

Risk Estimation Analytical Methods

Alan Hochberg, ProSanos Corporation

19 March 2009

Overview

This document describes *risk estimation*, the process of calculating an Incidence Rate Ratio (IRR) which compares the rates of a particular adverse event between two drugs, known as the *target* and the *comparator*. Some statistics associated with the IRR, such as confidence limits, are also computed. Propensity score adjustment is employed to minimize the impact of confounding clinical and demographic factors on the calculated IRR.

The input to the process is a data frame of subject eras. These describe periods of time when a subject was exposed to either the target or comparator drug. An *outcome* variable for each era describes whether or not the adverse event of interest occurred within the era. (At the user's option, a *surveillance window* of additional time may be added to the era, during which the subject is considered at risk. An event during this window will be counted.) *Outcome* is a binary variable; only one event within an era is recorded¹. A *follow-up time* (in days) indicates the length of the era. The follow-up time always corresponds to the amount of time the subject is at risk for the outcome event. Era descriptions also contain the subject's age at the start of the era; their gender; and the calendar year at the start of the era (represented as [year – 2000] to avoid numerical

¹ The outcome is modeled by Poisson regression. Note that the limitation to a single event per era causes the data to deviate from a true Poisson formulation. As long as the event rate is low enough that the probability of two or more events per era is small, the error introduced by this approximation is negligible.

overflow when exponentiating). Finally, eras are described by a sparse list of thousands of binary covariates, each of which corresponds to the administration of a drug (in addition to the target or comparator), or the occurrence of a medical condition which is coded as term in the Medical Dictionary for Regulatory Activities (MedDRA®)².

Certain aspects in the design of the risk estimation calculation are fixed, and cannot be controlled by the SAEfetyWorks software user. For instance, a retrospective cohort study design is always employed, and sampling within strata is used for propensity score adjustment. Other aspects of a given risk estimation are controlled by user-settable parameters. These are described below as data inputs in the various steps where they appear.

The computations for risk estimation are performed in C++. Routines for mathematical and statistical functions are components of the Numerical Algorithms Group (NAG) library (Numerical Algorithms Group, Oxford, UK), 64-bit version CLW6A08DA. Specifically, the routine **nag_glm_binomial** is used for the propensity score logistic regression model; **nag_random_continuous_uniform_ab** is used for sampling with replacement; and **nag_glm_poisson** is used for the outcome model. Other NAG routines are used for calculation of probability distributions and for matrix algebra.

For purposes of mathematical description, the steps of risk estimation are as follows:

- Target and comparator cohort selection

² MedDRA® is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA).

- Outcome definition
- Exposure covariate selection
- Exposure covariate evaluation
- Propensity score model creation
- Propensity score model assessment
- Outcome model covariate selection
- Outcome model creation without interactions
- Outcome model creation with interactions
- Risk estimation result calculation

Target and comparator cohort selection

Input data:

Data frame \mathcal{F} containing N_s subject eras. Each era i is described by:

t_i = the duration of the era in days

a_i = subject age in years at the start of the era

g_i = subject gender, 1=male, 0=female

r_i = calendar date of the start of the era

x_{ij} = for a set of N_d drugs \mathcal{D} , $j=1..N_d$, $x_{ij}=1$ if the subject received drug j during era i , 0 otherwise. For a set of N_c medical conditions \mathcal{C} , $j=N_d+1..N_d+N_c$, $x_{ij}=1$ if the subject experienced medical condition j during era i , 0 otherwise.

A given subject may appear in more than one era if they have more than one non-contiguous period of exposure to the target or comparator drug. For purposes of these calculations, multiple appearances of a subject are treated independently. For conciseness, we will refer to a data frame of “subjects”, but this term should be taken to mean “subject eras”.

Process steps:

Conditions are set on the elements of each row of the data frame, in order to define a *target* and a *comparator* cohort. At least one of the conditions involves selection on one

or more of the drug covariates x_{ij} where $j=1..N_{\mathcal{D}}$. The process of cohort selection can be very simple, for instance, selection on the basis of a single binary drug covariate for the target and a different drug covariate for the comparator, or it can be complex. The software accommodates selections based on Boolean logic involving a number of covariates, and also involving ranges of the continuous variables for *age* and *year*. Timing relationships for the appearances of drugs and conditions may also be involved. Complete descriptions of the cohort selection processes are given in the Users Guide.

Output data:

A data frame \mathcal{C} of cohorts, consisting of the subset of rows from data frame \mathcal{F} which meet the mutually-exclusive target or comparator cohort criteria. The content of each row of \mathcal{C} is the same as that of the corresponding row of \mathcal{F} , except for the presence of an additional variable:

$\varepsilon_i = 1$ if subject era i meets the user-specified criteria for inclusion in the target cohort, 0 if the subject era meets the criteria for inclusion in the comparator cohort.

Outcome definition

Input data:

Data frame \mathcal{C} containing N_c subject eras.

Process steps:

Conditions are set on the elements of each row of the data frame. At least one of the conditions involves selection on one or more of the condition covariates x_{ij} where $j = N_c + 1 \dots N_c + N_d$. As with cohort selection, outcome definition can be as simple as defining the outcome on the basis of a single binary condition variable, or it can be complex, involving many variables and their time-dependencies. Complete information is contained in the Users Guide.

Output data:

The data frame \mathcal{D} of all subject eras is augmented to include an additional variable:

$y_i = 1$ if subject era i meets the user-specified criteria for the outcome of interest, 0 otherwise.

Exposure covariate selection

Input data:

The data frame \mathcal{D} .

Process steps:

Let the target cohort \mathcal{F} consist of all subject eras with $\varepsilon_i = 1$. This will be compared with a reference cohort \mathcal{R} which, at the user's option, may be either:

- 1) The subset of \mathcal{C} with $\varepsilon_i = 0$ (“Comparator cohort”)
- 2) All patients in \mathcal{F} who were exposed to the Comparator drug, regardless of whether or not they fully met the Comparator Cohort definition criteria. (“Entire background”)

A crudely matched cohort for comparison \mathcal{M} is created as follows: For each subject era in \mathcal{F} , a matching subject era from \mathcal{R} is selected, such that the genders g of the eras are identical, and the calendar dates r match at either the “year” level, or the “year/month” level, as specified by the user. If at least one matching subject era is available, then the era from \mathcal{F} and the era from \mathcal{R} are added to \mathcal{M} . If more than one matching subject era is available and the target cohort is >5000 eras, a single matching era is chosen at random.

For target cohorts of ≤ 5000 eras, three-to-one matching is used. (Three comparator eras to one target era.)

For every candidate covariate $v \in \{x_j\}$, odds ratios are calculated with respect to exposure, outcome, gender, and every other candidate covariate, where the formula for the odds ratio is:

$$OR_{\mathcal{O}}(c,v) = AD / BC,$$

$$A = \sum_{i \in M} c_i v_i$$

$$B = \sum_{i \in M} c_i (1 - v_i)$$

$$C = \sum_{i \in M} (1 - c_i) v_i$$

$$D = \sum_{i \in M} (1 - c_i) (1 - v_i)$$

The odds ratio with respect to exposure is $OR_{\mathcal{O}}(\varepsilon, v)$; the odds ratio with respect to outcome is $OR_{\mathcal{O}}(y, v)$; The odds ratio with respect to gender is $OR_{\mathcal{O}}(g, v)$, and the odds ratio with respect to each covariate x_j is $OR_{\mathcal{O}}(x_j, v)$.

Wald confidence limits are estimated for each of the above odds ratios by calculating the approximate standard deviation of the log odds ratio:

$$\hat{\sigma}_{LOR} = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

Then the upper and lower 95% confidence limits for the OR are:

$$OR_+ = OR_M(c, v)e^{+1.96\hat{\sigma}_{LOR}}$$

and

$$OR_- = OR_M(c, v)e^{-1.96\hat{\sigma}_{LOR}}$$

respectively.

The user is presented with a set of recommended propensity score covariates from v which meet the criterion $OR_{\mathcal{M}}(\epsilon, v) > \tau_R$, where τ_R is a user-selectable parameter with a default value of 2.0. The user may also manually add covariates from a longer list of candidate covariates which meet the criterion $OR_{\mathcal{M}}(\epsilon, v) > \tau_C$, where τ_C is a user-selectable parameter with a default value of 1.1. Covariates in either of these lists must satisfy the “information content” criteria:

$$\min\left[\sum v, \sum (1 - v)\right] \geq N_{\min}$$

Where N_{\min} is a user-defined parameter with a default value of 10. A warning message is issued if $OR_{\mathcal{O}_k}(x_j, v) > \tau_m$, or $OR_{\mathcal{O}_k}(g, v) > \tau_m$, the mutual odds ratio parameter, where τ_m is a user-defined parameter with a default value of 40.0.

Age, Gender, and year of exposure (a , g , and r) are always selected for inclusion in the propensity score model as covariates.

Output data:

The data frame \mathcal{D} .

A set of covariates \mathcal{V} to be included in the propensity-score model.

Note: the approximately-matched cohort \mathcal{O}_k is not used subsequent to this step.

Propensity score model creation

Input data:

The data frame \mathcal{D} .

A set of covariates \mathcal{V} to be included in the propensity-score model.

Process steps:

For the subjects in \mathcal{D} , let the vector p be the probability of membership in the target cohort:

$$p = \Pr(\varepsilon = 1)$$

The propensity score model is a logistic regression model of the form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}$$

Where \mathbf{x} is a matrix. Each row of \mathbf{x} represents a subject era, and the columns are the covariates which describe the subject era:

$$\mathbf{x} = [a \quad g \quad r \quad x_j \in \mathcal{V}]$$

All variables selected by the user are included in the propensity score model. Unlike the outcome model described below, there is no stepwise selection process.

Model coefficients and standard errors are provided to the user. The model-predicted value of $s = \log[p/(1-p)]$ for each subject is their *propensity score*.

A propensity-score balanced set of subject eras \mathfrak{B} , is created as follows: At the user's option, subjects may be eliminated from \mathfrak{B} if they fall outside the propensity score range of common support, i.e. the range of propensity scores that is common to both the target and comparator cohorts. Subjects may also optionally be eliminated from the upper tail, or from both tails of the propensity score distribution, based on a specified percentile of the propensity score frequency distribution. The range of propensity scores for all subjects is divided into K quantiles. The number of propensity strata K is user-selectable with a default of 10. Let \mathfrak{S} represent the cohort within \mathfrak{D} which is larger, either the target cohort or the comparator cohort. Let \mathfrak{C} represent the smaller cohort. All subjects from \mathfrak{C} are placed in \mathfrak{B} . Then, for each of the K strata, a sample of subjects from \mathfrak{S} is chosen by sampling with replacement from the subjects in the K th propensity score stratum of \mathfrak{S} , and included in \mathfrak{B} . The size of the sample is chosen so that the numbers of subjects in each propensity score stratum from \mathfrak{S} and from \mathfrak{C} are equal.

A preliminary power calculation is presented to the user as follows³:

λ_x = The number of subject-eras having $\varepsilon=0$ and $y=1$.

r_x = The number of subject-eras having $\varepsilon=0$.

r_y = The number of subject-eras having $\varepsilon=1$.

$$d = r_y / r_x$$

$z_{1-\alpha}$ = the x-value for the cumulative normal distribution at $p=0.95$, which is 1.645

$z_{1-\beta}$ = the x-value for the cumulative normal distribution at $p=0.8$, which is 0.842.

If the user requests a calculation at 90% power instead of 80%, $z_{1-\beta}=1.282$.

$$K = z_{1-\alpha} + z_{1-\beta}$$

$$a = \lambda_x$$

³ Thode HC, Power and sample size requirements for tests of differences between two Poisson rates, *Statistician* 1997;46:227–230.

$$b = -\left(2\lambda_x + \frac{K^2}{d}\right)$$
$$c = \lambda_x - K^2$$
$$\Gamma = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

Then Γ is reported as an estimate of the minimum IRR which is distinguishable from 1.0 with 80% power at $p=0.05$. A refined power calculation is performed at the end of the entire risk estimation process.

Output data:

The balanced set of subject eras \mathfrak{B} .

The propensity score vector \mathbf{s} .

Propensity score model assessment

Input data:

The balanced set of subject eras \mathcal{B} .

Process steps:

For each covariate in \mathbf{x} , a dimensionless measure of covariate balance between -1 and 1 is calculated as follows:

$$\Delta = \frac{\bar{v}_t - \bar{v}_c}{\bar{v}_t + \bar{v}_c}$$

Where v_t is the value of the variable v (a column in \mathbf{x}) over the subjects in the target cohort (defined by $\varepsilon=1$), and v_c is the value of the variable v (a column in \mathbf{x}) over the subjects in the comparator cohort (defined by $\varepsilon=0$).

For each variable, the value of Δ is presented to the user, along with counts and percentages of subjects with that variable=1 in both cohorts.

Output data:

The balanced set of subject eras \mathcal{B} .

Outcome model covariate selection

Input data:

The balanced set of subject eras \mathfrak{B} .

Process steps:

For every candidate covariate $v \in \{x_j\}$, odds ratios are calculated with respect to exposure, outcome, gender, and each other candidate covariate, where the formula for the odds ratio is:

$$OR_{\mathfrak{B}}(\mathbf{c}, \mathbf{v}) = AD / BC,$$

$$A = \sum_{i \in \mathfrak{B}} c_i v_i$$

$$B = \sum_{i \in \mathfrak{B}} c_i (1 - v_i)$$

$$C = \sum_{i \in \mathfrak{B}} (1 - c_i) v_i$$

$$D = \sum_{i \in \mathfrak{B}} (1 - c_i) (1 - v_i)$$

The odds ratio with respect to exposure is $OR_{\beta}(\epsilon, \nu)$; the odds ratio with respect to outcome is $OR_{\beta}(y, \nu)$; The odds ratio with respect to gender is $OR_{\beta}(g, \nu)$, and the odds ratio with respect to each covariate x_j is $OR_{\beta}(x_j, \nu)$.

Wald confidence limits are estimated for each of the above odds ratios by calculating the approximate standard deviation of the log odds ratio:

$$\hat{\sigma}_{LOR} = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

Then the upper and lower 95% confidence limits for the OR are:

$$OR_{+} = OR_{\beta}(c, \nu)e^{+1.96\hat{\sigma}_{LOR}}$$

and

$$OR_{-} = OR_{\beta}(c, \nu)e^{-1.96\hat{\sigma}_{LOR}}$$

respectively.

The user is presented with a set of recommended propensity score covariates from ν which meet the criterion $OR_{\beta}(\epsilon, \nu) > \tau_R$, where τ_R is a user-selectable parameter with a

default value of 2.0. The user may also manually add covariates from a longer list of candidate covariates which meet the criterion $OR_{\mathcal{C}}(\varepsilon, \nu) > \tau_C$, where τ_C is a user-selectable parameter with a default value of 1.1. Covariates in either of these lists must satisfy the “information content” criteria:

$$\min\left[\sum \nu, \sum (1 - \nu)\right] > N_{\min}$$

Where N_{\min} is a user-defined parameter with a default value of 10. A warning message is issued if $OR_{\mathcal{B}}(x_j, \nu) > \tau_m$, or $OR_{\mathcal{B}}(g, \nu) > \tau_m$, the mutual odds ratio parameter, where τ_m is a user-defined parameter with a default value of 40.0. Covariates are further excluded if $OR_{\mathcal{B}}(\varepsilon, \nu) > \tau_\varepsilon$, the mutual odds ratio parameter, where τ_ε is a user-defined parameter with a default value of 10.0. This latter check prevents the inclusion of covariates that would compete with exposure for explanatory power, and therefore reduce the influence of drug exposure in the model.

Age, Gender, and year of exposure (a , g , and r) are included in the outcome model unless the user chooses an option to remove one or more of them.

Output data:

The balanced set of subject eras \mathcal{B} .

A set of covariates \mathcal{C} to be included in the outcome model.

Outcome model creation without interactions

Input data:

The propensity-score balanced data frame \mathfrak{B} .

A set of covariates \mathfrak{W} to be included in the outcome model.

The propensity score vector \mathbf{s} .

Process steps:

For the subjects in \mathfrak{B} , The outcome model is a Poisson regression model of the form

$$\log(\bar{y}) = \log(\mathbf{t}) + \beta_o + \boldsymbol{\beta}^T \mathbf{x}$$

At the user's option, the propensity score may be included in the model:

$$\log(\bar{y}) = \log(\mathbf{t}) + \beta_o + \beta_s \mathbf{s} + \boldsymbol{\beta}^T \mathbf{x}$$

Where \mathbf{t} is the time-at-risk variable (duration of the subject era) contained in \mathfrak{D} , and \mathbf{x} is a matrix, where each row represents a subject, and the columns represent the covariates:

$$\mathbf{x} = [a \quad g \quad r \quad x_j \in \mathfrak{W}]$$

Note that since, in this application, y is limited to values of 0 or 1, the data does not truly conform to the specification of a Poisson regression model. However, since generally for safety outcomes $y \ll 1$, the error introduced by this approximation is small.

Models are constructed using forward stepwise variable selection. The chi-squared test for reduction of deviance at $p=0.2$ is used as the stepwise variable inclusion criterion.

The total number of covariates accepted into the model is limited to a maximum specified by the user with a default value of 50, and is further limited by a constraint on the number of exposures ($\sum \varepsilon$) per variable included in the model. The default for the *exposures per variable* constraint is 50.

Model coefficients and standard errors are provided to the user.

Output data:

The balanced set of subject eras \mathfrak{B} .

A set of covariates \mathfrak{W} that met selection criteria for the outcome model.

The outcome model without interaction terms, embodied in β and associated statistics.

Outcome model creation with interactions

Input data:

The balanced set of subject eras \mathfrak{B} .

A set of covariates \mathfrak{W} that met selection criteria for the outcome model.

The propensity score vector \mathbf{s} .

Process steps:

For the subjects in \mathfrak{B} , The outcome model with interactions is a Poisson regression model of the form

$$\log(\bar{y}) = \log(\mathbf{t}) + \beta_o + \boldsymbol{\beta}^T \mathbf{x} + \boldsymbol{\gamma}^T \mathbf{w}$$

At the user's option, the propensity score can be included in the model:

$$\log(\bar{y}) = \log(\mathbf{t}) + \beta_o + \beta_s \mathbf{s} + \boldsymbol{\beta}^T \mathbf{x} + \boldsymbol{\gamma}^T \mathbf{w}$$

Where \mathbf{w} is a matrix containing all possible interactions between variables in \mathbf{x} :

$$\mathbf{w} = [(r - \bar{r})(s - \bar{s}) \quad r \in \mathbf{x}, s \in \mathbf{x}, r \neq s]$$

Interaction terms from \mathbf{w} are added to the original model using forward stepwise selection. In order to be added to the model, an interaction term must meet three criteria:

- 1) The absolute value of the t-statistic for the interaction term (ratio of its coefficient to its standard error) must be >1.96 .
- 2) The chi-squared value for the reduction in residual deviance due to entry of the term into the model must be statistically significant at $p=0.05 / c$, where c is the number of columns in \mathbf{w} .
- 3) Introduction of the term into the model must cause at least a 10% change in $\log(\text{IRR})$ from the formula below.

Output Data:

The outcome model, embodied in $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, along with \mathbf{w} and the variance-covariance matrix \mathbf{V} for the coefficients $\boldsymbol{\gamma}$.

Risk estimation result calculation

Input Data:

The outcome model, embodied in β , γ , along with \mathbf{w} and the variance-covariance matrix \mathbf{V} for the coefficients γ .

Process Steps:

The final IRR is calculated as:

$$IRR = \exp(\beta_{total}) = \exp\left(\beta_{\epsilon} + \sum_{\substack{\text{selected} \\ \text{interactions}}} \bar{w}_i \gamma_i\right)$$

Model coefficients and standard errors are provided to the user.

From the regression modeling calculations, we have \mathbf{V} , the variance-covariance matrix for the regression coefficients γ , and V_{ϵ} which is the variance for the exposure coefficient. A variance for exposure with interactions is calculated as:

$$\sigma^2 = V_{\epsilon} + \sum_{\text{interactions } i} \sum_{\text{interactions } j} \bar{w}_i \bar{w}_j V_{ij}$$

A variance inflation factor ρ for sampling with replacement is calculated from the formulas in the Appendix. The confidence limits on the IRR are then

$$\text{UCL} = \exp(\beta_{\text{total}} + 1.96\tau)$$

$$\text{LCL} = \exp(\beta_{\text{total}} - 1.96\tau)$$

Where

$$\tau = \sqrt{\rho \cdot \sigma^2}$$

Two power calculations are then performed. The following quantities are computed:

- 1) The power to detect the point-estimate IRR at an alpha of $\alpha=0.05$, given the variance structure of the estimate.
- 2) The minimum IRR which can be distinguished from the null hypothesis $\text{IRR}=1.0$ with 80% power and an alpha of $\alpha=0.05$.

For two-sided tests, our allowance for α on each side is gives $\alpha=0.025$, and $z_{\alpha} = 1.96$.

Letting $v = \frac{-\hat{\beta}_{\text{total}}}{\tau}$, the power to distinguish $\exp(\beta_{\text{total}})$ from 1.0 is:

$$P = 1 - \{\Psi(v + 1.96) - \Psi(v - 1.96)\}$$

where Ψ is the cumulative normal probability function. Setting $P=0.8$ and solving for v simply gives $v=-2.801$. So the value of β that is distinguishable from zero, corresponding to an IRR distinguishable from 1, is:

$$\beta_{crit} = \pm 2.801\tau$$

Where we use the positive sign if $\beta_{total} > 0$, and the negative sign if $\beta_{total} < 0$. From this we have:

$$IRR_{crit} = \exp(\beta_{crit}).$$

The power P and the value of IRR_{crit} are reported to the user. The above formulation of the model and power calculation works for a single interaction as well as multiple interactions, and works for the case of no interactions if the sums are taken to be zero in this case.

Finally, the absolute risk for the comparator group is calculated as:

$$r_{AC} = \frac{\sum y(1 - \varepsilon)}{\sum (1 - \varepsilon)}$$

And the absolute risk difference attributable to treatment is then calculated as:

$$r_{AD} = (\exp(\beta_{total}) - 1)r_{AC}$$

Output data:

The balanced set of subject eras \mathfrak{B} .

An expanded set of covariates \mathfrak{W}^* that includes interaction terms.

The outcome model, embodied in β, γ .

The IRR and confidence limits.

Power to detect the IRR.

Critical IRR detectable at 80% power.

Absolute risk for comparator group.

Absolute risk difference attributable to treatment.

Appendix: Sampling with Replacement Variance Inflation Calculation

Below we refer without loss of generality to the larger cohort as the “target-exposed cohort”, and the smaller cohort as the “comparator-exposed cohort”. Let the total number of target-exposed subjects in stratum i be T_i , $i=1..K$, and let the number of comparator-exposed subjects in stratum i be C_i .

For the target-exposed group, the number of adverse outcome events t_i in each stratum i is modeled as:

$$t_i = B(T_i, p_i)$$

Where $B(N,p)$ denotes a binomial-distributed random variable. Similarly for the comparator-exposed group,

$$c_i = B(C_i, q_i)$$

We can estimate the probability of an outcome event in the target population as:

$$\hat{p}_i = t_i/T_i$$

where t_i is the number of subjects in the target stratum who have the adverse event outcome of interest. And for the comparator cohort, we can estimate:

$$\hat{q}_i = c_i / C_i$$

where c_i is the number of subjects in the comparator stratum who have the adverse event outcome of interest.

In the case of sampling *without* replacement, and define a target-exposed sample which is the same size as the comparator-exposed population in each stratum. Letting “*” designate the target population sampled without replacement in this manner,

$$T_i^* = C_i$$

And the number of adverse-event outcomes in each stratum for the target-exposed population will then be:

$$t_i^* = B(C_i, p_i)$$

And the IRR is estimated for this case as:

$$R^* = \frac{\sum t_i^*}{\sum c_i} = \frac{\sum B(C_i, \hat{p}_i)}{\sum B(C_i, \hat{q}_i)}$$

[Note that these formulas for R and R^* are to show the derivation of the calculations below; these quantities are not actually calculated using these formulas.]

Now let us consider the case where $C_i \geq T_i$ for some one or more strata i , and the target population is now sampled with replacement.

We must first recognize the possibility that there will be some strata where $C_i < T_i$, and in these cases, the sample variance cannot be “deflated” by sampling. This fact is easily handled by incorporating the following rule into the computations below:

Set $C_i \leftarrow T_i$ if $C_i < T_i$

Letting the “prime” symbol denote the target population sampled with replacement, we have:

$$T'_i = C_i$$

And the corresponding number of adverse event outcomes is:

$$t'_i = \frac{C_i}{T_i} B(T_i, p_i)$$

The incidence ratio is now:

$$R' = \frac{\sum t'_i}{\sum c_i} = \frac{\sum (C_i/T_i) B(T_i, \hat{p}_i)}{\sum B(C_i, \hat{q}_i)}$$

We next want to calculate the factor by which the variance of the IR is inflated due to sampling with replacement:

$$\rho = \text{Var}(R') / \text{Var}(R^*)$$

Some identities and approximations from probability theory are incorporated into the calculation of ρ . An approximate formula for the variance of the ratio of two independent random variables with no mass at a value of zero can be derived from a Taylor-series expansion⁴:

$$\text{Var}(A/B) = \frac{E[A^2]}{E[B^2]} \left[\frac{\text{Var}(A)}{E[A^2]} + \frac{\text{Var}(B)}{E[B^2]} \right]$$

For a binomial-distributed variable $X=B(N,p)$, we employ the formulas:

$$E[X] = Np,$$

$$E[X^2] = Np(Np - p + 1)$$

⁴ Kendall M, Stuart A, Ord JK. *Kendall's Advanced Theory of Statistics, Volume 1* (6th ed.). Hodder Arnold: London, U.K., 1998, 351.

And

$$\text{Var}(X) = Np(1 - p)$$

Applying the formula

$$\left(\sum_i a_i \right)^2 = \sum_i a_i^2 + 2 \sum_{j \neq i} a_i a_j$$

To the means of independent random variables gives

$$E \left[\left(\sum_i a_i \right)^2 \right] = \sum_i E[a_i^2] + 2 \sum_{j \neq i} E[a_i] E[a_j]$$

So for the independent variables t^*_i , t'_i , and c_i , we define:

$$X' = E \left[\left(\sum t'_i \right)^2 \right] = \sum (C_i/T_i)^2 T_i \hat{p}_i (T_i \hat{p}_i - \hat{p}_i + 1) + 2 \sum_{j \neq i} (C_i/T_i) T_i \hat{p}_i T_j \hat{p}_j$$

$$X^* = E \left[\left(\sum t^*_i \right)^2 \right] = \sum C_i \hat{p}_i (C_i \hat{p}_i - \hat{p}_i + 1) + 2 \sum_{j \neq i} C_i \hat{p}_i C_j \hat{p}_j$$

$$X = E \left[\left(\sum c_i \right)^2 \right] = \sum C_i \hat{q}_i (C_i \hat{q}_i - \hat{q}_i + 1) + 2 \sum_{j \neq i} C_i \hat{q}_i C_j \hat{q}_j$$

And for the variances,

$$V' = \text{Var}\left(\sum t'_i\right) = \sum (C_i/T_i)^2 T_i \hat{p}_i (1 - \hat{p}_i)$$

$$V^* = \text{Var}\left(\sum t^*_i\right) = \sum C_i \hat{p}_i (1 - \hat{p}_i)$$

$$V = \text{Var}\left(\sum c_i\right) = \sum C_i \hat{q}_i (1 - \hat{q}_i)$$

Then, using our Taylor-series derived formula,

$$\text{Var}(R') = \left(\frac{X'}{X}\right) \left[\frac{V'}{X'} + \frac{V}{X}\right]$$

$$\text{Var}(R^*) = \left(\frac{X^*}{X}\right) \left[\frac{V^*}{X^*} + \frac{V}{X}\right]$$

And the variance inflation factor is

$$\rho = \text{Var}(R') / \text{Var}(R^*)$$

Risk Estimation Methods Summary

[The 600-word section below is intended for reference in drafting the “Methods” sections of papers that need to describe SAEfetyWorks. The idea is to describe the analytical methods concisely and without equations, and to include a reference to “supplementary material” to describe the mathematical and statistical details.]

The data used for risk estimation consists of a set of records describing *subject eras*, describing non-contiguous periods of time when a subject was exposed to either the target or comparator drug. Subject era records contain the age and gender of the subject, the calendar year of the event, a binary variable indicating whether the patient was exposed to the target or the comparator drug, another binary variable indicating whether or not the adverse event of interest occurred during the era. Additionally, each era is described by a list of thousands of binary covariates, corresponding to the presence or absence of the administration of a drug, or the occurrence of a medical condition which is coded as a term in MedDRA® (Medical Dictionary for Regulatory Activities).

From this input data, covariates for a propensity score model are selected as follows: A subset of the input data is selected by matching target and comparator patients on age, gender, and calendar year of era. Within this matched cohort, for each covariate, the odds ratio of the covariate with respect to drug exposure (target vs. comparator) is calculated. The user is presented with a sorted list of all covariates with an odds ratio greater than 1.1; covariates with an odds ratio greater than 2.0 are recommended for

inclusion in the propensity score model. The user may add additional covariates based on their judgment. Prior to creation of the model, covariates are screened for co-linearity as described in [this document]. The matched cohort created in this step of analysis is discarded after covariates are selected; subsequent calculations take place on the entire data set.

A propensity score for each subject era is determined from a logistic regression model, with drug exposure as the predicted variable and the selected covariates as predictors. The duration of the era enters the model as an offset term. [Subject eras in the tails of the propensity score distribution are trimmed [at the Nth percentile]/[to the interval of common support]].

A balanced dataset is created as follows: The propensity score range is stratified into [N quantiles]. If the comparator cohort is larger than the target cohort, then all target-cohort subjects are placed into the balanced dataset. Then, within each stratum, a sample of comparator-cohort subjects of equal size is selected and placed into the balanced dataset, using sampling with replacement. If the target cohort is the larger one, the role of target and comparator are reversed in the preceding steps.

The user is presented with graphical and numerical data describing covariate balance, and is permitted to refine covariate selection and revise the model until acceptable balance is achieved.

From the balanced dataset, covariates are selected for a Poisson regression model for the adverse event outcome of interest, as follows: For each candidate covariate, the odds ratio of the covariate with respect to the outcome is calculated. The user is presented with a sorted list of all covariates with an odds ratio greater than 1.1; covariates with an odds ratio greater than 2.0 are recommended for inclusion in the propensity score model. The user may add additional covariates based on their judgment. Prior to creation of the model, covariates are screened for co-linearity as described in [this document].

Using the selected covariates, a Poisson regression model for the outcome variable is created, using forward stepwise selection of variables. Interaction terms are added to the model in a subsequent forward stepwise selection, using inclusion criteria described in [this document]. The Incidence Rate Ratio is determined from the coefficient in the model of the term for drug exposure. The Wald confidence interval is calculated, and is corrected for the influence of sampling with replacement as described in [this document].