

# Fuzzy Joins with Proc SQL for Better Data Utilization



**Sunil Gupta**  
**Director, Statistical**  
**Programming**  
**Cytel**



# Proc SQL Tip: Fuzzy Join to Derive Ncart

**QPPB Dataset – generally limited data and not aligned with hematology data**

|            |           |       |   |    |         |                  |
|------------|-----------|-------|---|----|---------|------------------|
| % of PBMCs | URMC LABS | PL... | . | -5 | Day -5  | 2016-02-24T00... |
| % of PBMCs | URMC LABS | PL... | . | 7  | Day 7   | 2016-03-14T04... |
| % of PBMCs | URMC LABS | PL... | . | 8  | Week 2  | 2016-03-21T12... |
| % of PBMCs | URMC LABS | PL... | . | 9  | Week 4  | 2016-04-04T13... |
| % of PBMCs | URMC LABS | PL... | . | 11 | Month 3 | 2016-06-10T10... |
| % of PBMCs | URMC LABS | PL... | . | 12 | Month 6 | 2016-08-29T11... |
| % of PBMCs | URMC LABS | PL... | . | 13 | Month 9 | 2016-12-01T09... |



**Fuzzy Join both datasets to Derive Ncart**



**WBC, MONOLE and LYMLE Dataset – needs to be transposed**

|        |                   |            |      |         |      |        |     |   |   |        |                  |
|--------|-------------------|------------|------|---------|------|--------|-----|---|---|--------|------------------|
| LYMLE  | Lymphocytes (...) | HEMATOLOGY | 41.7 | %       | 41.7 | %      | 22  | C | 8 | Week 2 | 2016-03-21T00... |
| MONOLE | Monocytes (%)     | HEMATOLOGY | 11.9 | %       | 11.9 | %      | 4   | C | 8 | Week 2 | 2016-03-21T00... |
| WBC    | Leukocytes        | HEMATOLOGY | 2.5  | 10^3/uL | 2.5  | 10^9/L | 4.5 | C | 8 | Week 2 | 2016-03-21T00... |
| LYMLE  | Lymphocytes (...) | HEMATOLOGY | 13.8 | %       | 13.8 | %      | 22  | C | 9 | Week 4 | 2016-04-04T13... |
| MONOLE | Monocytes (%)     | HEMATOLOGY | 6.3  | %       | 6.3  | %      | 4   | C | 9 | Week 4 | 2016-04-04T13... |
| WBC    | Leukocytes        | HEMATOLOGY | 4.0  | 10^3/uL | 4    | 10^9/L | 4.5 | C | 9 | Week 4 | 2016-04-04T13... |

# Write more Robust Code

## - Fuzzy Join to Derive Ncart

| QPPB<br>Date 1 | WBC, MONOLE and LYMLE<br>Date 2                                                      | $\text{Ncart} = 1000 * (\text{WBC}) * (\text{MONOLE} + \text{LYMLE}) / 100 * \text{QPPB} / 100$ |
|----------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Date 1         | All non-missing on Date 1                                                            | Derive Ncart<br>(Traditional Proc SQL Join)                                                     |
| Date 1         | All non-missing and shortest distance on Date 2 within +/- 3 days if LBDY <= 35      | Derive Ncart<br>(Fuzzy Proc SQL Join)                                                           |
| Date 1         | All non-missing and shortest distance on Date 2 within +/- 10 days if LBDY > 35      | Derive Ncart<br>(Fuzzy Proc SQL Join)                                                           |
| Date 1         | Individual non-missing and equal distance than take average of WBC, MONOLE and LYMLE | Derive Ncart<br>(Fuzzy Proc SQL Join)                                                           |



# Key Steps: Fuzzy Join to Derive Ncart

## Nine Key Steps

- 1 - create lbdymin lbdymax var that define windows based on +/- 3 days if lbdy  $\leq$  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  $\leq$  35 and +/- 10 days if lbdy  $>$  35;
- 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;

# Fuzzy Join to Derive Ncart

```
data lb_newa1;  
set lb_new;  
by usubjid aperiod lbdy lbdy;
```

**\* 1 - create lbdymin lbdymax var that define windows based on +/- 3 days if lbdy <= 35 and +/- 10 days if lbdy > 35;**

```
if lbdy <= 35 then do; lbdymin = lbdy - 3;  
    lbdymax = lbdy + 3; end;  
else if lbdy > 35 then do; lbdymin = lbdy - 10;  
    lbdymax = lbdy + 10; end;
```

```
rename monole=monole_i wbc=wbc_i lymle=lymle_i;
```

**\* 2 - identify non-missing records for monole, wbc, lymle and nmcyt;**

```
if nmiss(monole, wbc, lymle) = 0 or wbc = 0 then nmcyt=1;
```

**\* 3 - filter only non-missing records;**

```
if nmcyt=1;  
drop QPPB; run;
```

*continued...*

# Fuzzy Join to Derive Ncart

- \* 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;
- \* 6 - select only the closest to original date;

```
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;
```

*continued...*

# Fuzzy Join to Derive Ncart

**\* 7 - average if same distance from original date;**

```
proc sql;  
  create table lb_newa3 as  
  select usubjid, aperiod, lbdt, lbdy, mean(monole_i) as monole_i2,  
  mean(wbc_i) as wbc_i2, mean(lymle_i) as lymle_i2, min(qppb) as qppb1  
  from lb_newa2 group by usubjid, aperiod, lbdt, lbdy;  
quit;
```

**\* 8 - check for duplicate same distance from target;**

```
proc sort data =  
lb_newa3 nodupkey out=lb_newa3_nodup dupout=lb_newa2_dups;  
by usubjid aperiod lbdt lbdy;  
run;
```

*continued...*

# Fuzzy Join to Derive Ncart

```
proc sort data=lb_newa3;  
  by usubjid aperiod lbdt lbdy;  
run;
```

```
data lb_newb;  
  merge lb_new lb_newa3;  
  by usubjid aperiod lbdt lbdy;
```

**\* 9 - replace missing values with fuzzy join values based on windows;**

```
if lymle = . then lymle=lymle_i2;  
if monole = . then monole=monole_i2;  
if wbc = . then wbc=wbc_i2;
```

```
  drop lymle_i monole_i wbc_i lymle_p1 lymle_n1 monole_p1 monole_n1  
wbc_p1 wbc_n1 lbdy_p1 lbdy_n1 impute;  
run;
```

# Fuzzy Joins with Proc SQL for Better Data Utilization



**Sunil Gupta**  
**Director, Statistical**  
**Programming**  
**Cytel**

