PharmaSUG China 2025 - Paper SA-110

Statistical Methods and Programming Practice of Study-size Adjusted Risk Difference in Safety Pooling

Jun Zhai, Johnson & Johnson

ABSTRACT

In recent years, health authorities like FDA have laid increasing emphasis on appropriate statistical approaches for integrating safety data across trials be considered to account for differences in trial design, e.g. study-size adjustment in risk difference/risk ratio/odds ratio.

This paper will briefly introduce the rationale behind, as well as some recommended statistical methods and programming practice in handling integrated design applied with study-size adjusted risk and risk difference in safety pooling studies.

INTRODUCTION

When presenting summary metrics (e.g. risk difference, risk ration, odds ratio) during integrating multiple clinical studies, using crude pooling methods, i.e. pooled data directly without stratifying or weighting based on individual study, may lead to a statistics phenomenon called Simpson's paradox (Chuang-Stein and Beltangady 2011), and cause potential misinterpretations of data.

Simpson's paradox depicts a possible scenario that data trends appeared in individual studies may disappear or reverse after integration. An example in Table 1 clearly illustrates that the trend with OR > 1 appears in each individual studies turn to OR < 1 after crude pooling (Nilsson 2024).

Study	**Placebo n/N (%)**	**Treated n/N (%)**	**Odds Ratio**
1	44/6127 (0.72)	88/10324 (0.85)	1.2
2	18/2643 (0.68)	30/2650 (1.13)	1.7
3	82/2936 (2.79)	54/1483 (3.64)	1.3
crude pooling	144/11706 (1.23)	172/14457 (1.19)	0.97
study-size adjusted pooling	adjusted % = 1.05	adjusted % = 1.36	Mantel-Haenszel OR = 1.31

Table 1. Example of Simpson's Paradox

To avoid Simpson's paradox, it's widely recommended to apply study-size adjustments or stratified by trials when integrating multiple clinical trials, such as in Integrated Summary of Safety (ISS) and Summary of Clinical Safety (SCS) from FDA guidance in 2018 and PHUSE white paper in 2017.

However, this statistical practice may not yet have been fully adopted by sponsors in safety pooling submissions, nor had the FDA made it a mandatory requirement during the review process over the past decades. Until as of late, FDA has placed a renewed emphasis on Simpson's paradox in 2024 webinar: "Integrated Safety Analyses in Drug Marketing Applications: Avoiding Common Mistakes", and begins to widely request sponsors to consider study-size adjustment statistical design in reporting adverse events when submitting ISS and SCS.

STATISTICAL METHODS

Assume there are I studies, for each individual study i, the estimated rate denoted as $\widehat{r_{ti}}$ for each treatment group t, where $t = 0$ for placebo control group, 1 for active treatment group. If there's additional treatment group, let $t = 2, 3, \ldots$

For proportion,

$$\widehat{r_{ti}} = \frac{n_{ti}}{\mathrm{N}_{ti}}$$

where n is the number of participants with events from each group; N is the total number of participants from each group.

For incidence per 100 subject-years of follow-up,

$$\widehat{r_{ti}} = 100 * \frac{n_{ti}}{T_{ti}}$$

where n is the number of participants with events or number of events from each group; T is the total subject-years of follow-up from each group.

And the risk difference for each study is defined by

$$\hat{d}_i = \widehat{r_{1i}} - \widehat{r_{0i}}$$

WEIGHT ADJUSTMENTS

For the calculation of adjusted rate and risk difference, weight w_i will be applied to account for each individual study. The estimated adjusted rate can be calculated as:

$$r_{adj} = \frac{\sum w_i \cdot \widehat{r_{ti}}}{\mathrm{W}}, \qquad W = \sum w_i.$$

And confidence intervals for adjusted rate is:

$$\frac{\sum w_i \cdot \widehat{r_{ti}}}{W} \pm z_{\alpha/2} \cdot \frac{1}{\mathrm{W}} \cdot \sqrt{\sum w_i^2 \cdot Var(\widehat{r_{ti}})}, where\ t = 0,1.$$

The adjusted risk difference and its confidence interval is defined by

$$d_{adj} = \frac{\sum w_i \cdot \hat{d}_i}{W}$$

$$(1-\alpha)100\%CI = \frac{\sum w_i \cdot \hat{d}_i}{\mathrm{W}} \pm z_{\alpha/2} \cdot \frac{1}{W} \cdot \sqrt{\sum w_i^2 \cdot [Var(\widehat{r_{0i}}) + Var(\widehat{r_{1i}})]}$$

Chuang-Stein and Beltangady (2011) suggested three approaches of weight choice for risk and risk difference.

Study-size based method

This method weighs the observed proportion in a study by the percentage of subjects in that study among the pooled population, which means:

$$w_i = N_i$$

Inverse variance method

This method sets weight as the inverse of the sample variance of risk difference.

$$w_i = Var\left(\widehat{d_i}\right)^{-1}$$

Cochran-Mantel-Haenszel (CMH) method

CMH weights can be defined by:

$$w_i = \frac{N_{0i}N_{1i}}{N_i}$$

Crowe et al (2016) recommended using study-size based weights and referred this kind of standardization as "study size-adjusted", due to its naturally easier to interpret and communicate. When the treatment allocation ratio is equal across all studies, study-size based weights are approximately the same as CMH method weights.

VARIANCE

For proportion, the variance can be defined by N total number of participants, and n number of participants with events.

$$Var(\widehat{r_{ti}}) = \frac{\widehat{r_{ti}}(1-\widehat{r_{ti}})}{N_{ti}} = \frac{n_{ti}(N_{ti}-n_{ti})}{N_{ti}^3}$$

For incidence rate per 100 subject-years, Scosyrev and Pethe (2022) introduced several methods to calculate variance.

Poisson

When the event rate is constant over time, it follows Poisson distribution, and the variance of incidence rate per 100 subject-years is:

$$Var(\widehat{r_{ti}}) = 100^2 * \frac{n_{ti}}{T_{ti}^2}$$

Asymptotically robust Wald-type

The incidence rate can also be viewed as a ratio of two correlated sample means, an asymptotic variance estimator can be obtained by delta method:

$$Var(\widehat{r_{ti}}) = 100^2 * (\frac{\overline{n_{ti}}}{\overline{T_{ti}}})^2[\frac{Var_{ti}(n)}{\overline{n_{ti}}^2 N_{ti}} + \frac{Var_{ti}(T)}{\overline{T_{ti}}^2 N_{ti}} - 2\frac{Cov_{ti}(n,T)}{\overline{n_{ti}}\,\overline{T_{ti}} N_{ti}}]$$

Asymptotically robust regression-based

The variance of incidence risk difference can also be inferred based on a generalized linear model (GLM) with identity link and Poisson or Negative Binomial likelihood functions, with a person-time offset. The regression model can be fit using SAS PROC REG step (Scosyrev and Pethe 2022).

```
data study;
        set study;
        n_prime = n/sqrt(T);      /* n: event number */
        T_prime = sqrt(T);        /* T: total follow-up time */
        trt_prime = trt*sqrt(T); /* trt: treatment group */
        u_prime = u*sqrt(T);      /* u: covariate */
        x_to_y = x/y;
run;
proc reg data=study;
```

```
        model n_prime = T_prime trt_prime u_prime / noint hcc hccmethod=3;
run;
proc reg data=study;
        model x_to_y = trt u / noint hcc hccmethod=3;
        weight T;
run;
```

CONTINUITY CORRECTION

When the event counts are low, e.g. analysis of uncommon adverse events, the confidence interval estimators may undercover the probability. A widely-used method is to apply continuity correction to approximate normal distribution for discrete distributions.

In SAS PROC FREQ statement, users can specify CORRECT riskdiff-option or the RISKDIFFC option to apply below continuity correction for Wald asymptotic confidence limits for proportion rate and risk difference:

$$\widehat{r_{ti}} \pm (z_{\alpha/2} \cdot \sqrt{Var(\widehat{r_{ti}})} + cc), where\ cc = 1/2N_{ti}$$

$$\widehat{d_i} \pm (z_{\alpha/2} \cdot \sqrt{Var(\widehat{d_i})} + cc), where\ cc = (1/N_{1i} + 1/N_{0i})/2$$

Bright and Soulakova (2016) reviewed three continuity correction methods that are applicable for incidence risk difference estimation.

$$\widehat{d_i} \pm (z_{\alpha/2} \cdot \sqrt{Var(\widehat{d_i})} + cc)$$

Yates

$$cc = (1/T_{1i} + 1/T_{0i})/2$$

Schouten

$$cc = 1/[2\max(T_{1i}, T_{0i})]$$

Hauck-Anderson

$$cc = 1/[2\min(T_{1i}, T_{0i})]$$

Scosyrev and Pethe (2022) proposed a new approach that considered a non-informative exponential prior, with correction factors are obtained by:

$$A_{Mi} = (1/T_{1i} - 1/T_{0i}),\ B_{Mi} = (1/T_{1i}^2 + 1/T_{0i}^2)$$

And confidence interval can be calculated as:

$$(\widehat{d_i} + A_{Mi}) \pm z_{\alpha/2} \cdot \sqrt{Var(\widehat{d_i}) + B_{Mi}}$$

To sum up, when designing statistical methods for safety pooling, it's recommended to consider weight adjustments by trials, continuity correction for uncommon AE analysis, and choice of variance for risk and risk difference, based on trial design, data nature and analysis needs. This paper only introduced several common methods, meanwhile in real practice integrated statistical analysis may adapt one or more methodologies that suit the study design best.

PROGRAMMING PRACTICE

From programming perspective, the statistical design of safety pooling analysis could be miscellaneous and easily confused. Not only should we cultivate a deeper understanding of SAP and keep close collaboration with statisticians, but we also need to exercise caution and add robustness during the design and implementation of analyses.

DISTINGUISHABLE OUTPUTS

Considering there could be different statistical methodologies designed for similar mock and same series of events of interests, to decrease the possible programming issues and ease the burden of reviewers, we may design distinguishable outputs and conduct programming activities considering below practices:

- Apply same pattern of table ID for similar mock, but use suffix to distinguish the differences
- Clearly state the statistics of interests and applicable methodology in titles and footnotes
- Create separate macros for each methodology with distinguishable names
- Create catalog of outputs and applicable macros for different methodologies, to assist supporting programmers and reviewers

ROBUST MACRO DESIGN

In order to make the macros easy to utilize and maintain by different team members, it's recommended to create separate macros at root level for methodologies that vary much, or need different kinds of input data and arguments, or may require further update. For methodologies with similar overall structure and inputs, it's also feasible to wrap them into one single macro, and use optional input arguments to achieve different statistical methodologies.

For example, a catalog of macros to calculate study-size adjusted statistics in real study is designed as Table 2 shows. Separate macros were designed for layout 1, since the selection of variance and continuity correction are all different and has the needs for ongoing updates. Meanwhile for layout 2, only one single macro, with the help of different input arguments, is sufficient and clear enough, given that the calculation of crude Wald statistics are conventional and similar across categories.

	Common AE (without continuity correction)		Uncommon AE (with continuity correction)	
	Proportion	Incidence	Proportion	Incidence
Layout 1 • Crude rate • Study-size adjusted rate and risk difference	%macroA	%macroB	%macroC	%macroD
Layout 2 • Crude rate	%macroE (cc=N, …);	%macroE (cc=N, fupvar=, …);	%macroE (cc=Y, …);	%macroE (cc=Y, fupvar=, …);
Layout 3 • Study-size adjusted Relative risk	%macroF	/	/	/

Table 2 Catalog of Macro Design

Despite the diverse choice of statistical components, i.e. weight adjustments, variance and continuity correction, we can still observe a universal formula when calculating the adjusted risk and risk difference for both proportion and incidence rate.

The point estimate is a sum of weighted estimation for each individual trials. The confidence interval considers standard error and continuity correction. The adjusted estimator without continuity correction can be defined by $f(est_i)$, and adjusted estimator with continuity correction can be defined as a joint function of $f(est_i)$ and cc of selection.

$$est_{adj} = G(f(est_i), cc)$$

$$f(est_i) = \frac{\sum w_i \cdot \widehat{est_\iota}}{W} \pm z_{\alpha/2} \cdot \frac{1}{\mathrm{W}} \cdot \sqrt{\sum w_i^2 \cdot Var(\widehat{est_\iota})}$$

$$where\ G\ is\ a\ function\ of\ continuity\ correction$$

Based on the general formula, a robust structure of macro design is proposed as Diagram 1 shows.

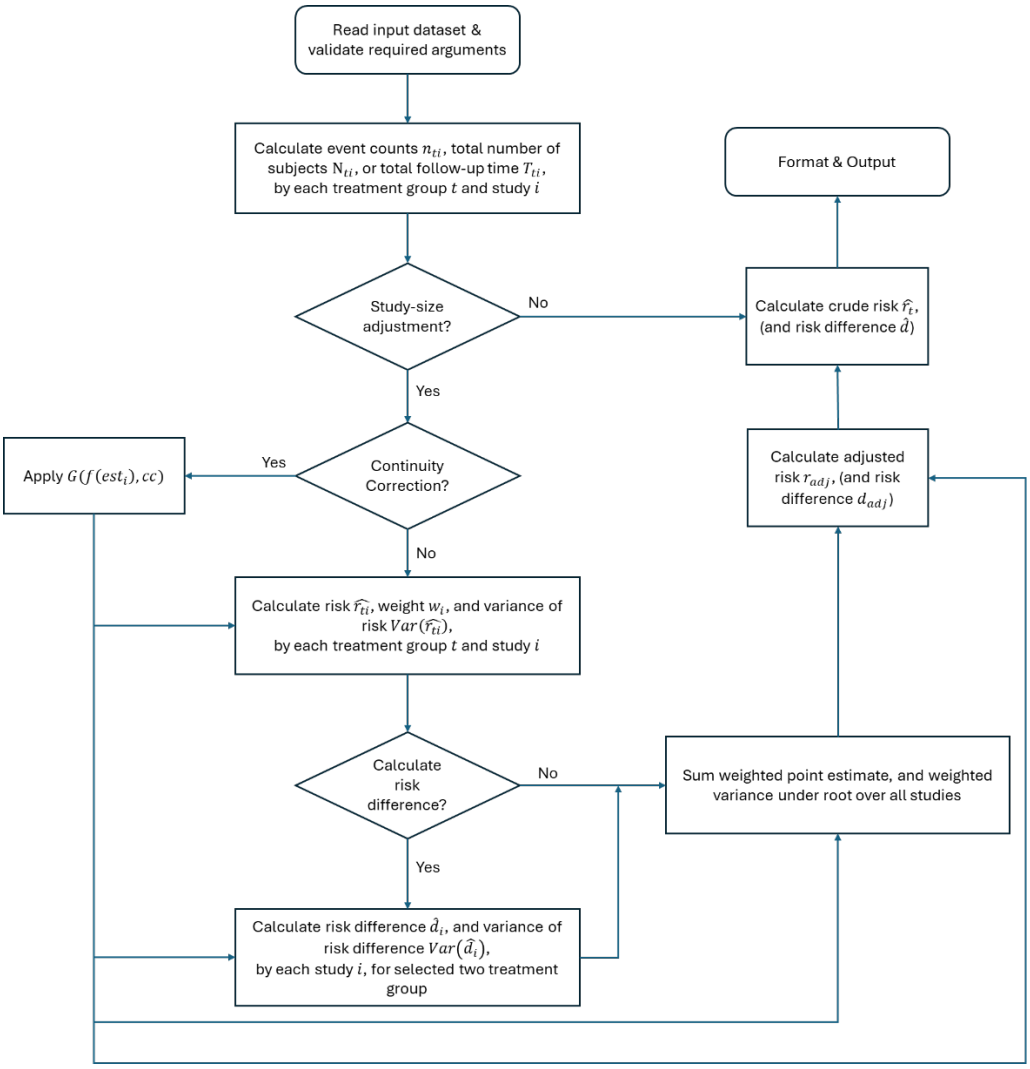

Diagram 1 Robust Macro Design

To take care of layouts with column of combined treatment groups, the macros should always summarize by both treatment group variables and column variables, especially when they are not equivalent.

It's also recommended to summarize event statistics in vertical structure, i.e. record by each treatment group and individual study, instead of adding columns of count and rate for each treatment group. Although it could be handy to use variable N0, N1, T0, T1 for calculation when there's only two treatment groups, we should also take multiple treatment groups into consideration when designing the macros.

AE BY SOC/PT

In addition to displaying one single row of adjusted risk and risk difference for subjects with one or more AE in the mock, it's also common to present the adjusted risk or risk difference by each AE system organ class (SOC) and preferred term (PT). The general idea is similar as Robust Macro DesignDiagram 1, we just need to repeat the process for AE events by each SOC and by PT respectively.

	Placebo	Active Treatment	Risk difference (95% CI)
Subjects with 1 or more adverse events (adjusted)	### (xx.x%)	### (xx.x%)	xx.x% (x.xx%, x.xx%)
95% CI	(x.xx%, x.xx%)	(x.xx%, x.xx%)	
[System organ class]	### (xx.x%)	### (xx.x%)	xx.x% (x.xx%, x.xx%)
[Preferred term]	### (xx.x%)	### (xx.x%)	xx.x% (x.xx%, x.xx%)

Table 3 Mock AE Table with Study-size Adjustment by SOC/PT

Normally for crude proportion or incidence by SOC/PT, there already exist well-established macros and templates within organization. And we may build additional functionalities based on them, which can lead to two approaches of macro design.

The first approach is to program based on input dataset: we calculate the adjusted risk and risk difference for every SOC and PT from input dataset, merge back each PT record under corresponding SOC, and then adjust the overall ordering or apply data cutoff based on mock requirements and organization standards. This approach has several underlying issues:

- When the design of methodology is complex, it may consume longer time and more server capacity than expected to compute for every SOC/PT terms from input dataset.
- If there's no available in-process standard macro to use, we need to apply data cut-off, format, and order each SOC/PT terminology by open code. It may waste extra efforts and cause potential discrepancies.

Hence, I recommend selecting the second approach when feasible, which is result-driven. We first create a crude AE table using standard macros as template shell. The ordering, formatting and cutoff are expected to be applied for each SOC/PT terms with standard macros. After that, we calculate the adjusted risk or risk difference for SOC/PT only appearing in the template, then merge the result back to the template, and finalize the table. This approach can save the efforts of unnecessary data wrangling or computation and avoid discrepancies during formatting.

DISCUSSION

USER-FRIENDLY CODE

As statistical programmers, we collaborate all the time. Therefore, it's always more important for us to write readable and easy-to-maintain code, rather than showing off skills and writing fancy chunks.

For notation and variables naming, be straightforward and distinguishable, use detailed annotations, and better keep aligned with how SAP defines. This can help reviewers, maintainers, even the author himself to quickly read through and understand codes.

If there's methodologies update or even back-and-forth revision, document each decision and pros in macro header, which may help team members or people who refer to it cultivate deeper understanding of the rationale behind.

EFFICIENT VALIDATION

The study-size adjusted risk difference macros are widely applied for AE analysis in safety pooling. On one hand, creating robust macros for different methodologies different mocks is already exhausting, on the other hand, the validation of the outputs can also be problematic. The most conservative way is always to double-program for everything. However, considering the resource and timeline in real life study, it's never a feasible choice.

Following the essence of risk-based quality control, our team eventually choose to double-program for selected important and unique mocks, apply code review and outputs cross-check from QC programmers, programming lead and statisticians. In this way, the correct methodology of choice and the quality of macros can be guaranteed to a certain extent. After that those macros can be treated as standard ones, and allowed to be directly utilized from QC side.

Despite the good validation of macros themselves, errors and issues can still happen when supporting programmers try to apply them, given the miscellaneous macros for different methodologies. It's essential that programming lead follow good practice in Distinguishable Outputs, and provide clear expectations and sufficient guidance to supporting programmers, especially those who are not familiar with SAP or pooling statistical design.

GENERALIZATION

This batch of macro can already satisfy analysis needs with same methodology and similar layouts, but still require further improvement and maintenance to achieve generalization, i.e. direct usage under different table layouts. Within our compound pooling efforts, this series of macros has been smoothly applied in ADR (adverse drug reaction) and BR (benefit-risk analysis) with limited revision. However, when utilized for complexed table layouts, it may give wrong results due to lack of consideration of combined columns and duplicate records.

When there's internal reporting standards established for study-size adjusted safety pooling analysis, this series of macros are expected to be further improved and has the potential to be generalized as global standard macros.

CONCLUSION

Health authorities, e.g. FDA, has been increasingly laid emphasis on the stratified design by study in integrated analysis, including avoid of Simpson's paradox. This paper briefly introduces the statistical design and programming practices of study-size adjustments in safety pooling analysis. It's recommended to consider weight adjustments, variance of choice and continuity correction in pooling statistical analysis plan based on the sample size and trial design of each individual trials. A general equation is summarized for study-size adjusted risk and risk difference calculation:

$$est_{adj} = G(f(est_i), cc)$$

$$f(est_i) = \frac{\sum w_i \cdot \widehat{est_\iota}}{W} \pm z_{\alpha/2} \cdot \frac{1}{\mathrm{W}} \cdot \sqrt{\sum w_i^2 \cdot Var(\widehat{est_\iota})}$$

$$where\ G\ is\ a\ function\ of\ continuity\ correction$$

From programming perspective, we need to implement good practice in designing distinguishable outputs and creating robust macros for study-size adjustments calculations. A general flowchart is proposed for a robust macro structure in section Robust Macro Design.

This paper also discusses some trade-off situations and room for improvement, such as how to perform efficient and qualified validation for complex statistical calculations, and how to improve the macros to make it more generalizable and standardized.

Overall speaking, this paper brings the topic of an increasingly essential health authorities review emphasis on study-size adjustment in safety pooling's statistical design, and share the fundamental statistical considerations and programming good practice.

REFERENCES

Chuang-Stein, C., and M. Beltangady. 2011. “Reporting Cumulative Proportion of Subjects with an Adverse Event Based on Data from Multiple Studies.” Pharmaceutical Statistics 10(1):3-7.

Crowe, B., C. Chuang-Stein, S. Lettis, and A. Brueckner. 2016. “Reporting Adverse Drug Reactions in Product Labels”. Therapeutic Innovation & Regulatory Science 50(4):455-463.

Scosyrev, E., and A. Pethe. 2022. “Confidence Intervals for Exposure-adjusted Rate Difference in Randomized Trials”. Pharmaceutical Statistics 21(1):103-121.

Bright C B., and J. N Soulakova. 2016. “Evaluating the Performance of Simultaneous Stepwise Confidence Intervals for the Difference Between Two Poisson rates”. Commun Stat Simul Comput. 45(7):2338-2361.

PHUSE Analysis & Display White Papers Project Team. 2017. “Analysis and Displays Associated with Adverse Events: Focus on Adverse Events in Phase 2-4 Clinical Trials and Integrated Summary Documents, Version 1.0.”

U.S. Food and Drug Administration. 2018. “Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products”. Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products | FDA

Nilsson, M.. 2024. “Common Mistakes When Pooling Clinical Trial Safety Data”. Integrated Safety Analyses in Drug Marketing Applications: Avoiding Common Mistakes - 03/07/2024 | FDA

Gubitz, J.. 2024. “Study-Size Adjusted Percentages in Integrated Adverse Event Displays”. https://www.rhoworld.com/study-size-adjusted-percentages-in-integrated-adverse-event-displays/

ACKNOWLEDGMENTS

I would like to give sincere thanks to my programming lead Angela Hutchinson and Sungmi Brown, as well as Shu Li, Chaohui Yuan and Ming-chun Hsu from statistician team, for their insightful inputs and support during the creation and validation of study-size adjusted macros in SCS efforts.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jun Zhai
Johnson & Johnson
jzhai@its.jnj.com