability is computed as:

$$P(M = 1) = w \cdot \frac{\text{rank}(X_1)}{n} + (1 - w) \cdot U(0, 1),$$

   where the same definitions apply: w moderates the contribution of the rank of X_1 relative to the random component.

5. **Strong MNAR:**

   Under the strong MNAR mechanism, the missingness is determined exclusively by X_1. The probability is defined by:

$$P(M = 1) = \frac{\text{rank}(X_1)}{n}$$

   .

In each case, once the probability is computed, it is adjusted so that its average equals the intended missing percentage. Finally, for every observation, a Bernoulli draw with the computed probability is performed to decide whether the value is missing ($M = 1$) or observed ($M = 0$).

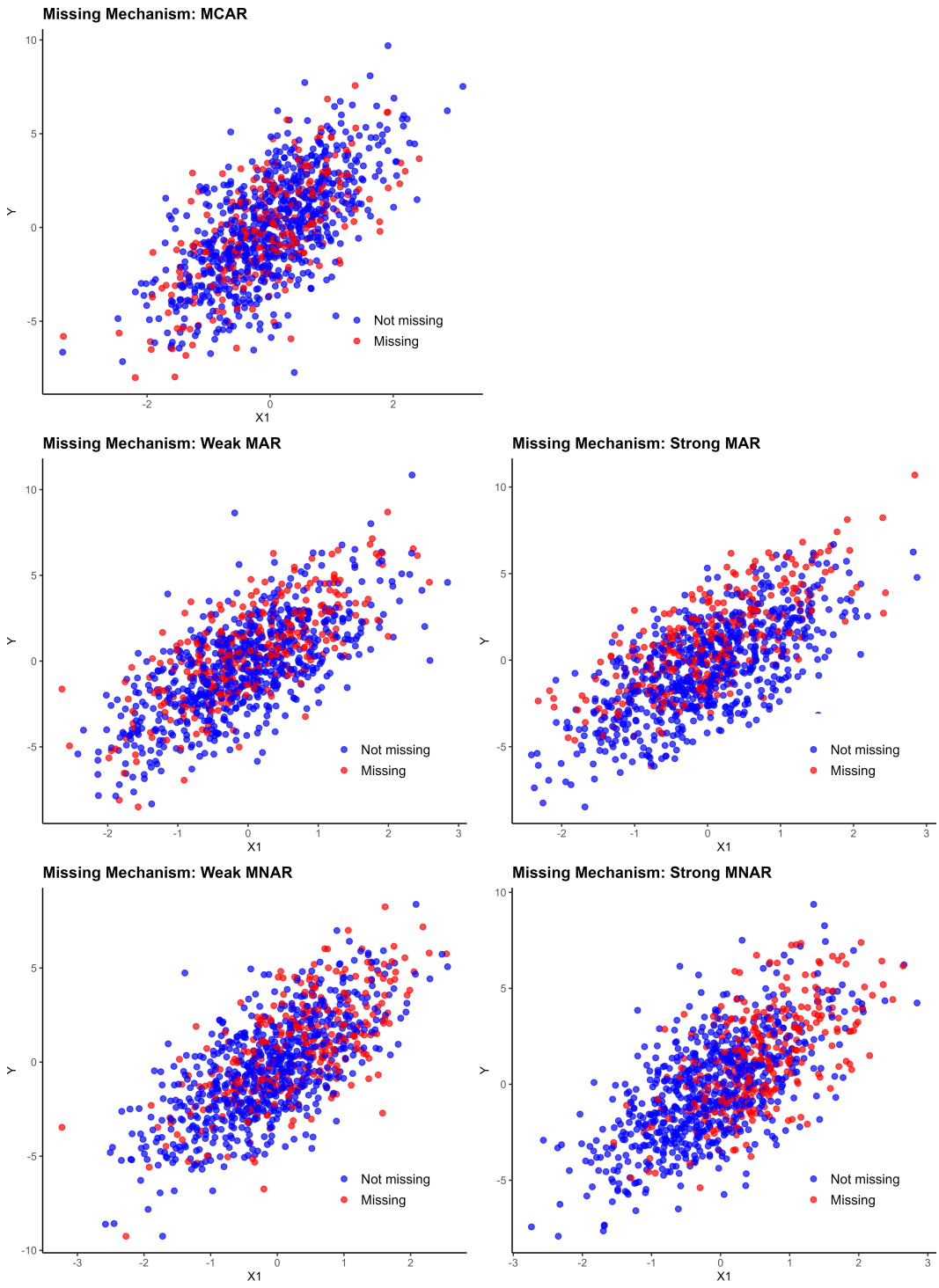

Figure 3: Visualization of the 5 Missing Mechanisms.

3 Estimands

The estimands in this study are metrics of predictive performance that evaluate the relationship between the predicted values ($\hat{Y}$) and the true outcome values (Y). These include:

1. **Root Mean Squared Error (RMSE):** RMSE evaluates the average size of error between the predicted and true values, providing a measure of overall prediction accuracy. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}.$$

   Lower RMSE values indicate more accurate predictions.

2. **Coefficient of Determination (R^2):** R^2 measures the proportion of variance in the true outcomes (Y) that is explained by the predicted values ($\hat{Y}$). It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

   R^2 ranges from 0 to 1, where $R^2 = 0$ indicates that the model explains none of the variability in Y, and $R^2 = 1$ indicates it explains all the variability perfectly.

3. **Visual Evaluation of Calibration Curves:** Calibration is assessed through visual inspection of calibration plots, which visualize the relationship between predicted and observed values. Deviations from the ideal 45 degree 1:1 line indicate systematic over- or underestimation, providing insights into the alignment of predictions with true outcomes.

These estimands were selected to assess the performance of congenial and uncongenial model combinations, evaluating different dimensions of predictive accuracy, explanatory power, and calibration.

4 Methods

4.1 Imputation Methods

To address the univariate missing data, we employed single imputation using chained equations via the R package `mice` [**?**]. Given our focus on predictive accuracy rather

than inference, we opted against multiple imputations. For all methods, the number of iterations per imputation was set to the default value of 5.

We evaluated five imputation strategies to generate single imputed datasets for model training: (1) predictive mean matching (PMM) with one donor (mice::pmm), (2) PMM with one donor extended to include quadratic terms (mice::quadratic), (3) regression imputation (mice::norm.predict), (4) regression imputation with manually specified squared terms, and (5) random forest imputation (mice::rf). For the PMM approaches, the number of donors was set to one. Additionally, we used regression imputation instead of the commonly used Bayesian regression imputation, as our primary objective was to achieve the most accurate imputed values for prediction, rather than to capture imputation uncertainty for inferential purposes.

Table 2: Imputation Strategies Employed

Full Name	Abbreviation	R/`mice` Command
Predictive Mean Matching	PMM	`mice::pmm`
PMM with Quadratic Terms	PMM-Q	`mice::quadratic`
Regression Imputation	RI	`mice::norm.predict`
Regression Imputation with Quadratic Terms	RI-Q	`mice::norm.predict` (with quadratic terms)
Random Forest Imputation	RF	`mice::rf`

4.2 Prediction Models

In our study, we employed two distinct predictive modeling approaches, both of which were trained on the imputed training dataset and subsequently applied to the test set. The first approach was a regression model that incorporated both linear and quadratic terms for all predictors, enabling us to capture potential non-linear effects while maintaining model interpretability.

The second approach involved a random forest algorithm. Recognizing the tendency of tree-based methods to overfit, we opted to use 5-fold cross-validation for model evaluation. To fine-tune the random forest, we conducted a comprehensive grid search across several hyperparameters, specifically varying the number of trees (500, 1000, and 1500), the nodesize (2, 5, and 8), and the number of predictors

sampled at each split (3, 5, and 7). The optimal model, determined by the cross-validation results, was then applied to the test dataset.

Table 3: Prediction Models Employed

Model Type	Abbreviation	R Package
Regression Model (Linear & Quadratic)	Reg	base R (`lm`)
Random Forest	RF	`randomForest`

4.3 Model Combinations

The five imputation models and the two substantive prediction models result in ten (5×2) distinct model combinations (see Table 4).

We considered six model combinations uncongenial. Five of these combinations differ in their assumptions regarding the complexity of relationships and occur when a random forest model—which is capable of modeling complex relationships—is paired with either regression-based or PMM, methods that assume simpler relationships. The final uncongenial model combination is classified as such because the regression imputation model assumes only linear effects, whereas the regression prediction model can also capture quadratic effects.

Two model combinations were labeled "Rather Uncongenial" and "Rather Congenial." Both of these include a PMM-based imputation method and a regression prediction model. We chose not to classify them strictly as uncongenial or congenial because PMM is often described as a semi-parametric method, viewed as a hybrid of parametric regression and the non-parametric k-nearest neighbor approach [**?**]. This results in PMM having less rigid assumptions than those of fully parametric regression, so the differences in assumptions are less pronounced.

Finally, there are two congenial model combinations in which the imputation model is essentially identical to the prediction model, sharing the same underlying assumptions.

Table 4: Model Combination Congeniality (using abbreviations)

Imputation Model	**Prediction Model**	**Congeniality**
PMM	Reg	Rather Uncongenial (RU)
PMM-Q	Reg	Rather Congenial (RC)
RI	Reg	Uncongenial (U)
RI-Q	Reg	Congenial (C)
RF	Reg	Uncongenial (U)
PMM	RF	Uncongenial (U)
PMM-Q	RF	Uncongenial (U)
RI	RF	Uncongenial (U)
RI-Q	RF	Uncongenial (U)
RF	RF	Congenial (C)

5 Results

5.1 RMSE and R^2 across all scenarios

Analysis of the mean RMSE across 100 iterations revealed that model combinations incorporating a regression prediction model performed consistently better than those employing a random forest prediction model. Specifically, even the worst performing regression-based combination – corresponding to the regression imputation model without quadratic effects (RI: RMSE = 1.247, SD = 0.679) – outperformed the best performing combination using a random forest prediction model – corresponding to the regression imputation model with quadratic effects (RI-Q: RMSE = 1.552, SD = 0.989).

Among the combinations that used a regression prediction model, the "rather congenial" and "rather uncongenial" approaches employing PMM-based imputation methods achieved the best performance (PMM-Q: RMSE = 1.078, SD = 0.187; PMM: RMSE = 1.114, SD = 0.261). These were followed by the congenial combination based on a regression imputation model that included quadratic terms (RI-Q: RMSE = 1.119, SD = 0.327). In contrast, the uncongenial combinations using a random forest imputation model (RF: RMSE = 1.220, SD = 0.519) and the regression imputation model without quadratic effects (RI: RMSE = 1.247, SD = 0.679) performed the worst in the regression prediction group.

For combinations using a random forest prediction model, the pattern differed. Specifically, uncongenial regression-based imputation methods (RI-Q: RMSE = 1.552, SD = 0.989; RI: RMSE = 1.568, SD = 0.982) outperformed the uncongenial PMM-based combinations (PMM: RMSE = 1.575, SD = 0.979; PMM-Q: RMSE = 1.590, SD = 1.042). Notably, the only congenial combination – where both imputation and prediction were based on random forests (RF) – yielded the poorest performance (RMSE = 1.637, SD = 1.094).

Table 5: Average RMSE and mean R^2 by Model Combination across all scenarios. Based on the RMSE the three best-performing combinations are highlighted from darkest (best) to lightest; the two combinations with the highest RMSE are unhighlighted.

Imputation Model	Prediction Model	RMSE	SD	R^2	SD	**Congeniality**
PMM	Reg	1.114	0.261	0.909	0.047	RU
PMM-Q	Reg	1.078	0.187	0.912	0.047	RC
RI	Reg	1.247	0.679	0.907	0.043	U
RI-Q	Reg	1.119	0.327	0.910	0.047	C
RF	Reg	1.220	0.519	0.906	0.045	U
PMM	RF	1.575	0.979	0.861	0.061	U
PMM-Q	RF	1.590	1.042	0.861	0.059	U
RI	RF	1.568	0.982	0.863	0.059	U
RI-Q	RF	1.552	0.989	0.865	0.059	U
RF	RF	1.637	1.094	0.860	0.058	C

5.2 RMSE and R^2 within different scenarios

Model performance varied substantially across the simulation scenarios. In nearly every case, the missing data mechanism had little influence on prediction accuracy or R^2. Accordingly, we concentrated our analysis on scenarios that differed in the type of the relationship between predictors and the outcome, the type of relationships among the predictors, and the strength of these relationships. Furthermore, in the following section we will focus mainly on the RMSE and not the R^2 because there is a very strong relationship between them.

5.2.1 Linear Relationships among Predictors and between Predictors and Outcome

When considering model combinations with a regression (REG) prediction model, those using PMM-based imputation models (PMM and PMM-Q) performed best across both high- and low-correlation scenarios. Overall, all combinations using REG prediction models showed similar performance across both correlation scenarios, although there was a slight improvement when using RF as the imputation model.

When examining combinations using a random forest (RF) prediction model, those using PMM as the imputation method performed best, followed by combinations based on RI (RI and RI-Q) in both high- and low-correlation scenarios. Interestingly, model combinations using PMM-Q consistently performed worse than other combinations, despite PMM performing very well. Additionally, for combinations using RF as the prediction model, the RMSE decreased in the high-correlation scenario compared to combinations using REG as the prediction model, where RMSE remained stable.

Across all evaluated combinations, congenial model combinations did not perform particularly well. They were consistently ranked third to fifth out of five combinations.

Table 6: Mean RMSE and mean R^2 of model combinations under scenarios with linear relationships among predictors and linear relationships between predictors and outcome. Based on the RMSE the three best-performing combinations are highlighted from darkest (best) to lightest; the two combinations with the highest RMSE are unhighlighted.

		Low Correlation				High Correlation				
Imputation Model	**Prediction Model**	RMSE	SD	R^2	SD	RMSE	SD	R^2	SD	**Congeniality**
PMM	REG	1.013	0.024	0.868	0.008	1.011	0.024	0.931	0.004	Rather Uncongenial
PMM-Q	REG	1.013	0.024	0.868	0.008	1.010	0.024	0.931	0.004	Rather Congenial
RI	REG	1.029	0.026	0.865	0.008	1.031	0.026	0.929	0.004	Uncongenial
RI-Q	REG	1.031	0.027	0.865	0.008	1.033	0.026	0.928	0.004	Congenial
RF	REG	1.033	0.027	0.864	0.008	1.019	0.024	0.930	0.004	Uncongenial
PMM	RF	1.122	0.029	0.840	0.010	1.076	0.026	0.922	0.005	Uncongenial
PMM-Q	RF	1.141	0.035	0.834	0.011	1.095	0.034	0.919	0.006	Uncongenial
RI	RF	1.125	0.030	0.837	0.010	1.084	0.027	0.921	0.005	Uncongenial
RI-Q	RF	1.125	0.030	0.837	0.010	1.084	0.027	0.921	0.005	Uncongenial
RF	RF	1.159	0.032	0.833	0.010	1.087	0.027	0.921	0.005	Congenial

5.2.2 Quadratic Relationships among Predictors and Linear Relationships between Predictors and Outcome

Table 7: Mean RMSE and mean R^2 of model combinations under scenarios with quadratic relationships among predictors and linear relationships between predictors and outcome. Based on the RMSE the three best-performing combinations are highlighted from darkest (best) to lightest; the two combinations with the highest RMSE are unhighlighted.

		Low Correlation				High Correlation				
Imputation Model	**Prediction Model**	RMSE	SD	R^2	SD	RMSE	SD	R^2	SD	**Congeniality**
PMM	REG	1.016	0.025	0.836	0.010	1.031	0.026	0.893	0.010	Rather Uncongenial
PMM-Q	REG	1.018	0.026	0.835	0.011	1.041	0.034	0.891	0.011	Rather Congenial
RI	REG	1.027	0.027	0.835	0.011	1.031	0.030	0.895	0.010	Uncongenial
RI-Q	REG	1.030	0.027	0.834	0.011	1.030	0.025	0.893	0.010	Congenial
RF	REG	1.040	0.029	0.830	0.011	1.055	0.034	0.888	0.011	Uncongenial
PMM	RF	1.130	0.030	0.800	0.013	1.117	0.039	0.875	0.011	Uncongenial
PMM-Q	RF	1.155	0.039	0.791	0.016	1.144	0.048	0.869	0.013	Uncongenial
RI	RF	1.124	0.029	0.800	0.013	1.105	0.040	0.877	0.011	Uncongenial
RI-Q	RF	1.125	0.029	0.800	0.013	1.112	0.038	0.876	0.011	Uncongenial
RF	RF	1.174	0.033	0.789	0.014	1.142	0.040	0.870	0.011	Congenial

5.2.3 Linear Relationships among Predictors and Quadratic Relationships between Predictors and Outcome

Table 8: Mean RMSE and mean R^2 of model combinations under scenarios with linear relationships among predictors and quadratic relationships between predictors and outcome. Based on the RMSE the three best-performing combinations are highlighted from darkest (best) to lightest; the two combinations with the highest RMSE are unhighlighted.

		Low Correlation				High Correlation				
Imputation Model	**Prediction Model**	RMSE	SD	R^2	SD	RMSE	SD	R^2	SD	**Congeniality**
PMM	REG	1.138	0.051	0.876	0.014	1.115	0.044	0.950	0.006	Rather Uncongenial
PMM-Q	REG	1.073	0.038	0.894	0.011	1.033	0.027	0.958	0.005	Rather Congenial
RI	REG	1.073	0.032	0.894	0.011	1.030	0.030	0.958	0.005	Uncongenial
RI-Q	REG	1.076	0.037	0.893	0.011	1.052	0.049	0.956	0.006	Congenial
RF	REG	1.078	0.039	0.890	0.011	1.072	0.037	0.954	0.006	Uncongenial
PMM	RF	1.579	0.099	0.769	0.023	1.519	0.164	0.915	0.013	Uncongenial
PMM-Q	RF	1.487	0.097	0.793	0.018	1.460	0.175	0.922	0.014	Uncongenial
RI	RF	1.513	0.096	0.787	0.020	1.435	0.166	0.923	0.013	Uncongenial
RI-Q	RF	1.511	0.096	0.787	0.020	1.458	0.169	0.921	0.014	Uncongenial
RF	RF	1.508	0.098	0.789	0.019	1.477	0.165	0.920	0.013	Congenial

5.2.4 Quadratic Relationships among Predictors and between Predictors and Outcome

Table 9: Mean RMSE and mean R^2 of model combinations under scenarios with quadratic relationships among predictors and quadratic relationships between predictors and outcome. Based on the RMSE the three best-performing combinations are highlighted from darkest (best) to lightest; the two combinations with the highest RMSE are unhighlighted.

		Low Correlation				High Correlation				
Imputation Model	**Prediction Model**	RMSE	SD	R^2	SD	RMSE	SD	R^2	SD	**Congeniality**
PMM	REG	1.108	0.054	0.930	0.009	1.479	0.603	0.987	0.014	Rather Uncongenial
PMM-Q	REG	1.103	0.052	0.930	0.009	1.332	0.439	0.988	0.011	Rather Congenial
RI	REG	1.167	0.090	0.924	0.013	2.585	1.270	0.959	0.032	Uncongenial
RI-Q	REG	1.127	0.115	0.926	0.018	1.575	0.768	0.982	0.022	Congenial
RF	REG	1.130	0.073	0.926	0.012	2.330	0.854	0.964	0.025	Uncongenial
PMM	RF	1.680	0.139	0.837	0.019	3.381	1.865	0.930	0.061	Uncongenial
PMM-Q	RF	1.678	0.140	0.838	0.019	3.557	1.970	0.924	0.066	Uncongenial
RI	RF	1.727	0.139	0.827	0.020	3.434	1.814	0.930	0.058	Uncongenial
RI-Q	RF	1.618	0.136	0.847	0.018	3.384	1.901	0.934	0.060	Uncongenial
RF	RF	1.693	0.162	0.839	0.023	3.857	1.882	0.919	0.060	Congenial

5.3 Calibration Plots

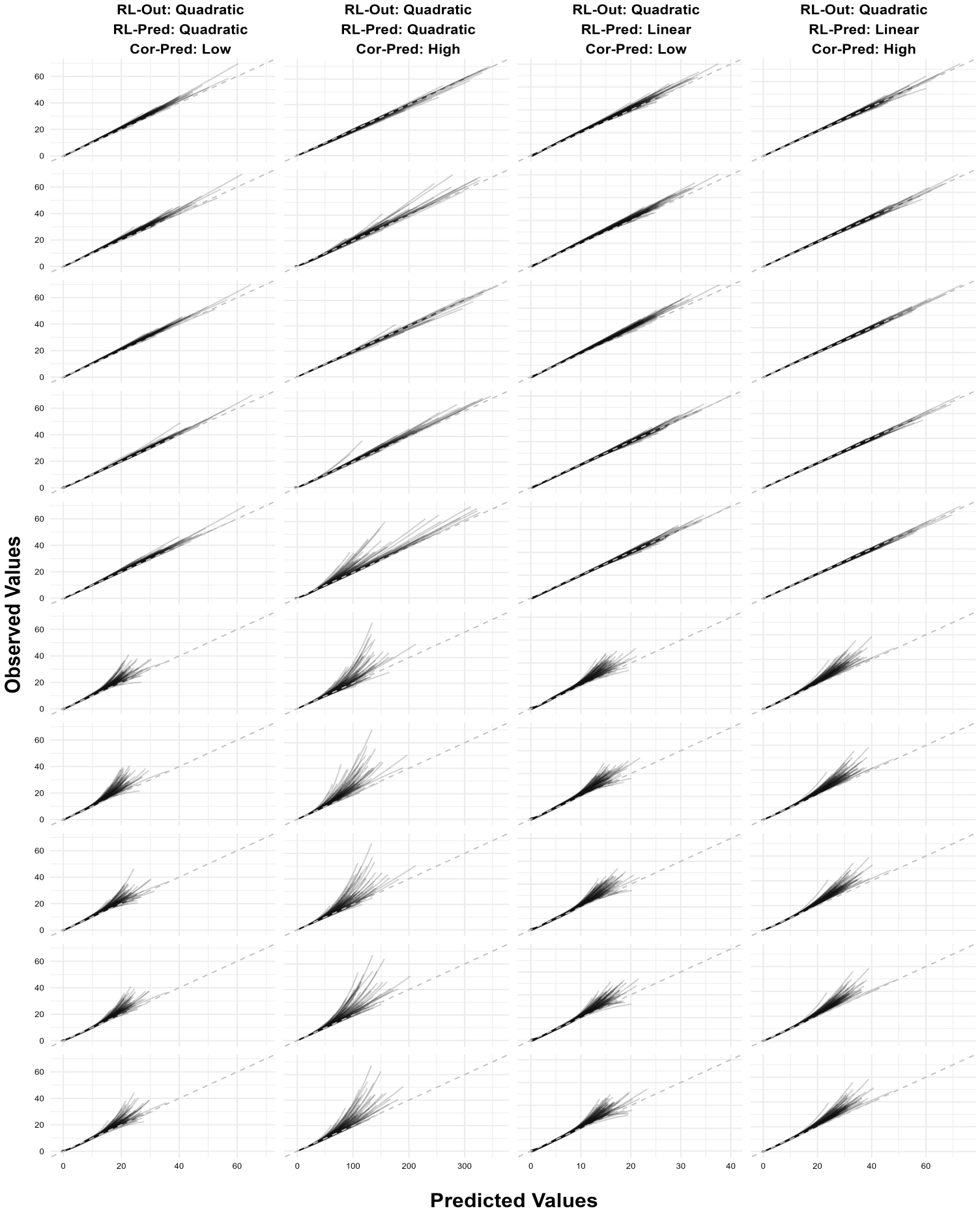

Figure 4: Calibration Plots of different scenarios with a quadratic outcome-relationship (all MCAR).

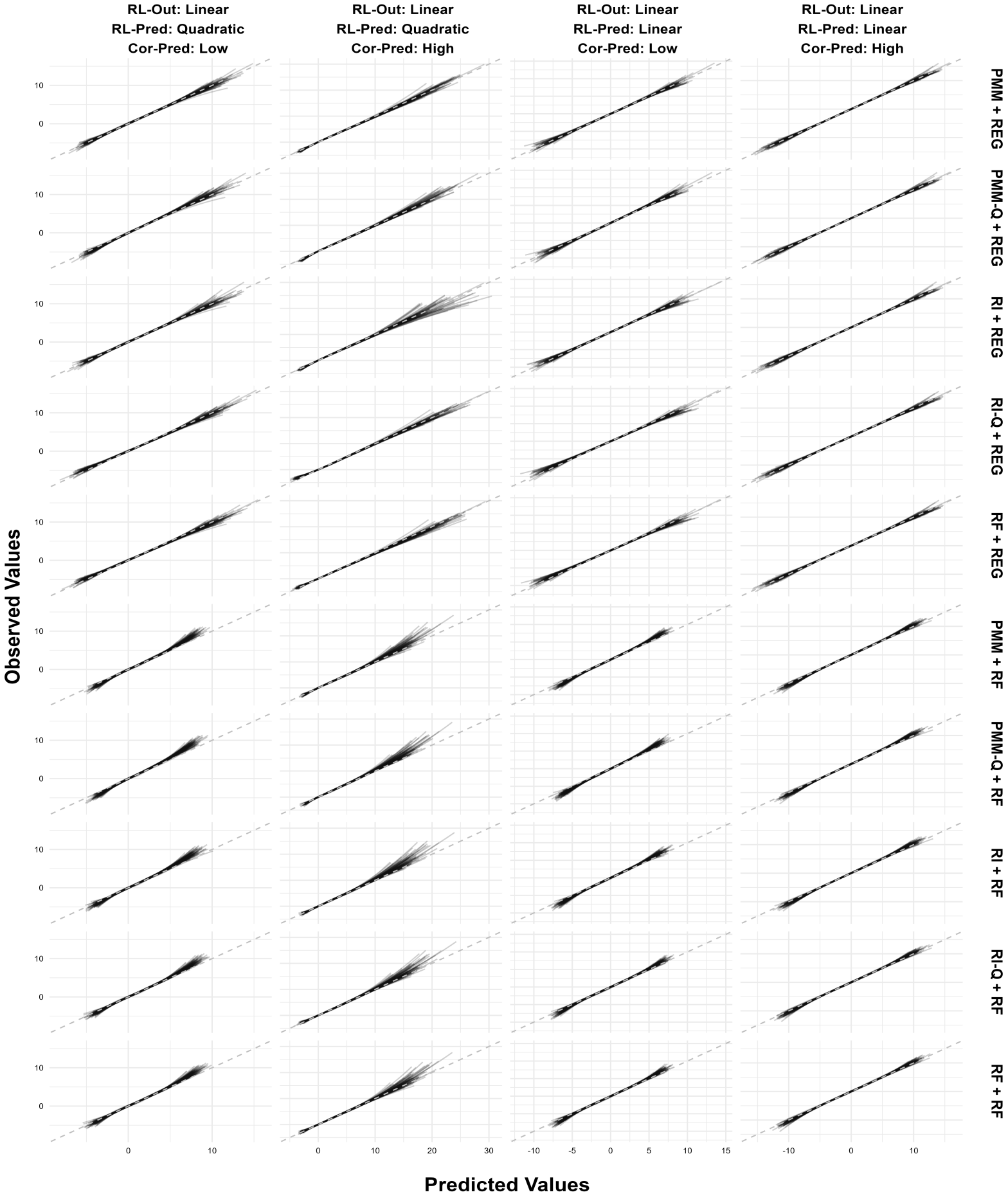

Figure 5: Calibration Plots of different scenarios with a linear outcome-relationship (all MCAR).

6 Key Findings

-linear linear rf (as prediction and imputation model) performs better with high correlation

-rf as prediction model always performes worse than lm

7 Appendix

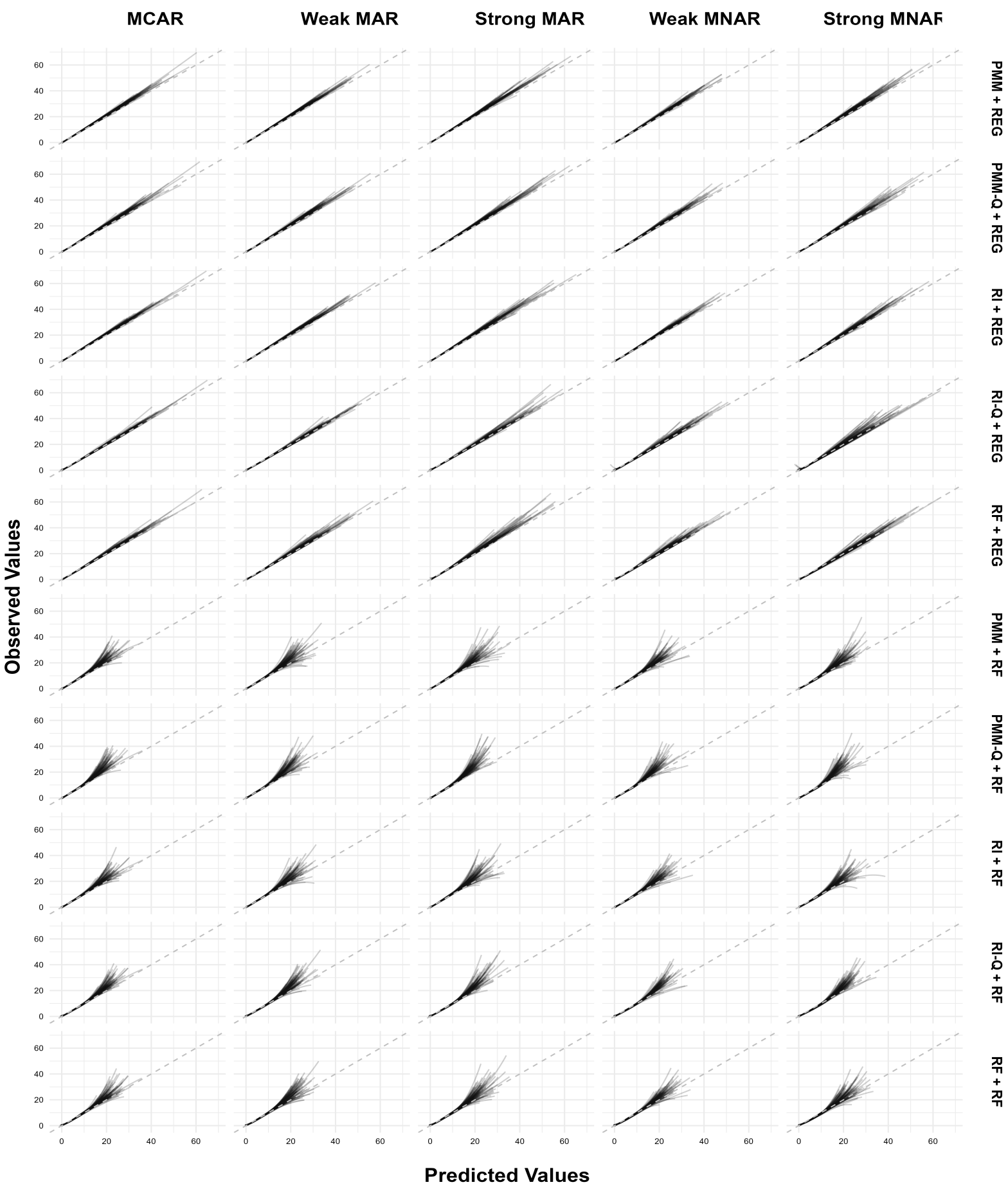

Figure 6: Calibration Plots of RL-Out: Quadratic, RL-Pred: Quadratic and Cor-Pred: Low.

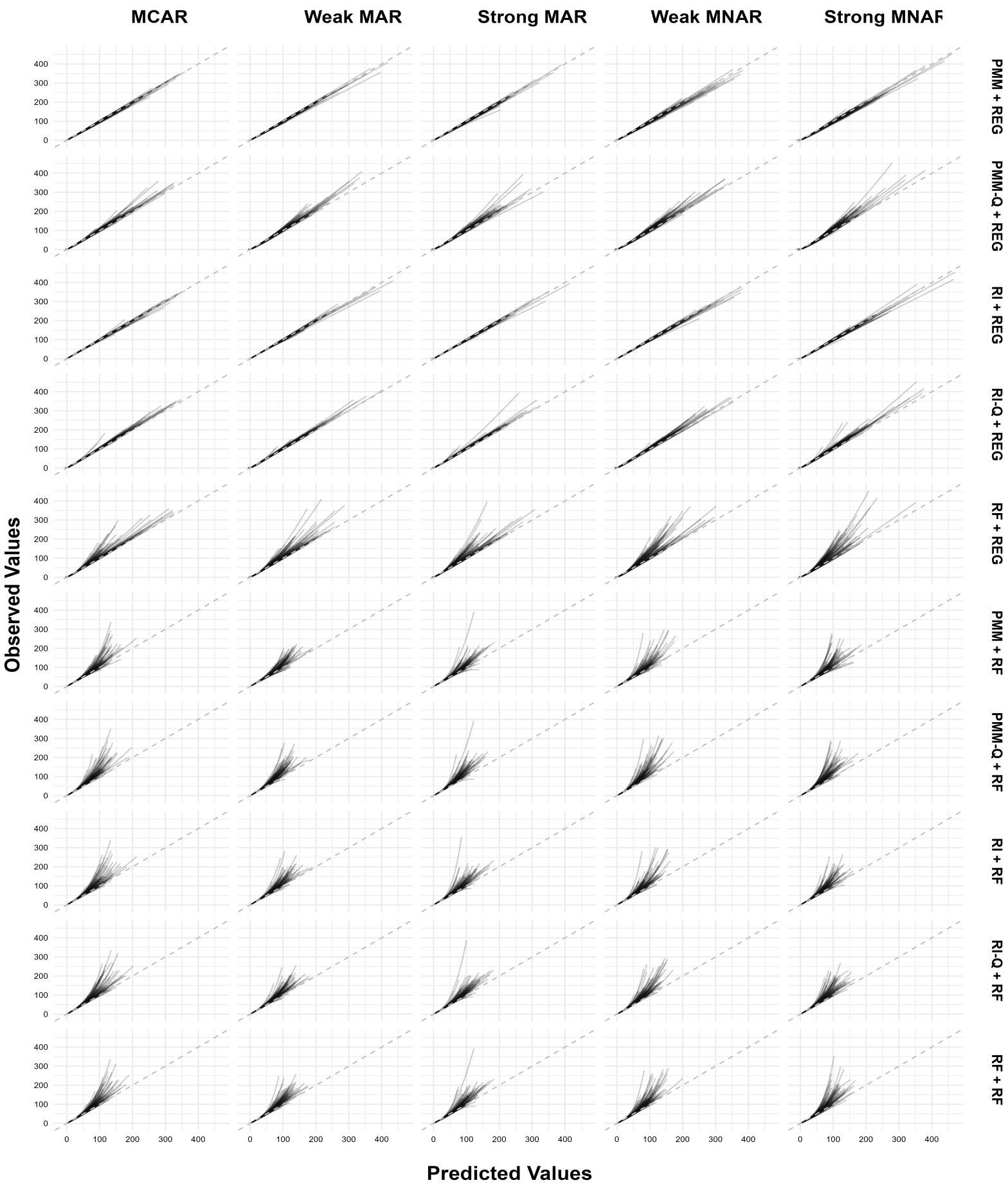

Figure 7: Calibration Plots of RL-Out: Quadratic, RL-Pred: Quadratic and Cor-Pred: High.

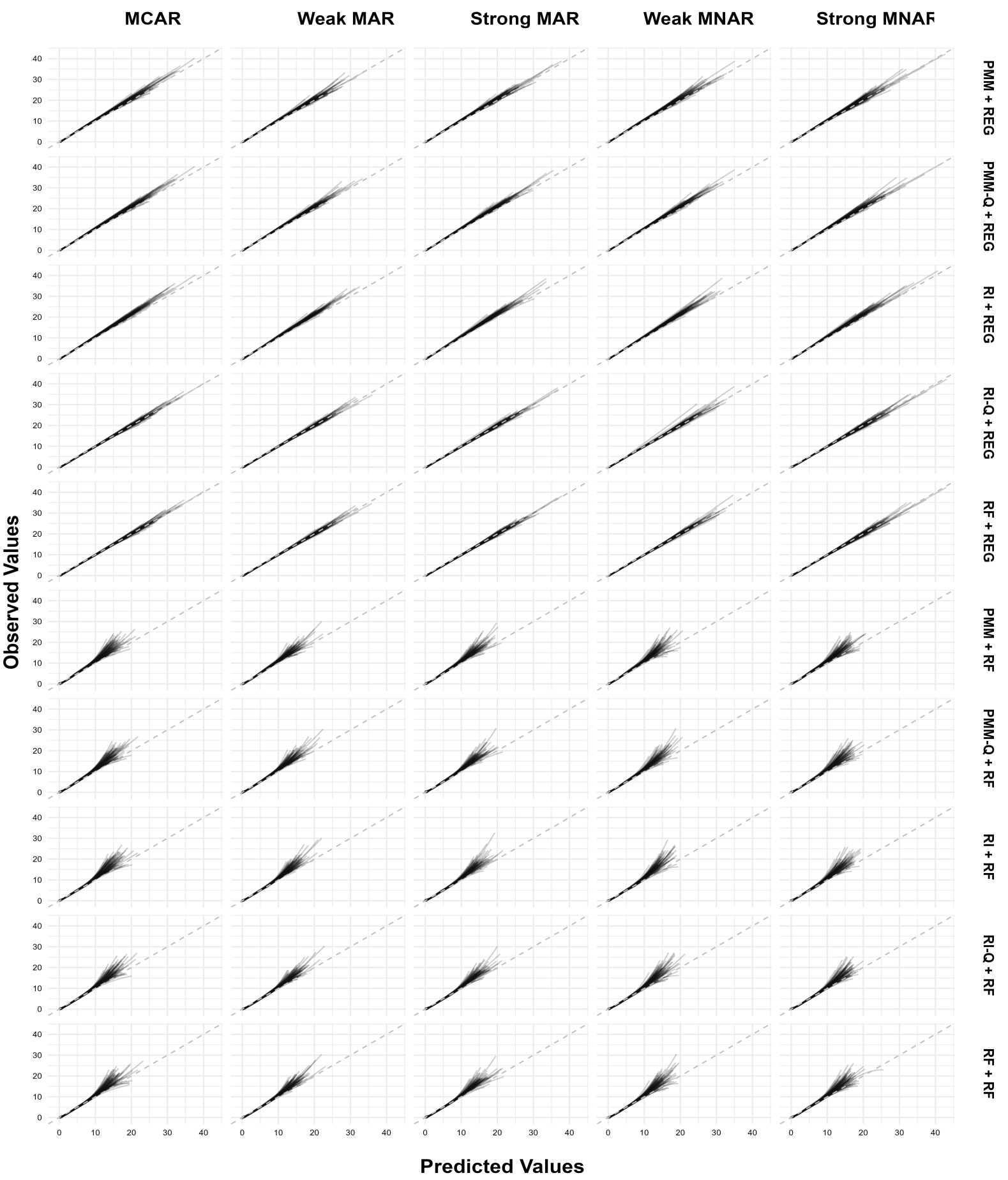

Figure 8: Calibration Plots of RL-Out: Quadratic, RL-Pred: Linear and Cor-Pred: Low.

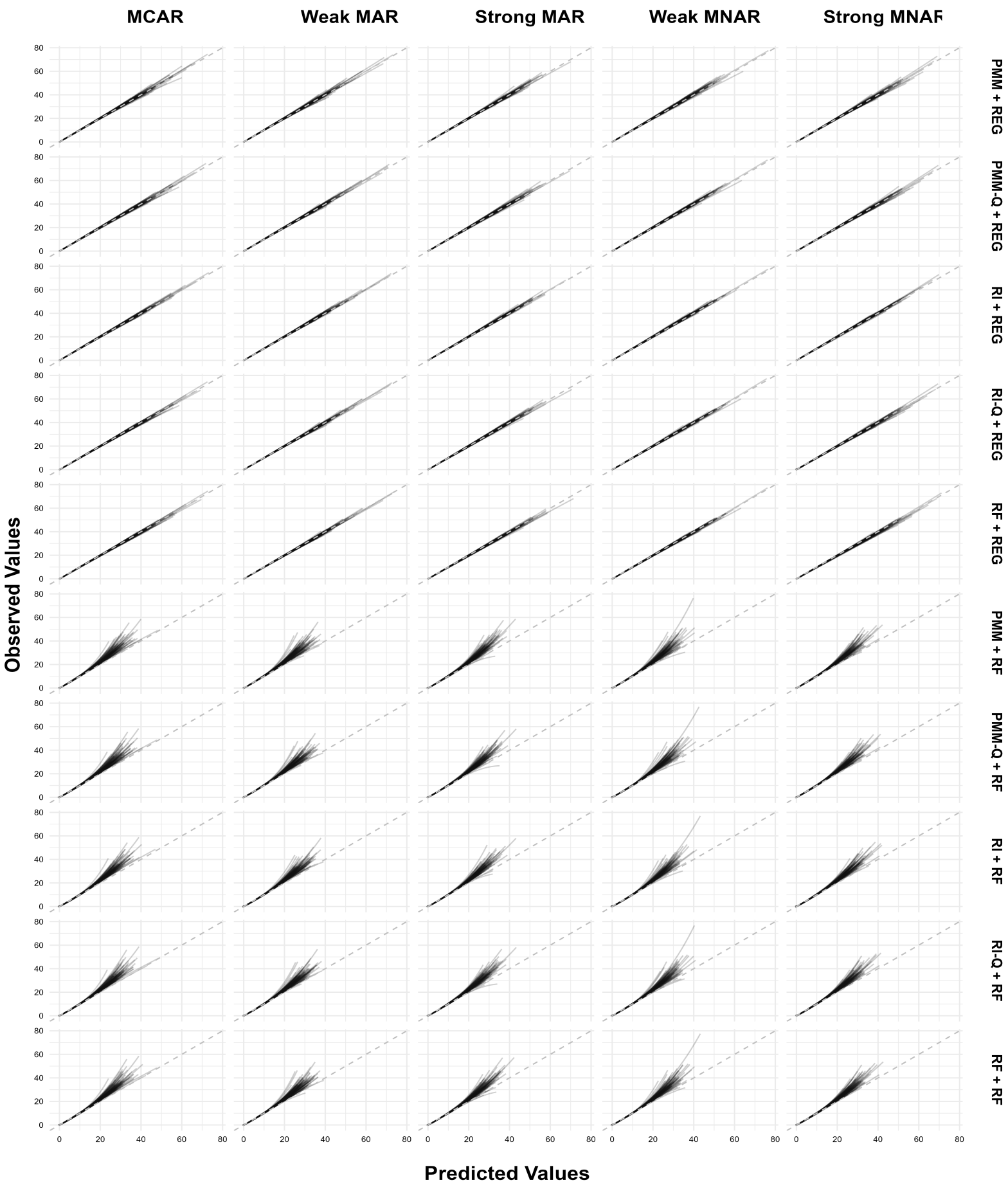

Figure 9: Calibration Plots of RL-Out: Quadratic, RL-Pred: Linear and Cor-Pred: High.

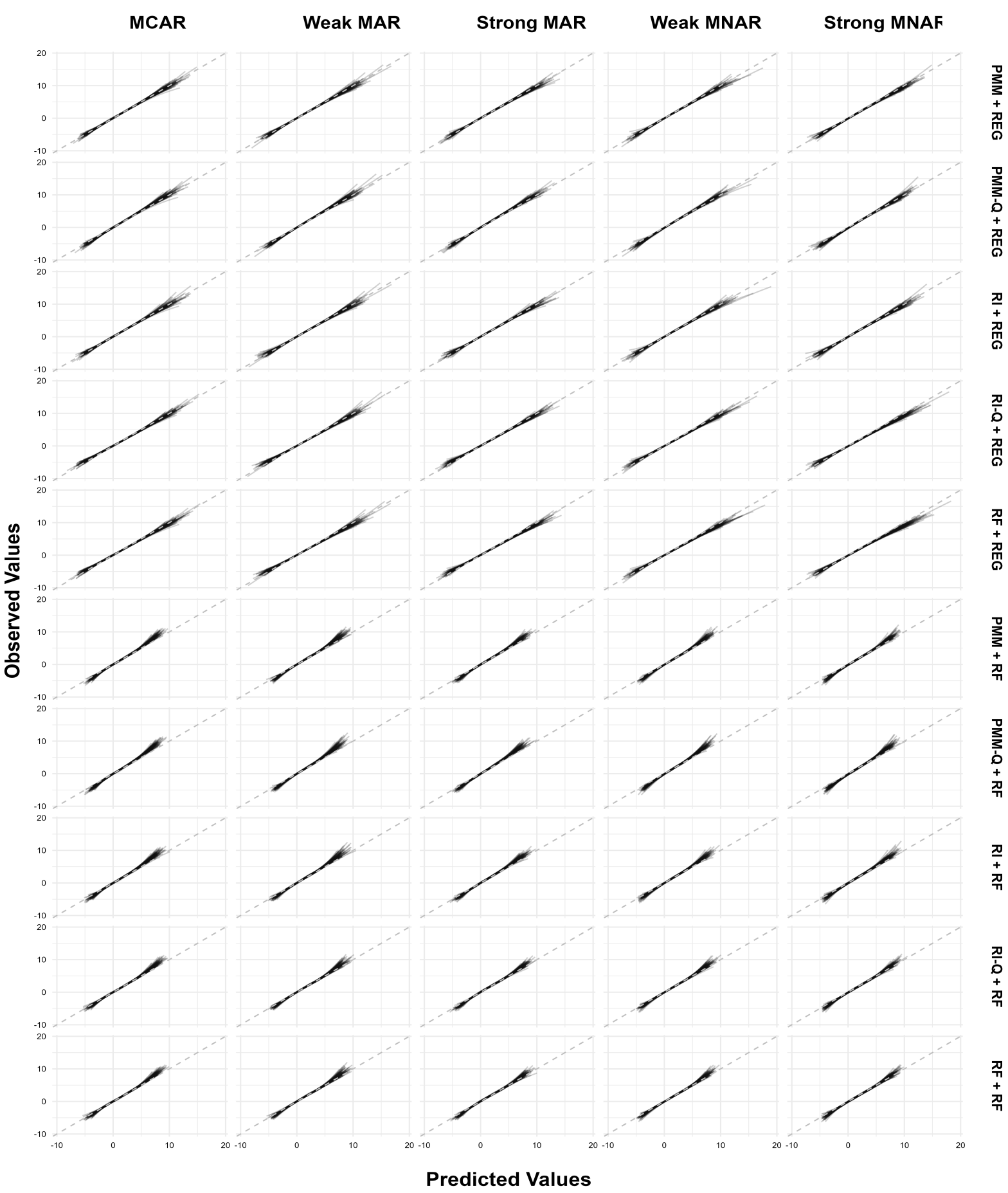

Figure 10: Calibration Plots of RL-Out: Linear, RL-Pred: Quadratic and Cor-Pred: Low.

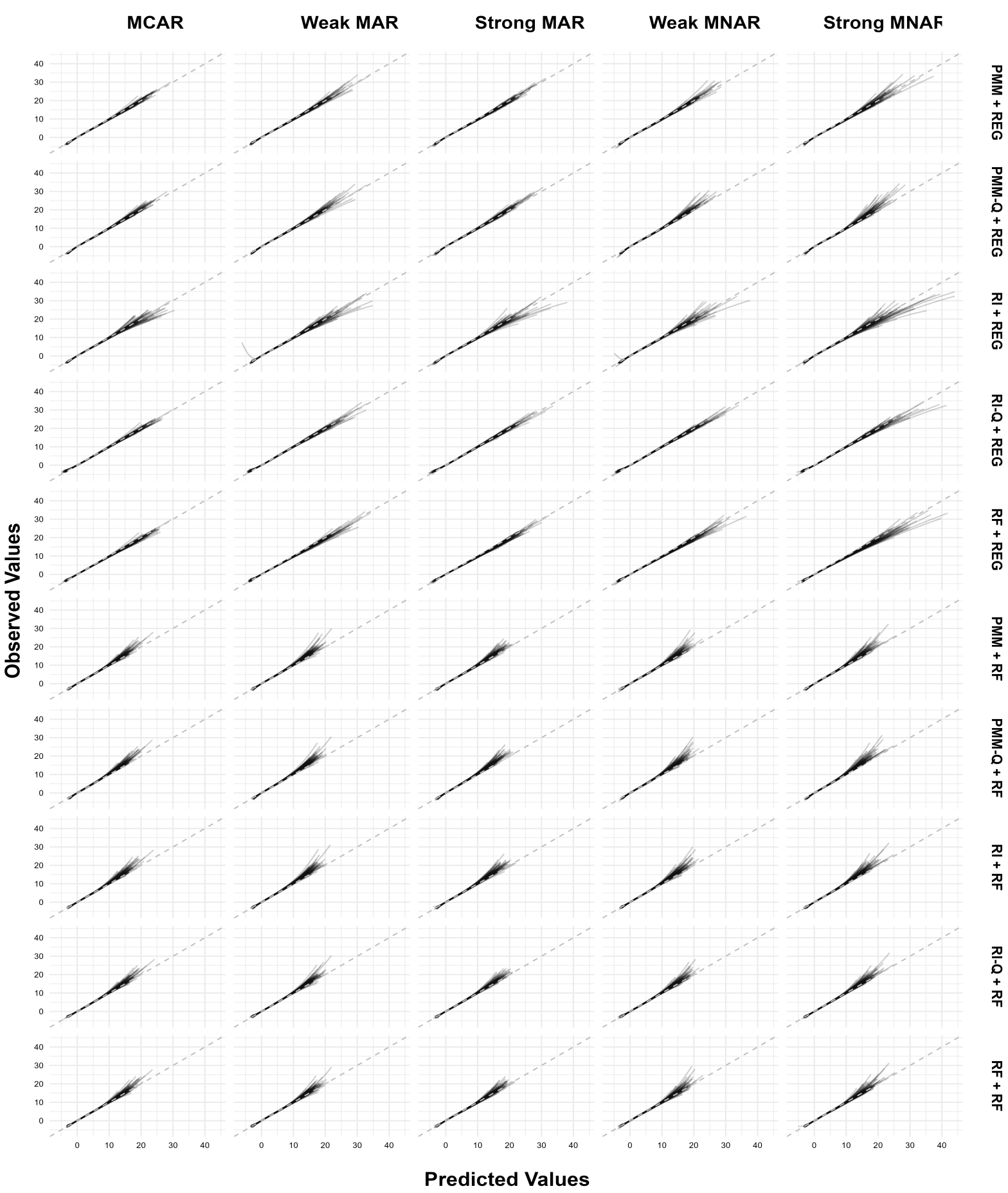

Figure 11: Calibration Plots of RL-Out: Linear, RL-Pred: Quadratic and Cor-Pred: High.

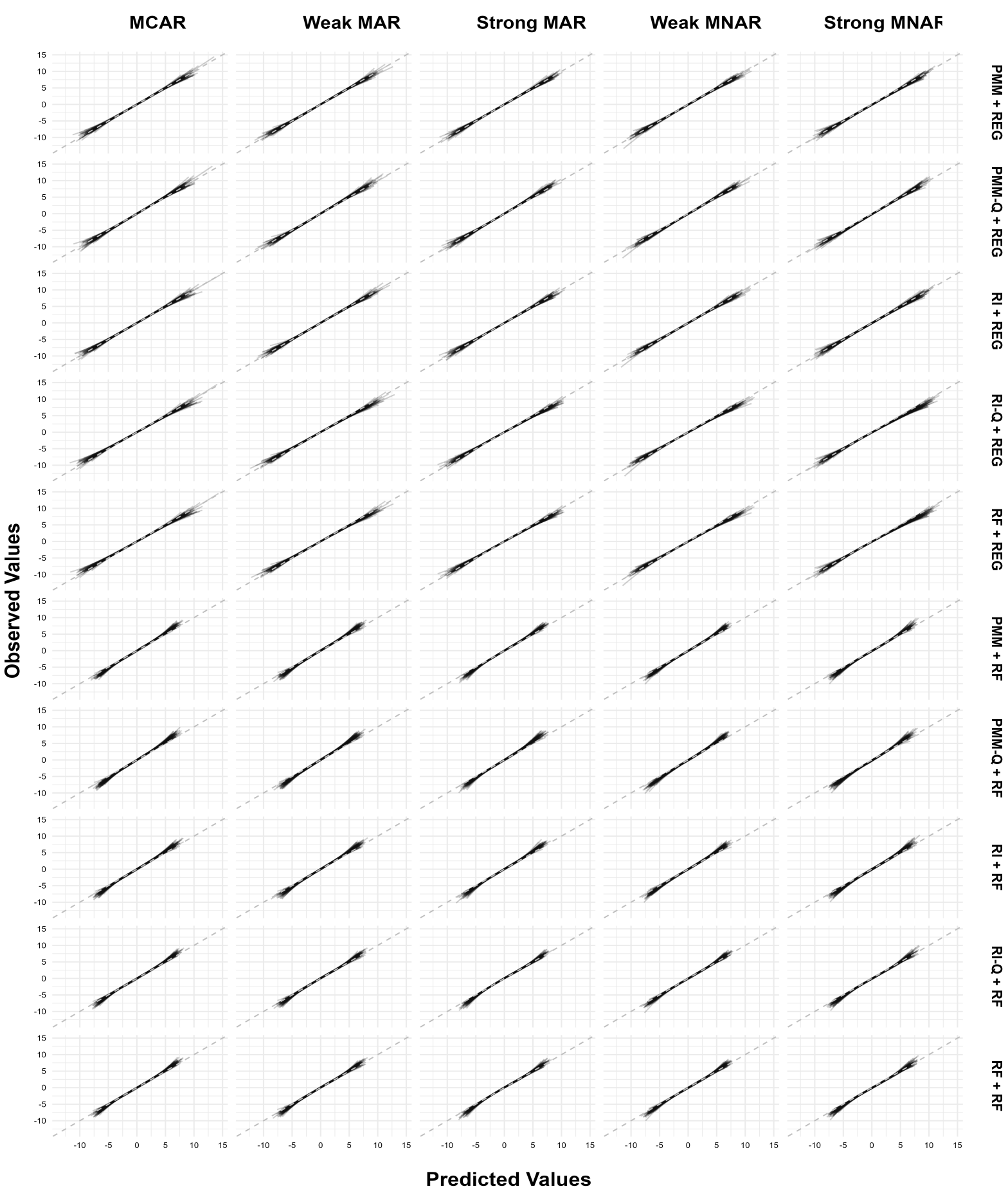

Figure 12: Calibration Plots of RL-Out: Linear, RL-Pred: Linear and Cor-Pred: Low.

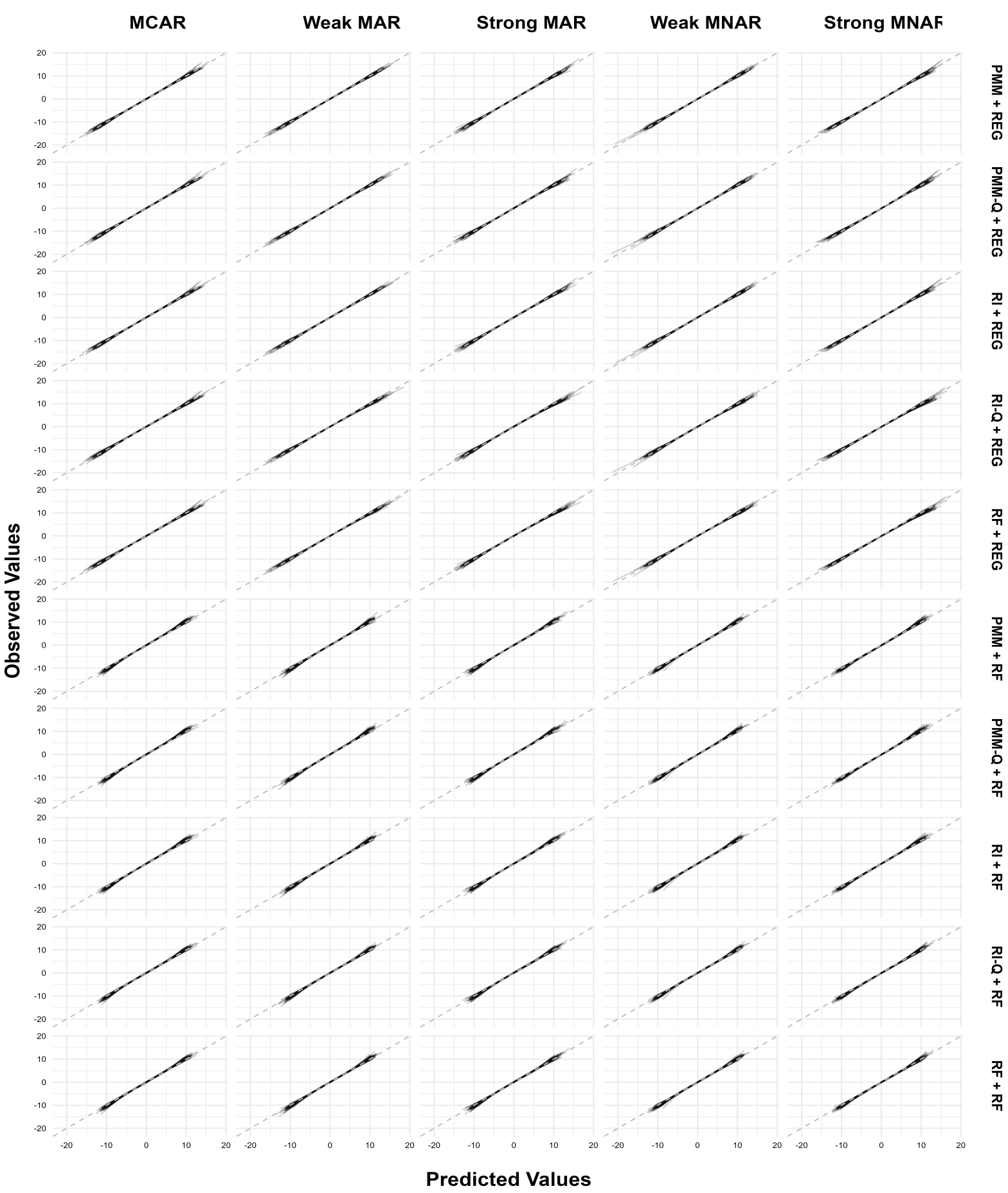

Figure 13: Calibration Plots of RL-Out: Linear, RL-Pred: Linear and Cor-Pred: High.