



Incidence Density Sampling for Nested Case-Control Study Designs

2019 PhUSE EU Connect Annual Conference

Quratul Ann & Rachel Tham

NEMEA Centre of Excellence for Retrospective Studies, Real-World Insights

12 November 2019

- + Introduction
- + Real-world data (RWD)
- + Retrospective study designs
- + Sampling approaches
- + Matching methods
- + Statistical analysis system (SAS) macro
- + Conclusion

To realistically evaluate treatments for safety studies, RCTs may not be possible or even unethical in some circumstances

- **Randomization is the gold standard for experimental design**
 - Balances both measured and unmeasured confounders
- **However, randomized controlled trials (RCTs) may not be possible in cases of:**
 - Rare disease
 - Long term disease conditions
 - Severe disease where no good treatment is available
 - Determining rates of adverse events (low frequencies in RCTs)

A proposed solution:

Use of real-world data (RWD) and incidence density sampling in a nested case-control study design



Examples: Oncology trials for rare or previously untreatable disease

RWD are a valuable source of information for investigating health-related outcomes in human populations

Why real-world data?

- RWD reflects the real-life utility of treatments and clinical experiences over time as the data is often longitudinal.
- RWD more often reflect the typical use of treatments in the clinical setting and tend to encompass patients with widely varying characteristics and co-morbidities.
- There has been the rapid growth in the availability and use of routine health care data to explore effectiveness and safety.
- Use of **RWD in post-authorization safety studies (PASS)** has become standard practice.

Real-world data includes:

Routine health care data captured in electronic health records (EHR), registries, claims data, and chart review studies.

Use of RWD provides opportunities to evaluate effectiveness and safety of treatments, but also encounters some challenges

Opportunities:

- Drug safety pharmacoepidemiologic studies commonly use RWD because they reflect real-life utility of drugs.
- Studies using RWD are capable of providing compelling insights about the safety of drugs when properly conducted.

Challenges:

- Absence of randomisation in RWD studies may introduce bias such as observer bias, selection bias and recall bias.
- RWD is susceptible to confounding, which may be unrecognised due to inadequate knowledge of how variables interrelate with the outcome of interest.

How is RWD often used in a study?

The exploration of RWD often uses the framework from retrospective study designs to explore exposure-outcome associations

RWD studies use retrospective study designs to reduce bias and explore exposure-outcome associations

Cohort study design (most common study design in epidemiological research):

- The entire cohort is followed and the rate of occurrence of new cases is measured to assess relationship between exposure and outcome.
- Involves exposure measurement based on assumptions which may not reflect real-life and needs a large and lengthy study for adequate statistical power.

Case-control study design:

- Compares a case with a matched control which may come from the source eligible cohort or another source which is otherwise similar to the case source.
- Liable to selection bias, the control group should be representative of the population of interest and it's prone to methodological errors in the analysis.

Nested case-control study design:

- We identify cases and match each to controls sampled from *all those at risk in the source eligible cohort*.
- **Minimises selection bias and recall bias** by identifying a control group that is representative of the population of interest within the same at risk cohort.

Case:

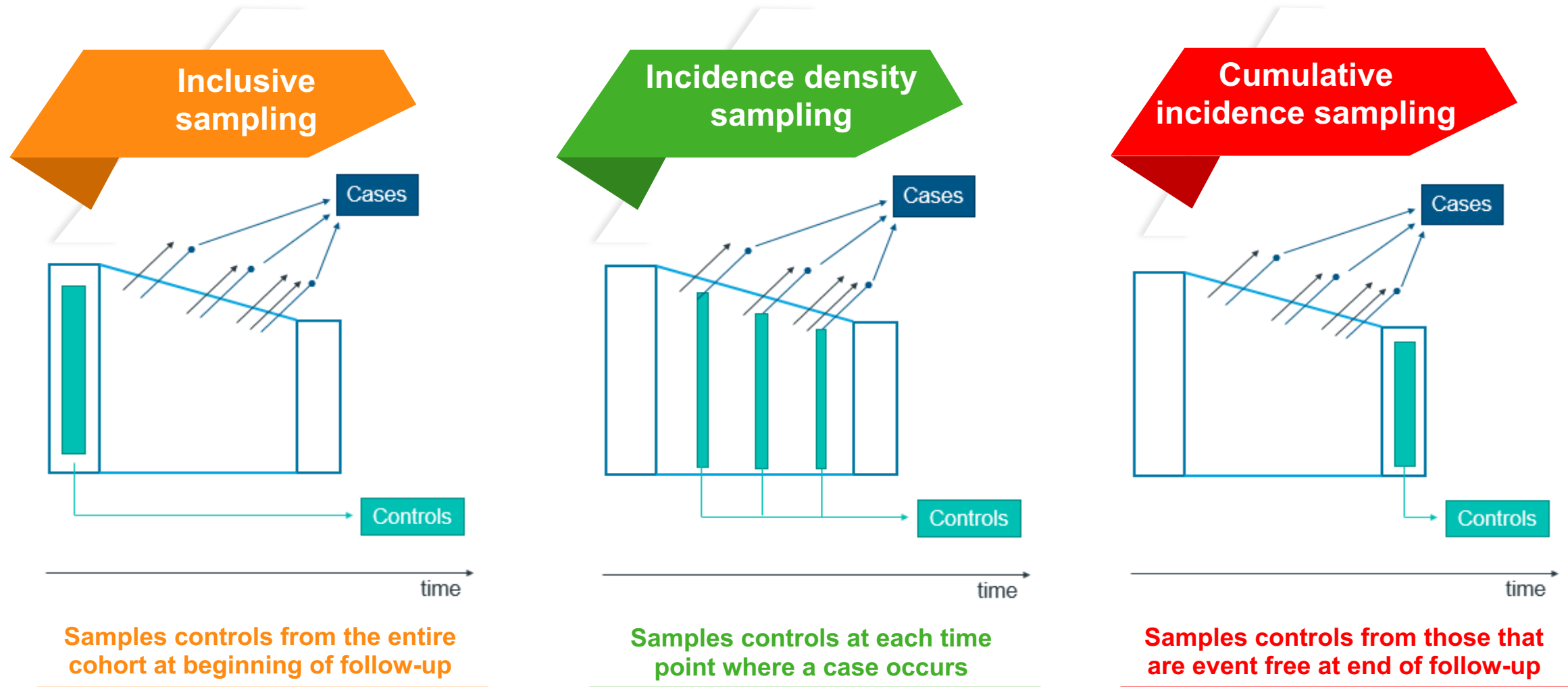
Subjects with the target disease or outcome of interest.

Control:

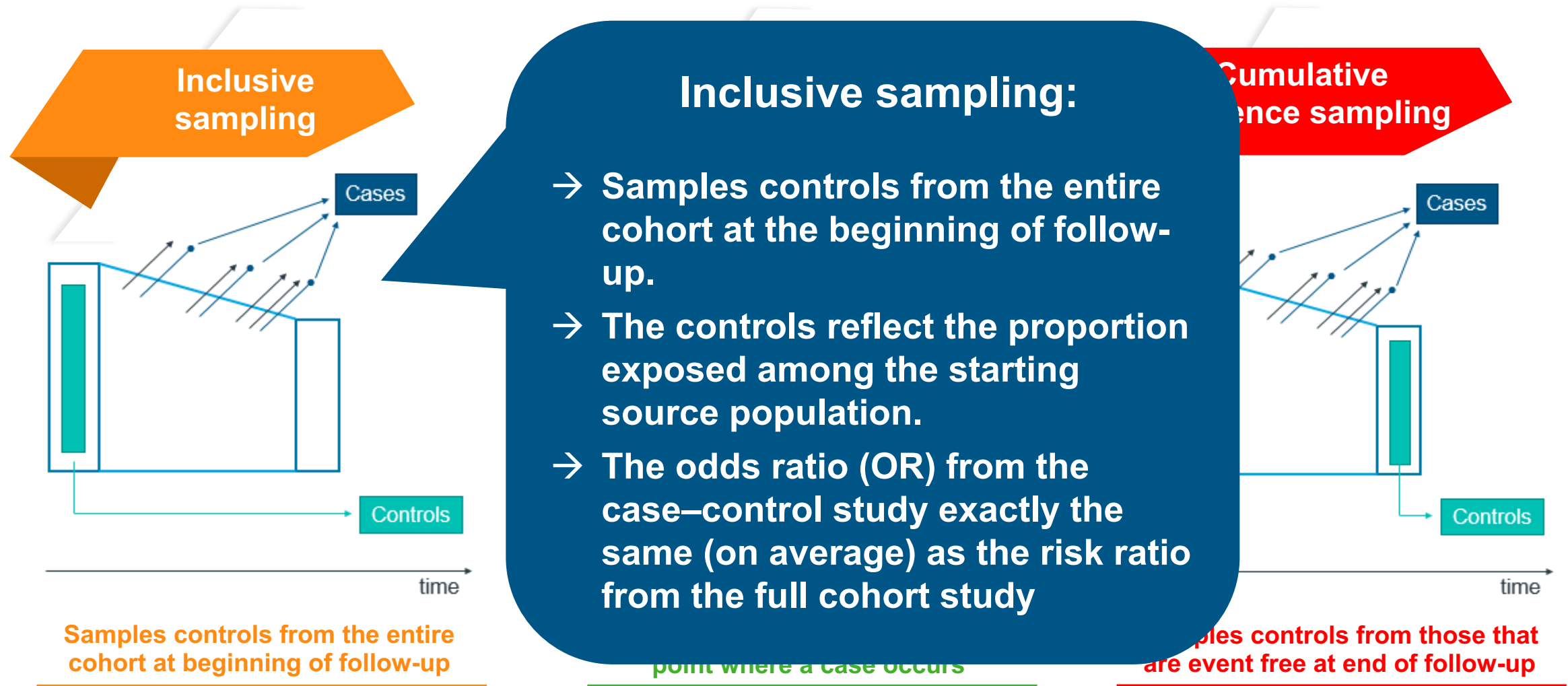
Subjects without the target disease or outcome of interest that are otherwise similar to the case.

The nested case-control is an extremely powerful study design that provides an enormous reduction in costs, data collection efforts, and analysis compared to a full study cohort approach

Sampling attempts to mimic randomisation by matching comparable controls to cases in nested case-control studies



Sampling attempts to mimic randomisation by matching comparable controls to cases in nested case-control studies



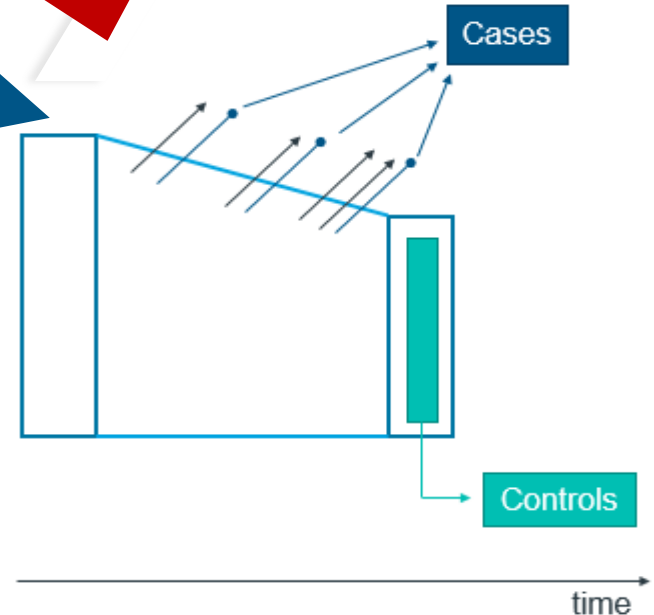
Sampling attempts to mimic randomisation by matching comparable controls to cases in nested case-control studies

Inclusion
sampling

Cumulative incidence sampling:

- Samples controls from those that are event free at the end of follow-up, and exposure status at start of follow-up identifies case/control status.
- Rare disease assumption must be met for the OR to be exactly the same as the full cohort study OR (and approximates the risk ratio/rate ratio).

Cumulative
incidence sampling



Samples controls from cohort at beginning of follow-up

point where a case occurs

Samples controls from those that are event free at end of follow-up

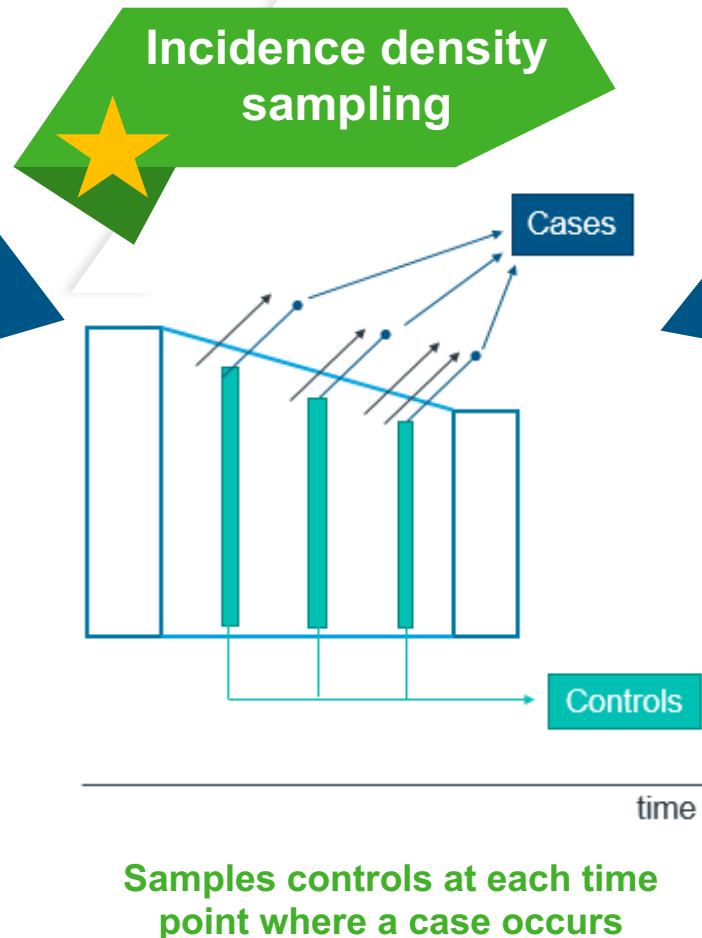
Sampling attempts to mimic randomisation by matching comparable controls to cases in nested case-control studies

Incidence density sampling is the recommended unbiased method for sampling controls

Incidence density sampling:

- Samples controls longitudinally at each point in time when a case occurs throughout the risk period
- Computationally complex

cohort at beginning of follow-up



Incidence density sampling:

- The OR from the case-control study will estimate the rate ratio from the cohort study *without the need of the rare disease assumption*
- Recommended unbiased method for sampling controls!

are event free at end of follow-up

Matching cases to controls enables a comparison of estimate treatment effects, this reduces bias due to confounding

Exact matching:

- Matches each case unit to all possible control units with exactly the same values on all the covariates included in the study
- Patients who do not find a match are excluded from analysis, which has implications for the power of the subsequent comparisons and on representativeness of the matched cohorts with regard to the true population

Multivariable matching:

- Each of the case unit is matched to one or more control units as similar as possible with respect to the given set of pre-treatment variables.
- Shows separate effect of each covariate considered in the analysis and balance the distribution of observed characteristics

Propensity score matching: (most common matching method in observational studies)

- Patients are matched on a single propensity score representing the probability of receiving the treatment given the observed baseline characteristics
- Balances the distribution of observed characteristics across cases and controls and useful when rare outcome (standard regression models may not converge)

Goal of matching:

To find one (or more) similar non-treated subjects for each treated subject with similar observable characteristics against whom the effect of the treatment can be assessed

Our 3-phase macro for incidence density sampling using RWD in nested case-control studies



How to prepare data for the macro

- Required variables in the input dataset

Variable name	Definition
Pat_ID	The unique identifier for the patient/participant included in the study
Index_Date	The start date for the beginning of the follow-up period for the Pat_ID
Index_Age	The individual (Pat_ID) age (in years) at their Index_Date
Censor_Date	The last date of follow up for the Pat_ID
Censor_Age	The individual (Pat_ID) age (in years) at their Censor_Date
Censor	Study disease flag (0 = no experience/non-case, 1 = experienced/case)

- Optional variables

Variable name	Definition/When is it required
DOB	Date of birth, required if AgeRange parameter is specified
[other matching vars]	Additional matching variables, required if specified in ExactVar or RangeVar

- Macro generated variables

Variable name	Definition
Match_ID	The unique identifier for a case and its matched controls (the do loop count)
CaseCont	A flag indicating if the row is a control (0) or case (1)

Macro notes and required parameters

- Macro call:

%IDSmatch (InTodata=, AgeRange=, ExactVar=, RangeVar=, nControl=, RandSeed=);

- Macro parameters:

Parameter	Definition
InToData	Input dataset that contains required and optional variables for matching
AgeRange	Range +/-"x" years from the cases start age, can be missing
ExactVar	Variable names for matching on exact same value as the case (NO COMMAS)
RangeVar	Variable name and acceptable range for matching to case (NO COMMAS)
nControl	Max number of controls per case. For all matches, specify 0
RandSeed	Sets the seed for generating random numbers

- Macro notes (see paper for full details):
 - This macro creates (or replaces tables if similarly named) and deletes named tables in the work library (see paper)
 - This macro generates (and will replace any) formats called: casecont (0 = control, 1 = case), MatchFmt (for the values of 0 to nControl, if nControl parameter > 0)
 - This macro will replace any global title and footnote statements

Sample macro report output and how to interpret it

- Summary statistics sampling report for number of matches per case

Number of Cases	Average number of matches per case	Standard deviation for the number of matches per case	Minimum number of matches per case	25th percentile for number of matches per case	Median number of matches per case	75th percentile for number of matches per case	Maximum number of matches per case
8142	24	31	0	2	11	34	213

Total number of cases in test: 8142

Total number of rows in test: 50000

Number of cases with no matches: 1041

- Matching report for n:1 matching

	n	%
Number of Matches per Case		
0	1041	12.79
1	606	7.44
2	442	5.43
3	361	4.43
4 +	5692	69.91
TOTAL	8142	100.00

Test is the name of the input dataset IDSmatch is run on

If there is a large proportion of cases without matches, the exact matching criteria may be too strict

This category contains all cases that have at least “n” matches. For this run of the macro, 4 was specified. So for a case to contribute to this category they had to have 4 to 213 sampled controls (213 is the max as seen above)

Conclusions

- Real-world data (RWD) provides a realistic utilisation of treatments and clinical experiences over time
- RWD can be used in situations where randomised clinical trials (RCT) may not be possible or even unethical and is currently the standard practice in post-authorization safety studies (PASS)
- RWD utilises retrospective study designs, such as the nested case-control, to reduce bias
 - ❖ The nested case-control is an extremely powerful study design that minimises selection bias and recall bias while provides an enormous reduction in costs, data collection efforts, and analysis
- Sampling is an analysis technique that attempts to mimic randomisation by selecting comparable controls for cases and thus adjusts for confounding
 - ❖ Incidence density sampling samples controls at each point in time where a case is identified and is the recommended unbiased method for sampling controls
- Matching is the procedure that is used to identify which controls are comparable to cases. Common matching techniques are: exact, multivariable and propensity score
- The 3-phase macro we have designed is for incidence density sampling using exact matching with RWD in nested case-control studies

References

- Beaumont, J. J., Steenland, K., Minton, A., & Meyer, S. (1989). A Computer Program for Incidence Density Sampling of Controls in Case-Control Studies Nested within Occupational Cohort Studies. *American Journal of Epidemiology*, Vol. 129, No. 1.
- Gately, R., & Turnbull, B. (2009). Incidence-Density Sampling: Alternative Programming Approaches to Achieve the Same End. *NESUG 2009 - Pharmaceuticals, Health Care, and Life Sciences*.
- Kiri, V. A. (2012). Programming Challenges of Sampling Controls to Cases from the Dynamic Risk Sets in Nested Case Control Studies. *PhUSE*.
- Matthews, R., & Brill, I. (n.d.). SAS® Programs to Select Controls for Matched Case-Control Studies. *SUGI 30*.
- Pearce, N. (1989). Incidence Density Matching with a Simple SAS Computer Program. *International Journal of Epidemiology*, 981-984.
- Pearce, N. (2016). Analysis of matched case-control studies. *British Medical Journal (BMJ)*. doi: <https://doi.org/10.1136/bmj.i969>
- Prentice, R. I. (June 1986). On the Design of Synthetic Case-Control Studies. *Biometrics*, 42, 301 - 310 .
- Richardson, D. B. (2004). An incidence density sampling program for nested case-control analyses. *Occup Environ Med*. doi:10.1136/oem.2004.014472
- Robins, J. M., Gail, M. H., & Lubin, J. H. (Jun., 1986). More on "Biased Selection of Controls for Case-Control Analyses of Cohort Studies". *International Biometric Society*, 42, No. 2, 293 - 299. Retrieved from <http://links.jstor.org/sici?sici=0006-341X%28198606%2942%3A2%3C293%3AMO%22SOC%3E2.0.CO%3B2-C>
- Steyerberg, E. W. (2019). *Clinical Prediction Models*. Leiden University Medical Center, Leiden, The Netherlands: Springer Nature Switzerland AG 2019.
- Vandenbroucke, J. P., & Pearce, N. (2012). Case-control studies: basic concepts . *International Journal of Epidemiology* , 1480-1489.
- Vandenbroucke, J. P., & Pearce, N. (2012). Incidence rates in dynamic populations. *International Journal of Epidemiology*, 1472-1479.
- Wang, M.-H., Shugart, Y. Y., Cole, S. R., & Platz, E. A. (July 24, 2019). A Simulation Study of Control Sampling Methods for Nested Case-Control Studies of Genetic and Molecular Biomarkers and Prostate Cancer Progression. *American Association for Cancer Research*, *Cancer Epidemiol Biomarkers Prev* 2009;18:706–11. Retrieved from <https://cebp.aacrjournals.org/>
- Hennekens CH, Buring JE. Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

Questions?



Rachel Tham

Rachel.Tham@IQVIA.com

www.linkedin.com/in/racheltham1/

Quratul Ann

Quratul.Ann@IQVIA.com

www.linkedin.com/in/quratulann/

Appendix – Incidence Density Sampling SAS Macro



IDS code