

# A Many to Many Merge, Without SQL?

## Paper TU05

David Franklin  
Independent SAS Consultant  
TheProgrammersCabin.com

While you are waiting, some trivia ....

“The first U.S. Transcontinental Flight took 84 days, 17SEP1911 to 10DEC1911.”

# Introduction

- Merging two datasets is a common task
- Common format is either a 1-to-1, 1-to-many or many-to-many
- Many-to-many is almost exclusively handled thru a PROC SQL call
- Paper looks at a many-to-many with a datastep

# First some data ...

```
data ae; ** Adverse Event Data;  
  infile cards; input ptnum $ 1-3 @5 date date9. event $ 15-35;  
  format date date9.;
```

```
cards;  
001 16NOV2009 Nausea  
002 16NOV2009 Heartburn  
002 16NOV2009 Acid Indigestion  
002 18NOV2009 Nausea  
003 17NOV2009 Fever  
003 18NOV2009 Fever  
005 17NOV2009 Fever
```

Subjects 001 and 005 have a one-to-one match

Subject 002 has a many to many match on 16NOV2009, but a one to one match on 18NOV2009

```
;  
data cm; ** Concomitant Medication Data;  
  infile cards; input ptnum $ 1-3 @5 date date9. medication $ 15-35;  
  format date date9.;
```

```
cards;  
001 16NOV2009 Dopamine  
002 16NOV2009 Antacid  
002 16NOV2009 Sodium bicarbonate  
002 18NOV2009 Dopamine  
003 18NOV2009 Aspirin  
004 19NOV2009 Aspirin  
005 17NOV2009 Aspirin
```

Subject 003 has no match for 17NOV2009 but does have a match for 18NOV2009

Subject 004 has no match.

```
;  
run;
```

Paper TU05, PharmaSUG2011 : David Franklin

# *Lets look ahead to our final dataset*

- we are doing an inner join, i.e. where there is no match the observation is excluded

Obs	ptnum	date	event	medication
1	001	16NOV2009	Nausea	Dopamine
2	002	16NOV2009	Heartburn	Antacid
3	002	16NOV2009	Heartburn	Sodium bicarbonate
4	002	16NOV2009	Acid Indigestion	Antacid
5	002	16NOV2009	Acid Indigestion	Sodium bicarbonate
6	002	18NOV2009	Nausea	Dopamine
7	003	18NOV2009	Fever	Asprin
8	005	17NOV2009	Fever	Asprin

# Looking at the merge in detail ...

ptnum	date	event
001	16NOV2009	Nausea
002	16NOV2009	Heartburn
002	16NOV2009	Acid Indigestion
002	18NOV2009	Nausea
003	17NOV2009	Fever
003	18NOV2009	Fever
005	17NOV2009	Fever

AE

Subjects 001 and 005 have a one-to-one match

Subject 002 has a many to many match on 16NOV2009, but a one to one match on 18NOV2009

ptnum	date	medication
001	16NOV2009	Dopamine
002	16NOV2009	Antacid
002	16NOV2009	Sodium bicarbonate
002	18NOV2009	Dopamine
003	18NOV2009	Asprin
004	19NOV2009	Asprin
005	17NOV2009	Asprin

CM

Subject 003 has no match for 17NOV2009 but does have a match for 18NOV2009

Subject 004 has no match.

Obs	ptnum	date	event	medication
1	001	16NOV2009	Nausea	Dopamine
2	002	16NOV2009	Heartburn	Antacid
3	002	16NOV2009	Heartburn	Sodium bicarbonate
4	002	16NOV2009	Acid Indigestion	Antacid
5	002	16NOV2009	Acid Indigestion	Sodium bicarbonate
6	002	18NOV2009	Nausea	Dopamine
7	003	18NOV2009	Fever	Asprin
8	005	17NOV2009	Fever	Asprin

# Maybe a MERGE Statement?

```
** SAS Program;  
data all0;  
  merge ae cm;  
  by ptnum date;  
title1 "Merge using the MERGE statement -- this fails";  
proc print data=all0;  
run;
```

\*\* SAS Output;

Merge using the MERGE statement -- this fails

Obs	ptnum	date	event	medication
1	001	16NOV2009	Nausea	Dopamine
2	002	16NOV2009	Heartburn	Antacid
3	002	16NOV2009	Acid Indigestion	Sodium bicarbonate
4	002	18NOV2009	Nausea	Dopamine
5	003	17NOV2009	Fever	
6	003	18NOV2009	Fever	Asprin
7	004	19NOV2009		Asprin
8	005	17NOV2009	Fever	Asprin

# The SQL – not the way we want

```
proc sql;
  create table all0 as
    select a.*, b.medication
    from ae a inner join cm b
    on a.ptnum=b.ptnum and a.date=b.date;
quit;
title1 "Merge using SQL -- most common way this is seen done";
proc print data=all0;
run;
```

```
** SAS Output;
Merge using SQL -- most common way this is seen done
```

Obs	ptnum	date	event	medication
1	001	16NOV2009	Nausea	Dopamine
2	002	16NOV2009	Heartburn	Antacid
3	002	16NOV2009	Heartburn	Sodium bicarbonate
4	002	16NOV2009	Acid Indigestion	Antacid
5	002	16NOV2009	Acid Indigestion	Sodium bicarbonate
6	002	18NOV2009	Nausea	Dopamine
7	003	18NOV2009	Fever	Asprin
8	005	17NOV2009	Fever	Asprin

# Going Loopy!

- The dataset with POINT option with the SET statement is not often seen but it gives the most control -- key to this is that the dataset that takes each observation in AE and then tries to match this with each observation in CM (this is basically a loop within a loop)

```
** SAS Code;
data all1;
  set ae;
  drop _; ** Drop temporary variables;
  match=0; ** Match flag;
  ** Our loop within a loop -- output if match;
  do i=1 to xnobs;
    ** Need to rename the "merging" variables within CM dset;
    set cm (rename=(ptnum=ptnum date=date)) nobs=xnobs point=i;
    ** Have to rename matching variables so that they do not
      overwrite the original values in AE;
    if ptnum=ptnum and date=date then do;
      match=1; ** Yes, there is a match!;
      output;
    end;
  end;
run;

title1 "Merge using using the POINT option in a SET statement";
proc print data=all1;
run;
```

Paper TU05, PharmaSUG2011 : David Franklin



# Output

```
** SAS Output;
```

```
Merge using using the POINT option in a SET statement
```

```
Obs ptnum date event match medication
```

1	001	16NOV2009	Nausea	1	Dopamine
2	002	16NOV2009	Heartburn	1	Antacid
3	002	16NOV2009	Heartburn	1	Sodium bicarbonate
4	002	16NOV2009	Acid Indigestion	1	Antacid
5	002	16NOV2009	Acid Indigestion	1	Sodium bicarbonate
6	002	18NOV2009	Nausea	1	Dopamine
7	003	18NOV2009	Fever	1	Asprin
8	005	17NOV2009	Fever	1	Asprin

# Conclusion

- A many-to-many merge without using SQL was accomplished
- There are a couple of other ways to do this merge without SQL, notably a variation on the "Loading Unique Dataset into an Array" which was presented at PharmaSUG 2010 entitled "Countdown of the Top 10 Ways to Merge Data".
- However, the technique shown above is a shorter method with similar control (however the use of the array will do a many-to-many merge with the equivalent SQL OUTER JOIN).

# *Questions and Contact Information*

## Questions?

## Contact Information

David Franklin

Independent SAS Consultant

16 Roberts Road, Litchfield, NH 03052

Phone: 603-275-6809

Email: [dfranklin@TheProgrammersCabin.com](mailto:dfranklin@TheProgrammersCabin.com)

Web: <http://www.TheProgrammersCabin.com>

LinkedIn: <http://www.linkedin.com/in/davidfranklinnh>



# *Acknowledgement*

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 registered trademarks or trademarks of their respective companies.