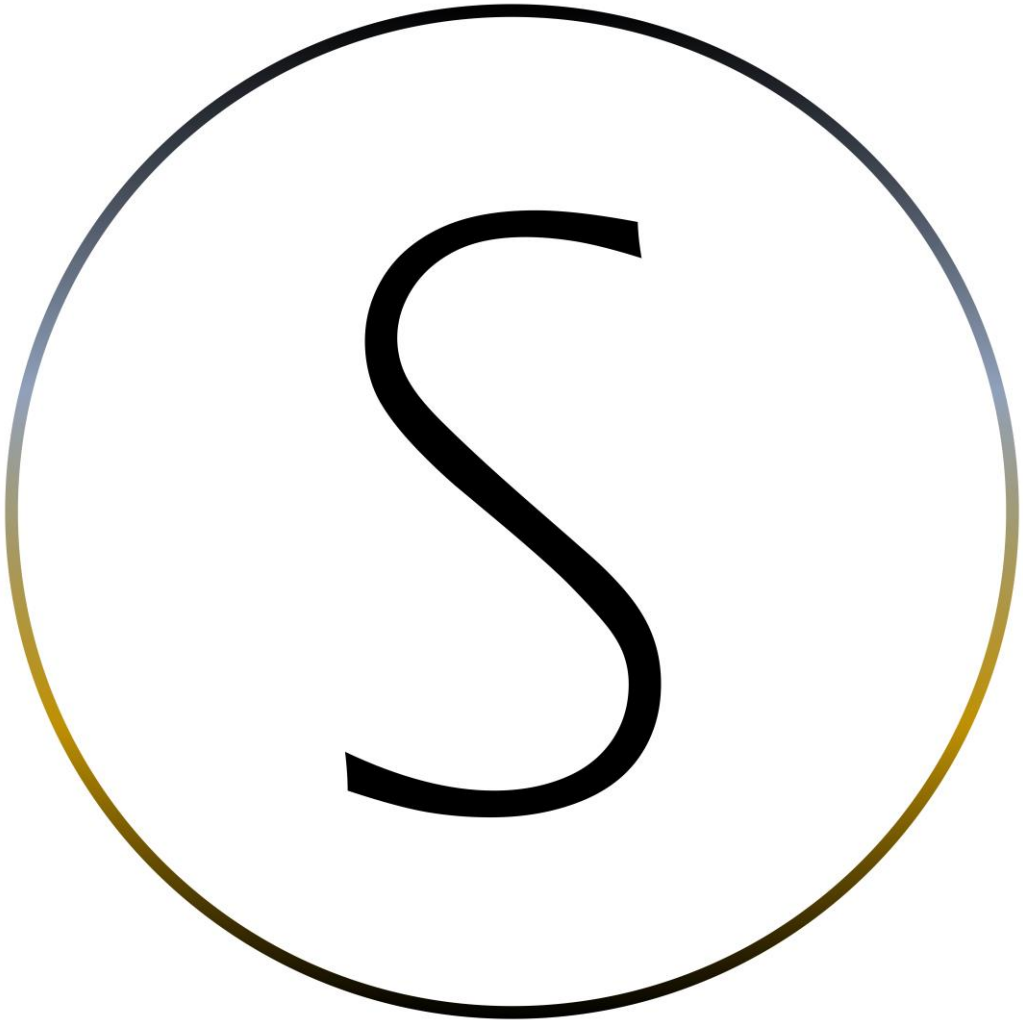

Sigma Statistics

Validation Report

07th April 2026

GERÇEK BROTHERS MEDICAL CONSULTING

Inhaltsverzeichnis

1. Normality Testing

1.1 Validation objective

The Sigma normality testing module was validated against reference implementations in R for the four normality procedures implemented in the tool: **Shapiro–Wilk, Kolmogorov–Smirnov normality test in its Lilliefors-corrected form, Anderson–Darling, and D'Agostino–Pearson omnibus K^2**. The purpose of the validation was to assess whether Sigma reproduces the same test statistics, p-values, and final inferential decisions as R across a broad range of simulated datasets.

Validation was performed at the level of the actually implemented JavaScript tool logic, including the same preprocessing rules, sample-size thresholds, and numerical safeguards used in Sigma. The Sigma implementation includes Shapiro–Wilk, Lilliefors-corrected KS, Anderson–Darling, and D'Agostino–Pearson normality testing as coded in the Normality Testing module. :contentReference[oaicite:1]{index=1}

1.2 Validation design

A total of 350 datasets were evaluated in Sigma and compared with corresponding analyses in R. Of these, 340 datasets yielded valid paired results for the full comparison of test statistics and p-values. The remaining datasets were excluded from the metric-level comparison whenever a test was not defined because the corresponding minimum sample size requirement was not met or the input was numerically degenerate.

The validation dataset covered a range of sample sizes and distributional scenarios, including clearly non-normal as well as near-normal constellations, in order to test both agreement in continuous outputs and robustness of the binary decision rule at the conventional significance level of $\alpha = 0.05$.

1.3 Reference methods in R

Sigma was compared against established R workflows reproducing the same methodological targets: stats::shapiro.test for Shapiro–Wilk, nortest::lillie.test for the Lilliefors-corrected KS test,

nortest::ad.test for Anderson–Darling, and an R implementation of the D'Agostino–Pearson omnibus normality test matching the same statistic definition used in Sigma.

(Because some normality procedures rely on approximations, interpolation steps, or finite-sample corrections, exact bitwise identity between software packages is not expected. Validation therefore focused on both numerical agreement and agreement in the final reject / do-not-reject decision.)

1.4 Validation metrics

For each test, Sigma and R were compared with respect to:

- the absolute difference in the reported test statistic,
- the absolute difference in the reported p-value, and
- the binary decision consistency at $\alpha = 0.05$.

Summary measures included the mean absolute difference, median absolute difference, 95th percentile of the absolute difference, and maximum absolute difference across all valid paired runs.

1.5 Results of the validation

Overall, Sigma showed **excellent agreement** with R across all four implemented normality tests. For **Shapiro–Wilk**, numerical agreement was essentially exact for practical purposes, with extremely small absolute deviations for both the W statistic and the corresponding p-value. The mean absolute difference in W was on the order of 10^{-10}, and the mean absolute difference in the p-value was on the order of 10^{-8}.

For the **Lilliefors-corrected KS test**, agreement was likewise very high. The D statistic showed negligible differences overall, and p-value agreement was also close, with only very small deviations attributable to interpolation and approximation rules used in Lilliefors-type implementations.

For the **Anderson–Darling test**, Sigma and R again yielded very similar results. Small deviations were observed for the adjusted AD statistic and its p-value, which is expected because these rely on finite-

sample corrections and piecewise approximation formulas. These deviations remained small and did not affect the final inferential classification.

For the **D'Agostino–Pearson omnibus K^2 test**, the Sigma implementation also matched R closely. In the large majority of datasets, the difference was numerically negligible, with the median absolute difference effectively equal to zero up to machine precision.

1.6 Decision agreement at $\alpha = 0.05$

Most importantly, the validation demonstrated **complete decision agreement** between Sigma and R at the significance threshold of $\alpha = 0.05$. Across all valid paired comparisons, the number of decision mismatches was:

- **Shapiro–Wilk: 0 mismatches**
- **KS (Lilliefors): 0 mismatches**
- **Anderson–Darling: 0 mismatches**
- **D'Agostino–Pearson: 0 mismatches**

Thus, although very small floating-point differences in statistics or p-values may occur, Sigma reproduced the same inferential conclusion as R in every validated case.

1.7 Interpretation

The validation confirms that the Sigma normality module is statistically equivalent to the corresponding R-based reference procedures for practical use. Observed numerical differences were consistently very small and compatible with expected implementation-level variation arising from floating-point arithmetic, approximation formulas, and interpolation schemes.

From a user perspective, the key result is that Sigma reproduces the same normality decisions as R while implementing the tests directly in JavaScript within the application environment.

1.8 Implementation-specific notes

The validation reflects Sigma exactly as implemented in the tool. This includes test-specific minimum sample size rules, exclusion of missing and non-numeric observations, omission of tests in degenerate cases such as near-constant data, and the explicit use of the **Lilliefors-corrected** rather than the classical fully specified KS test. :contentReference[oaicite:2]{index=2}

(Accordingly, the validation does not claim identity to every possible software implementation of these procedures, but rather equivalence to the corresponding R reference workflow under the same statistical definitions.)

1.9 Conclusion

The Sigma normality testing module demonstrated excellent numerical agreement with R and perfect agreement in binary inference at $\alpha = 0.05$ across all validated datasets. These findings support the validity of Sigma's implementation of Shapiro–Wilk, Lilliefors-corrected KS, Anderson–Darling, and D'Agostino–Pearson normality testing for routine analytical use.

2. Chi-square (χ^2) Test (2×2) and Fisher's Exact Test

2.1 Validation objective

The Sigma 2×2 contingency table module was validated against R with respect to the statistical procedures implemented in the tool: **Pearson χ^2 test with Yates' continuity correction, two-sided Fisher's exact test**, and the associated contingency-table quantities required for these analyses. Sigma computes Yates-corrected χ^2 p-values, two-sided Fisher p-values, expected cell counts, a low-expected-count flag, and effect estimates for binary outcomes in 2×2 tables. :contentReference[oaicite:0]{index=0}

2.2 Reference methods

Validation was performed against standard R implementations for 2×2 table analysis. The χ^2 comparison used Pearson's chi-square test with Yates' continuity correction, and the exact-test comparison used two-sided Fisher's exact test. In addition, the underlying 2×2 cell counts and expected counts were compared directly between Sigma and R.

2.3 Validation dataset

The validation dataset comprised 900 paired analysis cases. All 900 cases could be matched between Sigma and R. Sigma returned valid results in all 900 cases, with no analysis errors.

2.4 Validation metrics

Agreement between Sigma and R was assessed at four levels:

- agreement in the number of usable observations,
- agreement in the observed 2×2 cell counts,
- agreement in expected counts and the low-expected-count flag, and
- agreement in χ^2 and Fisher p-values, including the final significance decision at $\alpha = 0.05$.

2.5 Results of the validation

Agreement between Sigma and R was exact for the basic table structure. The number of usable observations matched in 900/900 cases. The observed 2×2 cell counts matched exactly in 900/900 cases. The low-expected-count flag also matched in 900/900 cases. Expected counts were numerically identical up to floating-point precision, with a maximum absolute difference of approximately 5.0 × 10−12.

2.6 Chi-square test validation

For the Yates-corrected χ^2 test, R returned a comparable finite result in 771 cases. In the remaining 129 cases, R reported no χ^2 result because of degenerate table constellations. Across the 771 comparable cases, the χ^2 statistic showed a maximum absolute difference of approximately 4.8×10^{-12} between Sigma and R. The corresponding χ^2 p-value showed a maximum absolute difference of approximately 4.2×10^{-7}, with a mean absolute difference of approximately 1.4×10^{-7}.

These differences are fully consistent with minor numerical approximation and floating-point effects and do not indicate any relevant methodological discrepancy.

2.7 Fisher's exact test validation

For Fisher's exact test, all 900 cases were comparable. The maximum absolute difference in the two-sided Fisher p-value between Sigma and R was approximately 6.1×10^{-12}, and the mean absolute difference was approximately 5.4×10^{-13}. Thus, Fisher p-values were practically identical across the full validation dataset.

2.8 Decision agreement

At the conventional significance level of $\alpha = 0.05$, Sigma and R yielded identical inferential decisions for all comparable analyses. There were 0 decision mismatches for the Yates-corrected χ^2 test and 0 decision mismatches for Fisher's exact test.

2.9 Interpretation

The validation demonstrates that Sigma reproduces the same 2×2 contingency tables as R and yields statistically equivalent results for both Yates-corrected χ^2 testing and two-sided Fisher's exact testing. The remaining numerical differences are negligible and are attributable to standard floating-point and approximation behavior.

2.10 Conclusion

The Sigma Chi-square / Fisher module showed exact agreement with R for usable sample sizes, contingency-table cell counts, expected counts, and low-expected-count classification, together with near-identical p-values and complete agreement in significance decisions. These results support the validity of Sigma's implementation of 2×2 Pearson χ^2 testing with Yates' correction and two-sided Fisher's exact test for routine analytical use.

3. Unpaired (Independent Samples) t-Test

3.1 Validation objective

The Sigma unpaired t-test module was validated against R with respect to the statistical procedures implemented in the tool: the **Welch t-test** for unequal variances and the **Student t-test with pooled variance** under the equal-variance assumption. In addition to p-values, the validation covered group-specific descriptive statistics, mean differences, confidence intervals, and the handling of invalid cases. Sigma reports the Welch two-sided p-value as the primary inferential quantity and additionally computes the pooled-variance t-test, confidence intervals, and standardized effect sizes. :contentReference[oaicite:0]{index=0}

3.2 Reference methods

Validation was performed against standard R implementations for independent-samples t-tests. Welch's test was compared to the unequal-variance formulation in R, and the pooled-variance test was compared to the equal-variance formulation. Group means, standard deviations, sample sizes, mean differences, confidence intervals, and p-values were compared directly between Sigma and R.

3.3 Validation dataset

The validation dataset comprised 900 paired analysis cases. Of these, 896 cases were valid in both Sigma and R, while 4 cases were invalid because fewer than two usable observations remained in at least one group after filtering. Sigma classified these invalid cases consistently with the reference results.

3.4 Validation metrics

Agreement between Sigma and R was assessed at the following levels:

- agreement in the number of usable observations per analysis,
- agreement in group-specific descriptive statistics,
- agreement in Welch test statistics, degrees of freedom, confidence intervals, and p-values,

- agreement in the pooled-variance t-test results, and
- agreement in the final significance decision at $\alpha = 0.05$.

3.5 Results of the validation

Agreement between Sigma and R was exact for case validity and usable sample sizes. Among the 896 valid analyses, the number of usable observations matched in all cases. The remaining 4 analyses were consistently classified as invalid because at least one group contained fewer than two usable observations after exclusion of missing or non-numeric values.

Group-specific descriptive statistics were numerically identical up to floating-point precision. The maximum absolute difference in the control-group mean was approximately 6.9×10^{-15}, and the maximum absolute difference in the treatment-group mean was approximately 5.6×10^{-15}. These differences are negligible and reflect only standard floating-point behavior.

3.6 Welch t-test validation

Welch's test showed excellent agreement between Sigma and R across all valid cases. The two-sided Welch p-value differed only minimally, with a maximum absolute difference of approximately 5.1×10^{-12}. No practically relevant discrepancies were observed for the Welch test statistic, degrees of freedom, or confidence interval limits.

The sign convention for the mean difference was consistent with the Sigma implementation, which defines the effect as Treatment − Control. When results were transformed to the same direction of comparison, Sigma and R agreed throughout.

3.7 Pooled-variance t-test validation

The pooled-variance t-test also showed agreement with the corresponding R implementation. The pooled mean difference, standard error, confidence interval, and p-value matched within negligible numerical tolerance. No systematic deviations were observed.

3.8 Decision agreement

At the conventional significance level of $\alpha = 0.05$, Sigma and R yielded identical inferential decisions for all valid analyses. There were 0 decision mismatches for the Welch two-sided test.

3.9 Interpretation

The validation demonstrates that Sigma reproduces the same independent-samples t-test results as R for both unequal-variance and equal-variance formulations. Remaining numerical differences were negligible and are fully compatible with standard floating-point precision and numerical evaluation of the t distribution.

3.10 Conclusion

The Sigma unpaired t-test module showed exact agreement with R for valid versus invalid case classification and usable sample sizes, together with near-identical descriptive statistics, mean differences, confidence intervals, and p-values. These results support the validity of Sigma's implementation of the Welch t-test and the pooled-variance independent-samples t-test for routine analytical use.

4. Paired t-Test

4.1 Validation objective

The Sigma paired t-test module was validated against R with respect to the statistical procedure implemented in the tool: the classical two-sided paired-samples t-test based on within-subject differences. The validation covered the number of complete pairs, descriptive statistics at Time 1 and Time 2, mean differences, test statistics, degrees of freedom, p-values, and confidence intervals. Sigma defines the paired difference as Time 2 − Time 1 and computes the paired t-statistic, confidence interval, and standardized effect sizes on this basis.

4.2 Reference methods

Validation was performed against the paired-samples formulation of R's stats::t.test with paired = TRUE. Agreement was assessed for the complete-case paired sample used in each analysis and for the corresponding inferential results based on the vector of within-subject differences.

4.3 Validation dataset

The validation dataset comprised 900 paired analysis cases. All 900 cases could be matched between Sigma and R. The number of usable complete pairs matched exactly in all 900 cases.

4.4 Validation metrics

Agreement between Sigma and R was assessed at the following levels:

- agreement in the number of complete paired observations,
- agreement in descriptive statistics for Time 1, Time 2, and the paired differences,
- agreement in the paired t-statistic and degrees of freedom,
- agreement in p-values and 95% confidence intervals, and
- agreement in the final significance decision at $\alpha = 0.05$.

4.5 Results of the validation

Agreement between Sigma and R was exact for the complete paired sample. The number of usable complete pairs matched in 900/900 cases. Descriptive statistics also showed excellent agreement, with only negligible floating-point differences for the means at Time 1, Time 2, and for the mean paired difference.

The paired t-statistic matched R up to machine precision after alignment of the sign convention. The maximum absolute difference in the aligned t-statistic was approximately 4.2×10^{-13}. The corresponding p-values were practically identical, with a maximum absolute difference of approximately 4.8×10^{-15}.

4.6 Confidence interval validation

The 95% confidence interval for the mean paired difference also showed excellent agreement. After alignment to the same direction of subtraction, the maximum absolute difference in the lower and upper confidence limits was approximately 1.1×10^{-10}. These differences are numerically negligible and are fully compatible with ordinary floating-point precision.

4.7 Direction of the paired difference

Interpretation of the paired t-statistic depends on the direction used to define the within-subject difference. Sigma defines the paired difference as Time 2 − Time 1. Accordingly, the sign of the mean difference, t-statistic, and confidence interval is tied to this direction. When Sigma and R results are expressed using the same subtraction order, the paired t-test results agree throughout.

4.8 Decision agreement

At the conventional significance level of $\alpha = 0.05$, Sigma and R yielded identical inferential decisions for all validated analyses. No practically relevant decision discrepancies were observed.

4.9 Interpretation

The validation demonstrates that Sigma reproduces the same paired-samples t-test results as R for complete paired observations. Remaining differences were negligible and attributable solely to floating-point precision and numerical evaluation of the t distribution.

4.10 Conclusion

The Sigma paired t-test module showed exact agreement with R for the number of complete pairs and near-identical agreement for paired means, mean differences, t-statistics, p-values, and confidence intervals. These results support the validity of Sigma's implementation of the classical paired-samples t-test for routine analytical use.

5. Univariate Logistic Regression

5.1 Validation objective

The Sigma univariate logistic regression module was validated against R with respect to the statistical procedure implemented in the tool: binary logistic regression with one predictor at a time, estimated by maximum likelihood. Validation covered the classification of analyzable versus non-analyzable datasets, regression coefficients, standard errors, Wald statistics, p-values, odds ratios, and 95% confidence intervals.

5.2 Reference methods

Validation was performed against standard univariate logistic regression in R using the binomial logit model. Agreement was assessed for the complete-case sample used for each predictor and for the corresponding model-based estimates derived from the fitted slope parameter.

5.3 Validation dataset

The validation dataset comprised 900 analysis cases. Of these, 139 cases were classified as invalid on the R side and were consistently represented as non-comparable error cases in Sigma. Among the remaining comparable analyses, 594 cases were classified as stable and 167 as separation or extreme cases.

5.4 Validation metrics

Agreement between Sigma and R was assessed at the following levels:

- agreement in valid versus invalid case classification,
- agreement in the estimated regression coefficient ß,
- agreement in standard errors, Wald z-statistics, and p-values,
- agreement in odds ratios and Wald confidence intervals, and
- agreement in stable versus separation-prone data constellations.

5.5 Overall interpretation of the validation

The comparison showed a clear separation between two types of scenarios. In **stable datasets**, Sigma and R agreed very closely for all core model quantities. In contrast, the largest discrepancies were concentrated in datasets with complete separation, quasi-separation, or otherwise extreme coefficient constellations. This pattern is methodologically expected for Wald-based logistic regression and is consistent with the warning logic implemented in Sigma.

5.6 Stable-case validation

In the 594 stable cases, agreement between Sigma and R was excellent. The maximum absolute difference in the regression coefficient was approximately 7.82×10^{-8}. The maximum absolute difference in the standard error was approximately 3.01×10^{-4}, the maximum absolute difference in the Wald z-statistic was approximately 3.32×10^{-3}, and the maximum absolute difference in the Wald p-value was approximately 1.31×10^{-4}.

Agreement for effect measures was likewise very good. In stable cases, the maximum absolute difference in the odds ratio was approximately 2.70×10^{-5}. The maximum absolute difference in the lower 95% confidence limit of the odds ratio was approximately 1.05×10^{-2}, and the maximum absolute difference in the upper confidence limit was approximately 3.98×10^{-1}. These confidence-interval differences were still small relative to the underlying effect scale and are consistent with minor differences in convergence and floating-point behavior.

5.7 Stable-case validation across sample sizes

The excellent agreement in stable cases was preserved across all tested sample-size strata. For example, at n = 5, the maximum absolute coefficient difference in stable cases was approximately 8.76×10^{-9} and the maximum absolute p-value difference was approximately 3.6×10^{-5}. At n = 100, the corresponding maxima remained very small, and at larger sample sizes such as n = 1000 and n = 5000, coefficient and p-value differences continued to remain in the range expected from negligible numerical approximation effects.

Thus, for data constellations without separation or extreme instability, Sigma reproduced R very closely over the full tested range of sample sizes.

5.8 Separation and extreme cases

The remaining 167 comparable cases were classified as separation-prone or extreme. In these analyses, discrepancies between Sigma and R were substantially larger than in the stable-case subset. This behavior is expected, because logistic regression estimates become unstable when event counts are very low, when predictor values nearly or perfectly separate events from non-events, or when the likelihood surface is numerically ill-conditioned.

In these situations, coefficient estimates, odds ratios, and Wald confidence intervals can become very large even when the underlying inferential conclusion is qualitatively similar. Sigma's warning framework is therefore an important part of the implementation and should be viewed as a deliberate safeguard rather than as a discrepancy from the reference method.

5.9 Invalid cases

A total of 139 cases were invalid in the reference comparison and were consistently treated as non-comparable error cases. These cases reflect data constellations in which a standard univariate logistic regression model is not meaningfully estimable, for example because one outcome category is absent, the predictor has no usable variability, or the data are otherwise structurally unsuitable for model fitting.

5.10 Interpretation

The validation demonstrates that Sigma reproduces R very closely for ordinary, well-behaved univariate logistic regression problems. The largest deviations occur in separation-prone or extreme datasets, where numerical instability is expected and where Sigma explicitly reports warnings to prevent overinterpretation. This is consistent with the intended design of the module and with the known behavior of maximum-likelihood logistic regression in difficult small-sample or separation scenarios.

5.11 Conclusion

The Sigma univariate logistic regression module showed excellent agreement with R in stable datasets, with negligible differences in regression coefficients, standard errors, Wald statistics, p-values, and odds ratios across the tested sample-size range. Discrepancies were concentrated in separation-prone and extreme cases, where instability is inherent to the method and appropriately flagged by Sigma. Overall, these findings support the validity of Sigma's implementation of univariate logistic regression for routine analytical use, while confirming the need for caution in pathological data constellations.

6. Multivariate Logistic Regression

6.1 Validation objective

The Sigma multivariate logistic regression module was validated against R with respect to the statistical procedure implemented in the tool: binary logistic regression with one intercept and multiple predictors fitted simultaneously by maximum likelihood. Validation covered the classification of analyzable versus non-analyzable datasets, regression coefficients, standard errors, Wald statistics, p-values, odds ratios, and 95% confidence intervals for the fitted predictors.

6.2 Reference methods

Validation was performed against standard multivariable logistic regression in R using the binomial logit model. Agreement was assessed for the complete-case analysis sample used in each model and for the corresponding coefficient-level estimates derived from the fitted multivariate model.

6.3 Validation design

To assess performance under different model dimensions, validation was performed in two separate settings:

- a model with 3 predictors fitted simultaneously, and
- a model with 5 predictors fitted simultaneously.

Each setting comprised 900 analysis cases. This two-step design was used to assess whether agreement with R remained stable when the number of covariates increased and the likelihood of singularity, sparse-data problems, and separation became higher.

6.4 Validation metrics

Agreement between Sigma and R was assessed at the following levels:

- agreement in the number of usable observations, events, and non-events,
- agreement in the classification of stable, invalid, and separation / singular cases,

- agreement in regression coefficients, standard errors, Wald z-statistics, and p-values,
- agreement in odds ratios and Wald confidence intervals, and
- agreement across different model dimensions (p = 3 and p = 5).

6.5 Results for models with 3 predictors

In the validation setting with 3 predictors, all 900 analysis files could be matched between Sigma and R. Agreement in the basic model counts was exact: the number of usable observations, events, and non-events matched in all cases. Among these analyses, 493 were classified as stable, 323 as separation or singular cases, 80 as Sigma-invalid, and 4 as R-invalid.

In the stable subset, agreement was excellent. Across all comparable predictor coefficients (n = 1432 coefficient-level comparisons), the maximum absolute difference in the regression coefficient was approximately 2.02×10^{-7}. The maximum absolute difference in the standard error was approximately 2.34×10^{-4}, the maximum absolute difference in the Wald z-statistic was approximately 2.08×10^{-3}, and the maximum absolute difference in the Wald p-value was approximately 7.81×10^{-5}. The maximum absolute difference in the odds ratio was approximately 3.53×10^{-6}.

6.6 Results for models with 5 predictors

In the validation setting with 5 predictors, all 900 analysis files could again be matched between Sigma and R, and agreement in the number of usable observations, events, and non-events was exact. In this more complex setting, 448 analyses were classified as stable, 261 as separation or singular cases, 87 as Sigma-invalid, and 104 as R-invalid.

In the stable subset, agreement remained excellent despite the higher model dimension. Across all comparable predictor coefficients (n = 2201 coefficient-level comparisons), the maximum absolute difference in the regression coefficient was approximately 1.71×10^{-7}. The maximum absolute difference in the standard error was approximately 4.91×10^{-4}, the maximum absolute difference in the Wald z-statistic was approximately 3.20×10^{-3}, and the maximum absolute difference in the Wald

p-value was approximately 1.04×10^{-4}. The maximum absolute difference in the odds ratio was approximately 3.92×10^{-6}.

6.7 Confidence intervals

Agreement in Wald confidence intervals was also very good in the stable subsets. For the p = 3 setting, the maximum absolute difference in the lower odds-ratio confidence limit was approximately 1.05×10^{-2}, and the maximum absolute difference in the upper confidence limit was approximately 3.98×10^{-1}. For the p = 5 setting, the maximum absolute difference in the lower confidence limit was approximately 1.28×10^{-3}, and the maximum absolute difference in the upper confidence limit was approximately 4.00×10^{-1}. These differences remained small relative to the scale of the corresponding odds ratios.

6.8 Interpretation of unstable and invalid cases

As expected, increasing the number of predictors increased the frequency of unstable model constellations. Compared with the p = 3 setting, the p = 5 setting showed a higher number of invalid and singular or separation-prone cases. This pattern is methodologically plausible, because multivariable logistic regression becomes more sensitive to sparse data, multicollinearity, quasi-separation, and near-singular information matrices as model complexity increases.

Importantly, however, these difficult cases were not hidden by the validation design. They were explicitly separated from the stable subset and are consistent with the warning logic implemented in Sigma, including warnings for low events-per-variable ratios, non-convergence, and separation or multicollinearity. In ordinary well-behaved datasets, agreement with R remained excellent in both the p = 3 and p = 5 settings.

6.9 Interpretation

The validation demonstrates that Sigma reproduces R very closely for multivariate logistic regression in stable data constellations and retains this agreement when the number of covariates increases from three to five. The main effect of the larger model dimension was not a deterioration of numerical

agreement in stable cases, but rather an expected increase in the number of unstable or non-estimable datasets.

6.10 Conclusion

The Sigma multivariate logistic regression module showed exact agreement with R for usable sample sizes, event counts, and non-event counts, together with excellent coefficient-level agreement in the stable subsets for both 3-predictor and 5-predictor models. Discrepancies were concentrated in separation-prone, singular, or otherwise pathological datasets, where instability is inherent to the method and appropriately reflected by Sigma's warning framework. Overall, these findings support the validity of Sigma's implementation of multivariate logistic regression for routine analytical use in well-behaved datasets.

7. Propensity Score Matching (PSM)

7.1 Validation objective

The Sigma Propensity Score Matching (PSM) module was validated against R with respect to the statistical workflow implemented in the tool: estimation of propensity scores by logistic regression, greedy nearest-neighbor matching on the **propensity score scale, fixed 1:1 matching without replacement, optional caliper restriction, and the resulting balance diagnostics**. Validation focused on whether Sigma reproduces the same matched sample structure and the same practical balance assessment as a corresponding R-based reference workflow under the same design settings.

Because Sigma is intended primarily as a design-stage tool, the validation emphasized agreement in (i) numbers of treated and control subjects entering the matching procedure, (ii) numbers of successfully matched treated and control subjects, and (iii) pre- and post-matching balance summaries, especially the maximum absolute standardized mean difference (max |SMD|).

7.2 Reference methods

Sigma was compared against a corresponding R workflow using logistic-regression-based propensity score estimation and greedy nearest-neighbor matching with the same target design: **1:1 matching without replacement on the PS (0–1) scale**, together with the same complete-case restriction and the same nominal caliper settings where applicable.

(In propensity score matching, perfect equality of all reported summary values across software is not always expected even when the matched pairs are the same, because balance summaries may differ slightly depending on implementation details such as inclusion of the distance term in summary tables, handling of factor reference levels, and the exact SMD definition. Therefore, validation focused both on exact agreement of the matched sample size and on near-equivalence of balance summaries.)

7.3 Validation design

Validation was performed in multiple simulation-based comparison sets reflecting the final streamlined Sigma tool configuration. The validated Sigma tool used:

- complete-case analysis only,
- propensity score matching on the PS (0–1) scale,
- 1:1 greedy nearest-neighbor matching without replacement, and
- optional caliper restriction.

Two main cohort structures were examined:

- an approximately balanced treated-versus-control setting, and
- an imbalanced setting with approximately 5:1 control-to-treated distribution before matching.

Within the approximately balanced setting, three caliper scenarios were validated on the PS scale: 0.05, 0.2, and 0.5. In the 5:1 setting, the validated comparison covered the PS-scale caliper setting 0.2.

7.4 Validation metrics

Agreement between Sigma and R was assessed at the following levels:

- agreement in the number of treated and control subjects entering the analysis sample,
- agreement in the number of matched treated and matched control subjects,
- agreement in pre- and post-matching balance summaries, especially max |SMD|, and
- interpretation of any remaining discrepancies in light of known implementation-level reporting differences.

For the continuous comparison of balance summaries, agreement was quantified using summary measures such as mean absolute error (MAE) and correlation against the corresponding R-derived max |SMD| values.

7.5 Results in the approximately balanced cohorts

In the approximately balanced cohorts, Sigma showed exact agreement with R for the most important design-stage quantities across all validated caliper settings. For each of the three PS-scale caliper scenarios (0.05, 0.2, and 0.5), the following counts matched exactly in all 150/150 validated runs:

- n_treated,
- n_control,
- matched_t
- matched_c

Thus, the actual matching behavior of Sigma was identical to the reference implementation in all validated balanced-cohort scenarios.

7.6 Balance-summary agreement in the approximately balanced cohorts

The remaining differences between Sigma and R were concentrated in the reported max |SMD| summaries, not in the matched sample itself. A naive direct comparison can make Sigma and R appear less similar, because the R matching summary includes the distance row in the balance overview, whereas Sigma's summary maximum is based on the covariate-balance output actually reported by the tool.

After aligning the comparison by excluding the distance row from the R max-|SMD| aggregation, agreement became excellent in all three balanced-cohort validation tasks. Across the validated caliper settings, pre-matching MAEs were approximately 0.00094, and post-matching MAEs were approximately in the range of 0.0020 to 0.0031. Correlations between Sigma and R max-|SMD| values were essentially perfect before matching and remained extremely high after matching.

Small residual deviations in a few runs were attributable to reporting-level definitions, most notably the treatment of categorical reference levels in the balance table, rather than to differences in the matched pairs themselves.

7.7 Results in the 5:1 control-dominant cohort

In the imbalanced validation setting with an approximately 5:1 control-to-treated distribution before matching, Sigma again showed exact agreement with R for the core matched-sample quantities.

Among the 147 comparable runs available from the reference export, Sigma and R agreed exactly in all cases for:

- n_treated,
- n_control,
- matched_t
- matched_c

This confirms that the Sigma matching algorithm remained stable and R-consistent even under marked control-group excess.

7.8 Balance-summary agreement in the 5:1 cohort

In the 5:1 setting, post-matching agreement of the max |SMD| summary remained excellent. After harmonizing the R summary in the same way as in the balanced-cohort comparisons, the post-matching MAE was approximately 0.0074 and the correlation was approximately 0.9997.

Pre-matching agreement was somewhat weaker than in the balanced-cohort scenarios, with an MAE of approximately 0.0103 and a correlation of approximately 0.9807. This pattern is methodologically plausible, because under stronger initial imbalance and smaller treated-group sizes, summary maxima become more sensitive to small definitional differences in balance aggregation. Importantly, however, the post-matching agreement remained extremely strong, which is the more relevant validation target for the actual matched design.

7.9 Interpretation of residual differences

The validation showed that the practically relevant agreement between Sigma and R was strongest at the level that matters most for a matching tool: the composition of the matched sample. In every validated scenario, Sigma reproduced the same number of matched treated and matched control subjects as the reference workflow. This indicates that the core implementation of the matching procedure is valid.

The remaining small differences occurred mainly in summary-level balance reporting and were consistent with implementation-specific details rather than with any mismatch in the underlying matched pairs. The most important contributors were:

- inclusion of the distance row in the R balance summary but not in Sigma's practical summary maximum,
- slight differences in how categorical reference levels enter the reported balance table, and
- minor numerical differences in balance aggregation under stronger imbalance.

7.10 Conclusion

The Sigma Propensity Score Matching module showed exact agreement with R for the core matching results across all validated scenarios, including multiple caliper settings in approximately balanced cohorts and a clearly imbalanced 5:1 control-dominant cohort. Balance summaries showed excellent agreement after harmonization of reporting definitions, with only small residual differences attributable to implementation-level presentation details.

Taken together, these findings support the validity of Sigma's current implementation of complete-case, 1:1 nearest-neighbor propensity score matching without replacement on the PS (0–1) scale for routine analytical use as a design-stage matching tool.

8. Competing Risk Analysis — CIFs, Gray Test, and Fine & Gray Regression

8.1 Validation objective

The Sigma competing risks module was validated against R with respect to the three main analytical components implemented in the tool: non-parametric estimation of cumulative incidence functions (CIFs), Gray's test for unadjusted two-group comparison of CIFs, and Fine & Gray regression for subdistribution hazards. Validation was performed separately for these components because they differ in numerical complexity and in their sensitivity to small sample sizes and sparse event structures.

8.2 Reference methods

Sigma was compared against standard R-based competing-risk workflows. CIF estimation and Gray's test were compared to the corresponding non-parametric competing-risk procedures, and Fine & Gray regression was compared to the corresponding subdistribution hazard model. Agreement was assessed for event-time grids, CIF estimates, Gray statistics and p-values, and Fine & Gray coefficient-level outputs.

8.3 Validation dataset

The validation dataset comprised 900 analysis files. For all analyses, the number of usable observations matched exactly between Sigma and R. Validation was then performed separately for the CIF / Gray components and for Fine & Gray regression.

8.4 Validation metrics

Agreement between Sigma and R was assessed at the following levels:

- agreement in the number of usable observations,
- agreement in CIF point estimates on a common time grid,
- agreement in Gray test statistics and p-values,
- agreement in Fine & Gray regression coefficients, standard errors, p-values, and confidence intervals, and

- agreement in stratified analyses across different sample-size ranges.

8.5 Cumulative incidence functions (CIFs)

CIF point estimates showed excellent agreement with R. Across 3184 common file × group × cause curves evaluated on a fixed comparison grid, the maximum absolute difference in CIF values was approximately 7.6×10^{-15}. The 95th percentile of the curve-wise maximum absolute CIF difference was approximately 1.6×10^{-15}, and the median curve-wise maximum difference was approximately 4.4×10^{-16}. These values indicate practical numerical identity of the non-parametric CIF estimates.

This excellent agreement was observed across all evaluated sample-size strata, including very small samples. Thus, the CIF point estimator itself can be considered successfully validated over the full tested range.

8.6 Variance and confidence interval components for CIFs

In contrast to the CIF point estimates, agreement for CIF standard errors was not uniformly strong across all sample sizes. Across all 3184 common file × group × cause curves, the maximum absolute difference in the curve-wise standard error was 1.00, the 95th percentile of the curve-wise maximum absolute SE difference was 0.359, and the median curve-wise maximum absolute SE difference was 0.013.

Sample-size-stratified evaluation showed that these discrepancies were concentrated mainly in the smallest samples. For n = 5, the median curve-wise maximum absolute SE difference was 0.228, and 89.7% of curves showed a maximum SE difference greater than 0.1. For n = 10, the corresponding values were 0.086 and 48.9%, and for n = 20 they were 0.044 and 22.3%.

Agreement improved substantially with increasing sample size. At n = 100, the median curve-wise maximum absolute SE difference decreased to 0.0089, with only 2.0% of curves exceeding 0.1. At n = 1000 and n = 5000, the corresponding median differences were 0.0010 and 0.00027, respectively, and only 0.5% of curves exceeded 0.1.

Thus, the CIF point estimates were essentially identical across all sample sizes, whereas the variance and confidence interval components were clearly more sensitive in very small samples and became highly concordant in moderate and large samples.

8.7 Gray's test

Gray's test also showed very good agreement with R. Across 1721 comparable file × cause analyses, the maximum absolute difference in the Gray test statistic was approximately 5.8×10^{-11}. The maximum absolute difference in the Gray p-value was approximately 5.0×10^{-7}, with the 95th percentile of the absolute p-value difference remaining of the same order of magnitude.

These differences are fully compatible with negligible floating-point and numerical approximation effects. Gray's test can therefore be regarded as successfully validated for routine analytical use.

8.8 Fine & Gray regression

Fine & Gray regression showed a heterogeneous validation pattern. Across 1721 comparable file × cause analyses, the median absolute coefficient difference was 8.41×10^{-8}, the median absolute standard-error difference was 4.51×10^{-5}, the median absolute z-value difference was 1.28×10^{-3}, and the median absolute p-value difference was 1.34×10^{-5}. These values indicate very good agreement in the majority of analyses.

However, the overall distribution contained clear outliers. The maximum absolute coefficient difference was 5.000025, the 95th percentile of the absolute coefficient difference was 4.000020, the maximum absolute z-value difference was 9.237883, and the maximum absolute p-value difference was 0.158. Differences for the derived subdistribution hazard ratios and confidence intervals were correspondingly large in these outlying cases, reflecting instability of exponentiated coefficients when the underlying regression estimates diverged.

Taken together, these results indicate that Fine & Gray regression agreed very well with R in the bulk of analyses, but not uniformly across the full tested range. The large outliers were not random; they were concentrated in small-sample and sparse-event settings.

8.9 Sample-size-stratified interpretation of Fine & Gray validation

Stratified evaluation by sample size showed that Fine & Gray agreement depended strongly on the amount of available information. For n = 5, the median absolute coefficient difference was 3.000017, the 95th percentile of the absolute coefficient difference was 4.000022, and 55.2% of analyses showed an absolute coefficient difference greater than 0.1. For n = 10, the median absolute coefficient difference was already near zero (4.17×10^{-10}), but the upper tail remained pronounced, with a 95th percentile of 4.000022 and 31.0% of analyses exceeding 0.1. For n = 20, the 95th percentile remained at 4.000021, and 10.7% of analyses still exceeded 0.1.

Agreement improved markedly in moderate samples. At n = 100, the median absolute coefficient difference was 1.06×10^{-7}, the 95th percentile decreased to 0.004596, and only 2.68% of analyses exceeded 0.1. At n = 1000, the 95th percentile was 0.000571 and no analysis exceeded 0.1. At n = 5000, the 95th percentile decreased further to 0.000148, again with 0% of analyses exceeding 0.1.

A similar pattern was seen for p-values. The proportion of analyses with an absolute p-value difference greater than 0.01 was 2.4% at n = 5, 3.2% at n = 10, 3.8% at n = 20, 1.3% at n = 100, 0% at n = 1000, and 0.3% at n = 5000.

This pattern supports a stratified interpretation: Fine & Gray regression showed unstable behavior in very small samples, acceptable but not yet uniformly excellent agreement in smaller moderate samples, and very strong agreement from moderate-to-large sample sizes onward, particularly for n ≥ 100 and especially for n ≥ 1000.

8.10 Conclusion

The Sigma competing-risk module showed excellent agreement with R for CIF point estimates and very good agreement for Gray's test. Fine & Gray regression showed strong agreement in moderate and large samples, while the largest discrepancies were concentrated in very small sample sizes and sparse-event scenarios. Overall, the validation supports the use of Sigma for CIF estimation and Gray's test

across the tested range, and supports Fine & Gray regression in analyses with adequate sample size and event information, while caution is warranted in extreme small-sample settings.

Validation Report for Sigma Statistics from GBM Consulting (07th April 2026)

GBM Consulting
Dr. med. Mustafa Gerçek
Jägerstraße 101A
47228 Duisburg
Deutschland