

Fuzzy Joins with Proc SQL for Better Data Utilization

Sunil Gupta, Cytel, Simi Valley, CA, USA

ABSTRACT

Often when working with lab data, especially during the early measurement collection period, there is limited data. One of the challenges of utilizing limited data is to make the most of it for new measurement calculations. When new lab calculations depend on several lab components, it is often required that lab dates be first aligned. The challenge exists when the lab dates are within plus or minus three days of each other, for example. How exactly can you align these lab dates to calculate new lab tests?

In this presentation, I am introducing a robust fuzzy join method to apply logic for joining based on dates within a window. With this approach, lab data is better utilized without having to be collected on the same lab date.

INTRODUCTION

Instead of using traditional Proc SQL joins, which is based on lab dates matching, fuzzy Proc SQL joins were applied to better align records from two lab datasets with different lab dates but within a window range.

The objective is to calculate Ncart using WBC, MONOLE and LYMLE from one lab dataset with one set of dates with QPPB from a second lab dataset with another set of dates. Fuzzy Proc SQL joins enabled us to calculate more Ncart values since more alignments between the two lab datasets were possible.

$$Ncart = 1000 * (WBC) * (MONOLE + LYMLE) / 100 * QPPB / 100$$

FUZZY VISIT WINDOWS

QPPB DATE	WBC, MONOLE AND LYMLE DATE	NCART = $1000 * (WBC) * (MONOLE + LYMLE) / 100 * QPPB / 100$
DATE 1	ALL NON-MISSING ON DATE 1	DERIVE NCART
DATE 1	ALL NON-MISSING AND SHORTEST DISTANCE ON DATE 2 WITHIN +/- 3 DAYS IF LBDY <= 35	DERIVE NCART
DATE 1	ALL NON-MISSING AND SHORTEST DISTANCE ON DATE 2 WITHIN +/- 10 DAYS IF LBDY > 35	DERIVE NCART
DATE 1	INDIVIDUAL NON-MISSING AND EQUAL DISTANCE THAN TAKE AVERAGE OF WBC, MONOLE AND LYMLE	DERIVE NCART

NINE KEY STEPS FOR FUZZY JOINS

Below are the nine steps to perform the Fuzzy Proc SQL join. One of the benefits of this method is the ability to customize visit windows on how the fuzzy join is applied based on lbdy values.

- 1 - Create lbdymin lbdymax var that define windows based on +/- 3 days if lbdy <= 35 and +/- 10 days if lbdy > 35;
- 2 - Identify non-missing records for monole, wbc, lymle and nmcyt;
- 3 - Filter only non-missing records;
- 4 - Proc sql with overlap of lbdy and lbdymin and lbdymax;
- 5 - Join within +/- 3 days if lbdy <= 35 and +/- 10 days if lbdy > 35;

PhUSE US Connect 2018

- 6 - Select only the closest to original date;
- 7 - Average if same distance from original date;
- 8 - Check for duplicate same distance from target;
- 9 - Replace missing values with fuzzy join values based on windows;

KEY FUZZY PROC SQL JOIN

The Fuzzy Proc SQL Join below uses a combination of LEFT JOIN, DATE RANGE MATCHES, GROUP BY and HAVING CLAUSE based on CALCULATED variable. This is true power of aligning lab dates within a visit window range.

```
proc sql;
  create table lb_newa2 as
  select a.*, c.monole_i, c.wbc_i, c.lymle_i, c.lbdymin, c.lbdymax, abs(c.lbdy - a.lbdy) as lbdydif
  from lb_new as a left join lb_newa1 as c
  on (a.usubjid = c.usubjid and a.aperiod = c.aperiod) and
     ( (c.lbdymin <= a.lbdy <= c.lbdymax) )
  group by a.usubjid, a.aperiod, a.lbdt, a.lbdy having calculated lbdydif = min(calculated lbdydif);
quit;
```

SUMMARY

When lab data comes from multiple sources and not always on the same date, better data utilization often requires more creativity in data management. Fuzzy Proc SQL joins with the option to customize visit windows enables us to calculate more Ncart values than using the traditional Proc SQL join by lab date. As a result, more Ncart values are used in statistical analysis and modeling.

CONTACT INFORMATION

The author welcomes your comments and suggestions.

Sunil K. Gupta
Director, Statistical Programming
Cytel
213 Goldenwood Circle
Simi Valley, CA 93065
E-mail: Sunil.Gupta@Cytel.com

Sunil Gupta, MS, is an international speaker, best-selling SAS author, and a global CDISC corporate trainer. Sunil is a Director of Statistical Programming at Cytel and has over twenty-five years of experience in the pharmaceutical industry. Most recently, Sunil is a CDISC Oncology ADaM reviewer and has taught his CDISC online class at the University of California at San Diego and SAS Institute India. Sunil also taught his Sharpening Your SAS Skills online class for UCLA Extension. In 2011, Sunil launched his unique SAS resource blog, SASSavvy.com, for smarter SAS searches and teaches Data Science Using SAS. Sunil has MS in Bioengineering from Clemson University and a BS in Applied Mathematics from the College of Charleston.



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.