Relationships Between CDISC Variables Linking Domains (RELREC) (Part 3 of 3)

Susan Fehrer (BioClin, Inc.) and Russ Lavery

ABSTRACT

Many CDISC variables are related to each other and this paper is one of a short series of example-focused papers devoted to exploring relationships among CDISC variables. This paper focuses on using the RELREC domain to link or establish relationships among records in the AE and CM domains. RELREC is used for two purposes. First it can be used to establish relationships among observations in different data sets. The CDISC Implementation Guide (IG) says that one should use RELREC to establish those relationships collected on the CRF. This purpose will be illustrated in this paper.

The second use of RELREC is to define relationships between whole domains/data sets. This is required when a sponsor has divided data into two data sets and needs to be able to examine the data, together as part of analysis. This use of RELREC will not be discussed in this paper.

As this paper is part of a series on variable relationships, the paper will digress and also discuss relationships among variables in the three domains, AE, CM, and RELREC. Variables occur (with different two character prefixes and the same suffix) in multiple domains and this “repeating of variables in multiple domains” is a type of relationship. There will be sections of the paper that appear repetitive. This is intentional. Relationships among variables in a domain will be developed by explanation. The fact that a domains have similar structures, similar variables, is a form of relationship. We will develop this second type of relationship by repeating variable descriptions.

All information in this paper is already present in the CDISC Implementation Guide, but it is hoped that a short, topic-focused, example-based, discussion of these issues might be of use to the community.

INTRODUCTION

SAS® software users like “examples” and this paper explores the RELREC concept via two examples. One example involves buying shoes and the other example involves recording of adverse events and concomitant medication information. Specifically the paper concentrates on the relating/grouping of records in two different domains/tables/files via the RELREC domain.

Variables occur (with different two character prefixes and the same suffix) in multiple domains and this “repeating of variables in multiple domains” is a type of relationship. There will be sections of the paper that appear repetitive. This is intentional. Relationships among variables in a domain will be developed by explanation. The fact that a domains have similar structures, similar variables, is a form of relationship. We will develop this second type of relationship by repeating variable descriptions.

Understanding relationships among variables helps both in programming and in edit checking. Understanding relationships allows a QC reviewer to say not only, “this variable correctly contains information from the CRF”, but also to say “and since this variable has this value this other variable should be valued according to this rule”.

The RELREC data set was structured to allow study designers to establish relationships among records in two or more data sets. The CDISC group does not expect that programmers will, after the data are collected, use RELREC to create links between all or even many variables. Rather, it is expected that RELREC will be only used to establish links that are of research interest and that are collected on the CRF.

RELREC establishes links between rows of data in *different* domains (domain is “CDISC speak” for data sets or tables). If the relationship that needs to be established is within a *single* domain, RELREC is not needed to establish the relationship. A relationship among rows in a single domain can be defined, in a simple manner, using variables within the data set. Specifically; - -CAT, - -SCAT and - -GRPID can be used to establish relationships within a domain. The “- -“ here indicates that there are two characters missing from the variable name the two characters that indicate which domain contains the variables.

In this paper we will refer several times to the idea of a “key” and use that idea in the way that an IT person might use it. A “key” is one or more variable(s) that allows a computer to identify ONE AND ONLY ONE row in a data set. In SAS, keys are often used in merging files, a one to one by merge. Technically, “natural keys” are part of the data record and “synthetic keys” are generated by the computer system. CDISC keys are a combination of these two types of components. An understanding of natural and synthetic keys is not important to the use of RELREC.

Example #1 - Introduction to Keys and to a version of RELREC file - data from “The Shoe Sole” online shoe store. A very informal example of a “key” is shown below. It is intended to illustrate the principle of a RELREC file in a very simple business setting.

A mail order house selling shoes has kept track of its orders and shipments in two separate files. In our example, “The Shoe Sole” ordering and shipping departments are run using different programs. In our imaginary company the programs do not communicate. Each program needs to generate its own keys to keep track of its records and since the programs do not communicate, the keys are different.

Customer Number and Customer Order Number will identify a unique row, as a key in the Orders files. Customer Number and Customer Shipping Number will identify a unique row, as a key in the Shipping files. However; the values in the variables Customer Order Number and Customer Shipping Number are not common to both the files so the files cannot be merged by the keys.

Because of customer behavior and inventory levels, there is usually not a one-to-one relationship between orders and shipping in “The Shoe Sole” systems. Some orders are cancelled by the customer before shipping. Occasionally, shoes are not in stock. In that case, one order causes multiple shipments as the back-ordered shoes are received by “The Shoe Sole” and quickly shipped out to customers. Arrows in the graphic below show the true relationships between entries in the order and shipping files. These files do not have any common linking variables.

A relationship between the Order and Shipping files may be established by creating and valuing another file. We will call this file RELREC for **Rel**ationship among **Rec**ords file. We will enter key information from the observations to be “linked” into rows in RELREC. The way to link rows in RELREC itself done by valuing a variable called Relationship ID. The relationship ID number is the mechanism that links the rows in RELREC and through the keys on the rows in RELREC links rows in the Orders and Shipping files.

The logic is:

- All rows in RELREC having the same relationship ID value are related.
- A row in RELREC contains key information that identifies one row in another file.
- The rows in RELREC that are linked via a common value in RELID contain keys that describe rows in other files.
- If RELID shows that the rows in RELREC are linked, the observations identified by the keys on those rows are linked.

In this example and graphic, Russ has not placed many orders for shoes (only 5). He has never cancelled an order or had shoes backordered (ordering brown loafers creates little buyer regret and the shoes are almost always in stock). Because of his simple, and boring, purchase habits, his Customer Order number and Customer Shipping Order number happen to have the same values.

Susan (“I’ll take a pair in every color”) Fehrer has placed many orders for shoes; 124 shoe orders placed before this data pull. She has canceled several orders and has had several orders split-shipped because of “out of stock” conditions. As can be seen below, her 120th order corresponds with her 90th shipment.

Susan’s 122nd order involved some exotic shoes - several pairs of Taryn Rose dress walking shoes so she could comfortably walk through Lisbon, Portugal - making a “Stadtbummel” as she says. Those shoes had to be backordered and delivered in three shipments.

No matter how many files “The Shoe Sole” has, the four bracketed variables, RDOMAIN, USUBJID, IDVAR, and IDVARVAL, allow a user to identify a specific row of data anywhere in the computer system of the company.

The above RELREC file illustrates one way of linking files that do not have a common key. The first column identifies a file in the computer system. The second column identifies a customer in that system. Columns three and four work together, intimately, to identify a variable and then a value for that variable.

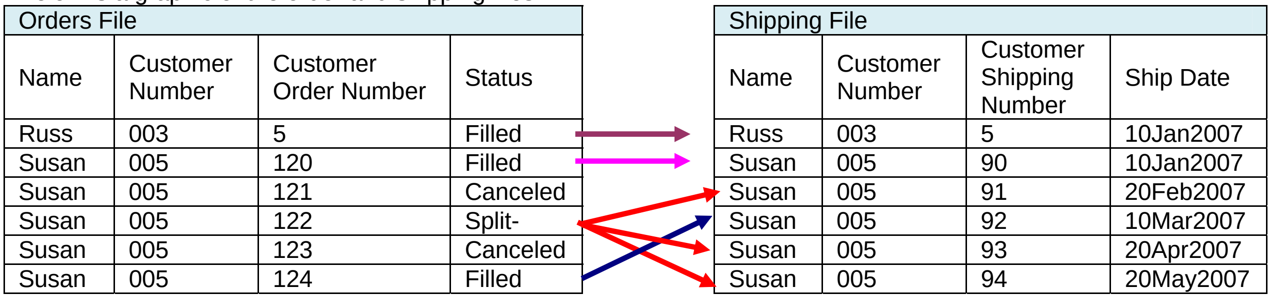
Combined, the four left columns say to go a system and then to a customer within that system. For that customer, go to a variable (this component happens to be a synthetic key) and look for a value in that variable. This process identifies one row of data in a different file. The far right column in RELREC is the variable that is used to link rows. The far right column establishes the relationships.

A simple relationship is shown by the rows for Russ. His fifth order was filled immediately and is linked to his fifth shipment. Those shoes shipped on 10Jan2007. Coincidentally, for both rows for customer Russ, the values of IDVARVAL are 5.

For a more complex example we can look at the third and fourth rows in RELREC. We read this as: the row in the orders table for Customer Number 005 (where customer Order Number has a value of 120) is linked to the row in the shipping table for Customer Number 005, where Customer Shipping Number has a value of 90. Susan's 90th order was shipped on 10Jan2007.

The most complex relationship is described in rows five to eight in the RELREC File. Susan, in her Order Number 122, ordered several pairs of some very trendy shoes (seen in Elle, I think). The Shoe Sole had just shipped all of their stock to their Rodeo Drive store and had none left for Susan when she placed her order. As a result the shoes were back-ordered and Susan's order (remember, "I'll take a pair in every color") was sent out in three shipments. Those shipments went out on 20Feb2007, 20Apr2007, and 20May2007.

Below is a graphic of the order and shipping files.



Orders File

Name	Customer Number	Customer Order Number	Status
Russ	003	5	Filled
Susan	005	120	Filled
Susan	005	121	Canceled
Susan	005	122	Split-
Susan	005	123	Canceled
Susan	005	124	Filled

Shipping File

Name	Customer Number	Customer Shipping Number	Ship Date
Russ	003	5	10Jan2007
Susan	005	90	10Jan2007
Susan	005	91	20Feb2007
Susan	005	92	10Mar2007
Susan	005	93	20Apr2007
Susan	005	94	20May2007

Figure 1

Figure 2 is a graphic of a RELREC file for the files in Figure 1

RELREC File

RDOMAIN or File	USUBJID or Customer Number	IDVAR This is included because variables do not have to have the same names in source files.	IDVARVAL or Value of the variable as it is in the original file	RELID or Relationship Number
Orders	**003**	**Customer Order Number**	**5**	**Relationship 1**
Shipping	**003**	Customer Shipping Number	**5**	**Relationship 1**
Orders	**005**	**Customer Order Number**	**120**	**Relationship 2**
Shipping	**005**	Customer Shipping Number	**90**	**Relationship 2**
Orders	**005**	**Customer Order Number**	**122**	**Relationship 3**
Shipping	**005**	Customer Shipping Number	**91**	**Relationship 3**
Shipping	**005**	Customer Shipping Number	**93**	**Relationship 3**
Shipping	**005**	Customer Shipping Number	**94**	**Relationship 3**
Orders	**005**	**Customer Order Number**	**124**	**Relationship 4**
Shipping	**005**	Customer Shipping Number	**92**	**Relationship 4**

Figure 2

The fourth column (from the left) is called IDVARVAL and bears further discussion. In our simple example, this one column contains information from two different variables. When the row in RELREC contains information from the Orders file, IDVARVAL is a value from the Customer Order Number variable. When the row contains information from the Shipping file, IDVARVAL contains a value from the Customer Shipping Number variable.

The linking of rows in RELREC is indicated by a common value of the variable RELID (or Relationship Number). Rows with the same value of RELID are related. Records in RELREC "are pointers" to records on other domains and through this "pointer" mechanism we establish relationships among observations in other domains.

Example #2 – A Example involving a Clinical Trial

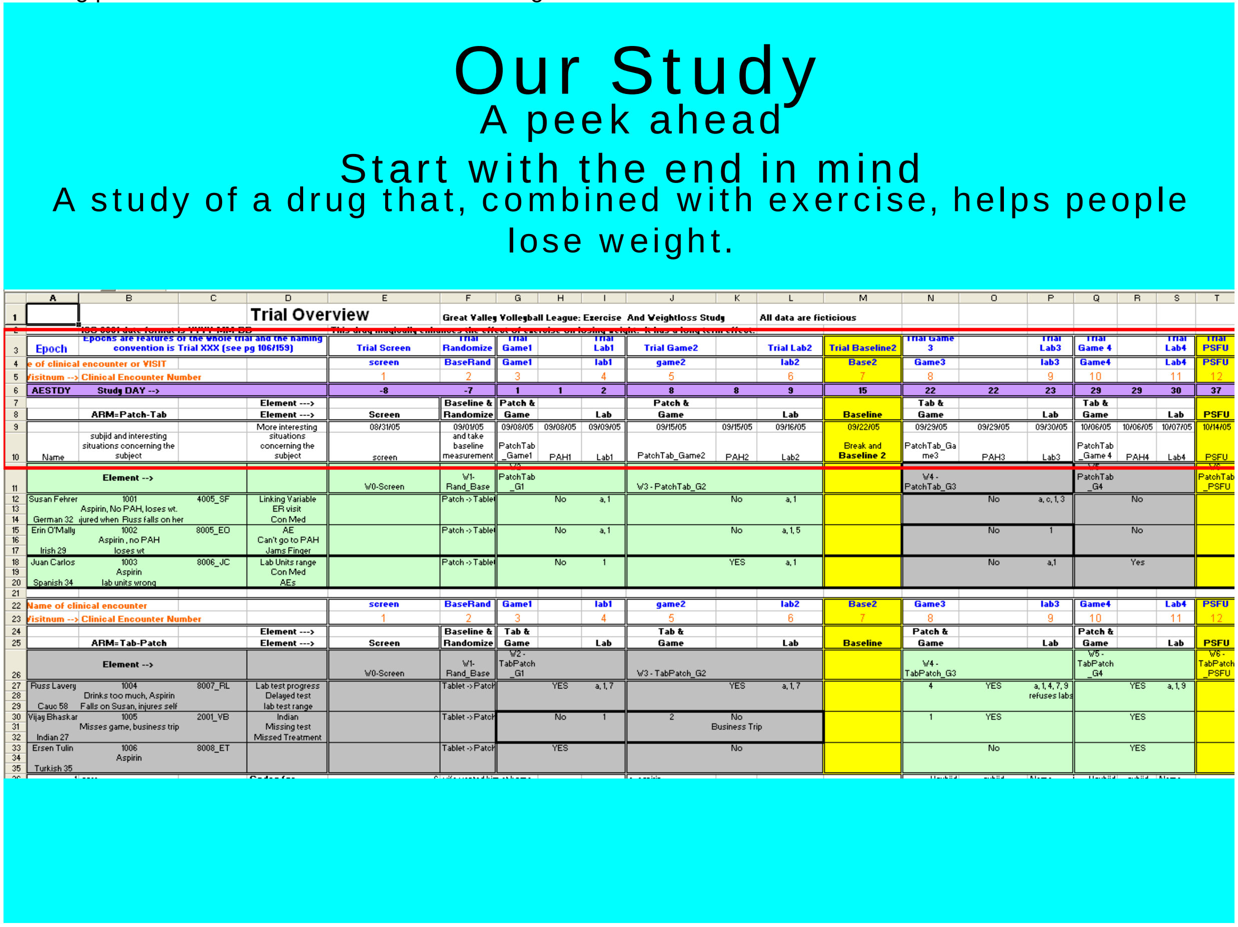
The next example uses RELREC in a clinical setting to link adverse events and concomitant medications(CM) taken for those Adverse Events (AEs).

The next few examples in this paper will be developed from an imaginary clinical study of a new weight loss drug. The drug can be administered via "skin-contact patch" or via an orally ingested tablet. This compound, hopefully, will have a maximum effect if it is part of a regimen that includes moderate exercise (in our study, playing in a volleyball league). Quirks in the process of recruiting resulted in all of the subjects being from the same geographic area and all subjects competing in the same volleyball league. After the game, most subjects go directly from the gym to the local "Schutzen Hall" to drink, brag, review the game, and talk about life with their friends.

Common adverse events associated with this study are hangovers from excessive alcohol consumption and minor strains. In this example, we will focus on and record the following AEs: subject Russ has several hangovers, sprains his ankle, and falls into subject Susan injuring her knee.

Below is a graphic of our six subject crossover study. Three subjects are assigned to the arm Tablet -> Patch and three to the Patch -> Tablet arm. Subjects play two games with one treatment modality and then crossover and play two games with the other treatment modality.

The two files to be "related" (AE and CM) have several variables with different prefixes and the same suffixes. The differing prefixes for variables with common endings are indicated with a "- -".



Our Study

A peek ahead

Start with the end in mind

A study of a drug that, combined with exercise, helps people lose weight.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1				Trial Overview		Great Valley Volleyball League: Exercise And Weightloss Study						All data are ficticious								
2		ISO 8601 date format is YYYY-MM-DD			This drug magically enhances the effect of exercise on losing weight. It has a long term effect.															
3	Epoch	Epochs are features of the whole trial and the naming convention is Trial XXX (see pg 106/159)			Trial Screen	Trial Randomize	Trial Game1		Trial Lab1	Trial Game2		Trial Lab2	Trial Baseline2	Trial Game 3		Trial Lab3	Trial Game 4		Trial Lab4	Trial PSFU
4	e of clinical encounter or VISIT				screen	BaseRand	Game1		lab1	game2		lab2	Base2	Game3		lab3	Game4		Lab4	PSFU
5	Visitnum -->	Clinical Encounter Number			1	2	3		4	5		6	7	8		9	10		11	12
6	AESTDY	Study DAY -->			-8	-7	1	1	2	8	8	9	15	22	22	23	29	29	30	37
7				Element --->		Baseline &	Patch &			Patch &				Tab &			Tab &			
8		ARM=Patch-Tab		Element --->	Screen	Randomize	Game		Lab	Game		Lab	Baseline	Game		Lab	Game		Lab	PSFU
9					08/31/05	09/01/05	09/08/05	09/08/05	09/09/05	09/15/05	09/15/05	09/16/05	09/22/05	09/29/05	09/29/05	09/30/05	10/06/05	10/06/05	10/07/05	10/14/05
10	Name	subjid and interesting situations concerning the subject		More interesting situations concerning the subject	screen	and take baseline measurement	PatchTab_Game1	PAH1	Lab1	PatchTab_Game2	PAH2	Lab2	Break and Baseline 2	PatchTab_Game3	PAH3	Lab3	PatchTab_Game 4	PAH4	Lab4	PSFU
11		Element -->			W0-Screen	W1-Rand_Base	W2 PatchTab_G1			W3 - PatchTab_G2				W4 - PatchTab_G3			W5 PatchTab_G4			W6 PatchTab_PSFU
12	Susan Fehrer	1001	4005_SF	Linking Variable		Patch -> Table		No	a, 1		No	a, 1			No	a, c, 1, 3		No		
13		Aspirin, No PAH, loses wt.		ER visit																
14	German 32	ijured when Russ falls on her		Con Med																
15	Erin O'Mally	1002	8005_EO	AE		Patch -> Table		No	a, 1		No	a, 1, 5			No	1		No		
16		Aspirin , no PAH		Can't go to PAH																
17	Irish 29	loses wt		Jams Finger																
18	Juan Carlos	1003	8006_JC	Lab Units range		Patch -> Table		No	1		YES	a, 1			No	a,1		Yes		
19		Aspirin		Con Med																
20	Spanish 34	lab units wrong		AEs																
21																				
22	Name of clinical encounter				screen	BaseRand	Game1		lab1	game2		lab2	Base2	Game3		lab3	Game4		Lab4	PSFU
23	Visitnum -->	Clinical Encounter Number			1	2	3		4	5		6	7	8		9	10		11	12
24				Element --->		Baseline &	Tab &			Tab &				Patch &			Patch &			
25		ARM=Tab-Patch		Element --->	Screen	Randomize	Game		Lab	Game		Lab	Baseline	Game		Lab	Game		Lab	PSFU
26		Element -->			W0-Screen	W1-Rand_Base	W2 - TabPatch_G1			W3 - TabPatch_G2				W4 - TabPatch_G3			W5 - TabPatch_G4			W6 - TabPatch_PSFU
27	Russ Lavery	1004	8007_RL	Lab test progress		Tablet -> Patch		YES	a, 1, 7		YES	a, 1, 7		4	YES	a, 1, 4, 7, 9		YES	a, 1, 9	
28		Drinks too much, Aspirin		Delayed test												refuses labs				
29	Cauc 58	Falls on Susan, injures self		lab test range																
30	Vijay Bhaskar	1005	2001_VB	Indian		Tablet -> Patch		No	1	2	No			1	YES			YES		
31		Misses game, business trip		Missing test							Business Trip									
32	Indian 27			Missed Treatment																
33	Ersen Tulin	1006	8008_ET			Tablet -> Patch		YES			No				No			YES		
34		Aspirin																		
35	Turkish 35																			

Figure 3 (shown larger in appendix)

Relationships within and among data sets are implemented with the following variables:

STUDYID – A unique study identifier. A submission, typically, will have many studies.
DOMAIN – The two character abbreviation for the domain (eg., AE,CM, DM).
USUBJID – A unique subject identifier
--SEQ (AESEQ, CMSEQ) – A sequence number to insure uniqueness within a data set.
--SPID (AESPID, CMSPID) –An optional sponsor defined reference number, perhaps printed on the CRF or in the sponsor's operational database.
--CAT (AECAT, CMCAT) – Defines a category of records (establishes links across subjects)
--SCAT (AESCAT,CMSCAT) – Defines a sub-category of records (establishes links across subjects)
--GRPID (AEGRPID, CMGRPID) - Used to tie together a block of records *within a subject*

In the relating file (RELREC):

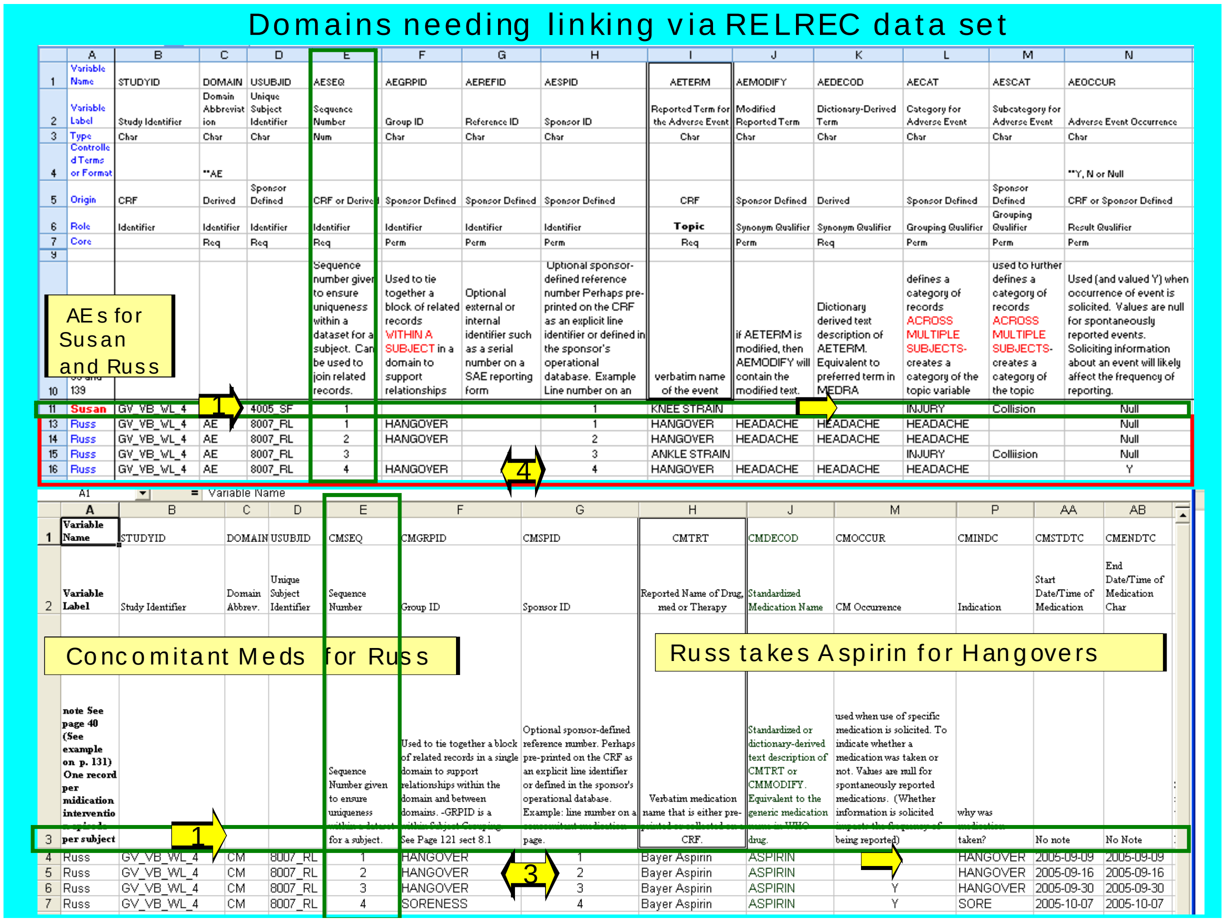
STUDYID - Unique study Identifier. A submission, typically, will have many studies.
RDOMAIN – Related Domain (in two character abbreviated format)
USUBJID – A unique subject identifier
IDVAR – This tells us to look at this variable in the domain mentioned above.
IDVARVAL - This tells us to look for this value in the ID variable in the domain mentioned above.
RELTYPE - This is used in linking whole data sets and discussion is beyond the scope of this paper.
RELID – The linking variable - all records that have the same RELID value are related.

It is important to discuss the context of the links before discussing the links themselves. Subject Russ drinks too much and is uncoordinated; he is sports challenged, if one is being politically correct.

In the AE Domain we see subject Russ has several hangovers on: 09-09-2005 after Game 1, 16-09-2005 after Game 2, and on 30-09-2005 after game 3. In Game 3 he stumbles and falls when he strains his ankle. When he falls, he falls into subject Susan causing her to strain her knee. Subject Susan's injury is displayed in the AE table. Our primary task will be to use the RELREC domain to link the AEs to the medicines recorded in the CM (Concomitant Medications) Domain.

AE Domain: Overall Variables – Key variables for linking files
STUDYID identifies the study to which this data set belongs. A submission will likely have many studies and all studies will likely have AE and CM domains. The "AE" value in the DOMAIN variable identifies this file as the AE file in a manner that is internal to the file, as opposed to identifying the file via the filename. This means that the domain information saying that a row is a row in AE can easily be used in computerized filtering and merging of files. The USUBJID variable establishes that a row contains data for *that particular* subject identified by the USUBJID. AESEQ is a variable that is a unique identifier, a synthetic key, *within a subject-domain* and is not unique within the domain as a whole as all subjects with an AE will have an AESEQ of 1. The combination of STUDYID, DOMAIN, USUBJID, and AESEQ uniquely identifies a row in the AE Domain of a particular study ***and creates a key***. This key identifies a unique piece of information, a row of data, for the submission. Within the study, in the AE Domain and Subject 8007_RL, AESEQ having a value of 1 is unique.

Since AESEQ is a synthetic key assigned inside USUBJID, subject 8007_RL can have an AESEQ=1 and subject 4005_SF can also have an AESEQ=1. It is likely, but not required, that the AESEQ variable could be used to order observations for a subject in order of increasing time. AESPID is an AE link to a sponsor defined ID number, likely a number on an AE page. It is possible but not required that sorting by this variable will sort a subject's AEs in order of ascending date and time.



Domains needing linking via RELREC data set

AEs for Susan and Russ

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Variable Name	STUDYID	DOMAIN	USUBJID	AESEQ	AEGRPID	AEREFID	AESPID	AETERM	AEMODIFY	AEDECOD	AECAT	AESCAT	AEOCCUR
2	Variable Label	Study Identifier	Domain Abbreviation	Unique Subject Identifier	Sequence Number	Group ID	Reference ID	Sponsor ID	Reported Term for the Adverse Event	Modified Reported Term	Dictionary-Derived Term	Category for Adverse Event	Subcategory for Adverse Event	Adverse Event Occurrence
3	Type	Char	Char	Char	Num	Char	Char	Char	Char	Char	Char	Char	Char	Char
4	Controlled Terms or Format		**AE											**Y, N or Null
5	Origin	CRF	Derived	Sponsor Defined	CRF or Derived	Sponsor Defined	Sponsor Defined	Sponsor Defined	CRF	Sponsor Defined	Derived	Sponsor Defined	Sponsor Defined	CRF or Sponsor Defined
6	Role	Identifier	Identifier	Identifier	Identifier	Identifier	Identifier	Identifier	**Topic**	Synonym Qualifier	Synonym Qualifier	Grouping Qualifier	Grouping Qualifier	Result Qualifier
7	Core		Req	Req	Req	Perm	Perm	Perm	Req	Perm	Req	Perm	Perm	Perm
9														
10	139				Sequence number given to ensure uniqueness within a dataset for a subject. Can be used to join related records.	Used to tie together a block of related records WITHIN A SUBJECT in a domain to support relationships	Optional external or internal identifier such as a serial number on a SAE reporting form	Optional sponsor-defined reference number Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database. Example Line number on an	verbatim name of the event	if AETERM is modified, then AEMODIFY will contain the modified text.	Dictionary derived text description of AETERM. Equivalent to preferred term in MEDRA	defines a category of records ACROSS MULTIPLE SUBJECTS- creates a category of the topic variable	used to further defines a category of records ACROSS MULTIPLE SUBJECTS- creates a category of the topic	Used (and valued Y) when occurrence of event is solicited. Values are null for spontaneously reported events. Soliciting information about an event will likely affect the frequency of reporting.
11	Susan	GV_VB_WL_4		4005_SF	1			1	KNEE STRAIN			INJURY	Collision	Null
13	Russ	GV_VB_WL_4	AE	8007_RL	1	HANGOVER		1	HANGOVER	HEADACHE	HEADACHE	HEADACHE		Null
14	Russ	GV_VB_WL_4	AE	8007_RL	2	HANGOVER		2	HANGOVER	HEADACHE	HEADACHE	HEADACHE		Null
15	Russ	GV_VB_WL_4	AE	8007_RL	3			3	ANKLE STRAIN			INJURY	Colliision	Null
16	Russ	GV_VB_WL_4	AE	8007_RL	4	HANGOVER		4	HANGOVER	HEADACHE	HEADACHE	HEADACHE		Y

A1 = Variable Name

Concomitant Meds for Russ

Russ takes Aspirin for Hangovers

	A	B	C	D	E	F	G	H	J	M	P	AA	AB
1	Variable Name	STUDYID	DOMAIN	USUBJID	CMSEQ	CMGRPID	CMSPID	CMTRT	CMDECOD	CMOCCUR	CMINDC	CMSTDTC	CMENDTC
2	Variable Label	Study Identifier	Domain Abbrev.	Unique Subject Identifier	Sequence Number	Group ID	Sponsor ID	Reported Name of Drug, med or Therapy	Standardized Medication Name	CM Occurrence	Indication	Start Date/Time of Medication	End Date/Time of Medication Char
3	note See page 40 (See example on p. 131) One record per midication intervention episode per subject				Sequence Number given to ensure uniqueness within a dataset for a subject.	Used to tie together a block of related records in a single domain to support relationships within the domain and between domains. -GRPID is a within Subject Grouping See Page 121 sect 8.1	Optional sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database. Example: line number on a concomitant medication page.	Verbatim medication name that is either pre-printed or collected on a CRF.	Standardized or dictionary-derived text description of CMTRT or CMMODIFY. Equivalent to the generic medication name in WHO drug.	used when use of specific medication is solicited. To indicate whether a medication was taken or not. Values are null for spontaneously reported medications. (Whether information is solicited impacts the frequency of being reported)	why was medication taken?	No note	No Note
4	Russ	GV_VB_WL_4	CM	8007_RL	1	HANGOVER	1	Bayer Aspirin	ASPIRIN		HANGOVER	2005-09-09	2005-09-09
5	Russ	GV_VB_WL_4	CM	8007_RL	2	HANGOVER	2	Bayer Aspirin	ASPIRIN		HANGOVER	2005-09-16	2005-09-16
6	Russ	GV_VB_WL_4	CM	8007_RL	3	HANGOVER	3	Bayer Aspirin	ASPIRIN	Y	HANGOVER	2005-09-30	2005-09-30
7	Russ	GV_VB_WL_4	CM	8007_RL	4	SORENESS	4	Bayer Aspirin	ASPIRIN	Y	SORE	2005-10-07	2005-10-07

Figure 4

AE Domain: Timing Variables

Timing variables are in many domains are related to two important timing variables in the DM (Demographic) Domain. These two important timing variables in DM are RFSTDTC and RFENDTC and are formatted as ISO 8601 date-time. They contain information as to when a subject enters and leaves the study.

The variable AESTDTC is formatted as an ISO 8601 value and indicates the date-time when the AE started. The variable AEENDTC is also formatted as an ISO 8601 value and indicates the date-time when the AE ended. AEENDTC should be greater than or equal to AESTDTC. ISO 8601 formatting can hold differing levels of precision if the subject does not remember the month, week, day, or time of the event (see paper by Shi-Tao Yeh for more coverage and note that the IG v3.1.2 has an expanded section on the ISO 8601 standard). AEDUR contains the duration of the event, ONLY if it is recorded on the CRF. If the duration of the event is calculated from the start and end dates for the event, AEDUR should not have a value. Likewise, if the CRF has values for AESTDTC and AEENDTC then AEDUR should not have a value on the CRF. AEDUR has a value only if an AE duration is collected on the CRF.

AESTDY and AEENDY are measured relative study start day for the study RFSTDTC. Variables ending with - -DY are sometimes called “study day variables”. Study day is a measure of how long the subject has been participating in the study and CANNOT have a value of zero. Study day variables for AEs can be calculated with the following formulas. All study day variables are calculated reference to the RFSTDTC variable in the DM Domain and have the same logic. The formulas use the date part of the date-time values.

AE**ST**DY= AE start study day = (AE**ST**DTC – RFSTDTC) + 1 (when AE**ST**DTC GE RFSTDTC);
AE**ST**DY= AE start study day = (AE**ST**DTC – RFSTDTC) + 0 (when AE**ST**DTC LT RFSTDTC);

AE**EN**DY= AE end study day = (AE**EN**DTC – RFSTDTC) + 1 (when AE**ST**DTC GE RFSTDTC);
AE**EN**DY= AE end study day = (AE**EN**DTC – RFSTDTC) + 0 (when AE**ST**DTC LT RFSTDTC);

AESTRF and AEENRF are for recording imprecise memories. AESTRF should not be used if the investigator was able to assign a start date to the AE and record that start date in AESTDTC. AEENRF should not be used if the investigator was able to assign an end date to the AE and record that end date in AEENDTC. Remember that ISO 8601 can record varying levels of precision, so the subject need not be able to recall a day, or month. AESTRF and AEENRF variables simply record if the AE happened BEFORE, DURING, or AFTER the subject's period of participation in the study as defined by the RFSTDTC and RFENDTC variables in the DM Domain.

This ability to make gross (before / during / after) assignments to time periods has been expanded in the IG v3.1.2. New start point and end point variables have been added to the domain so that an investigator may classify an event as before, during, coincident or after time points that are internal to the study, eg., a subject's business trip or vacation. The IG has rules for entering values into these variables that sound complex but are really basic logic. The complexity comes from trying to explain if "AFTER" is an allowed value. Please see Table 1 below.

	--relationship between reference time point and date of data collection--	
	Reference time point IS THE date of data collection	Reference time point BEFORE date of data collection
Allowable values of Start→	BEFORE, COINCIDENT, UNKNOWN (AFTER is not allowed since one cannot know the future)	BEFORE, COINCIDENT, UNKNOWN, *AFTER is allowed*
Allowable values of End→	BEFORE, COINCIDENT, ONGOING, UNKNOWN (AFTER not allowed since one cannot know the future)	BEFORE, COINCIDENT, UNKNOWN, *AFTER is allowed*

Table 1

AE Domain: Grouping Variables

- -GRPID, - -CAT, - -SCAT are variables that establish relationships within a domain. - -GRPID is used to establish relationships within a subject. - -CAT, - -SCAT are used to establish relationships involving several subjects. The relationship among - -CAT, - -SCAT, and - -GRPID can be seen in Figures 4 and 5. These variables allow easy grouping or selecting of rows that belong to one subject and also allow easy grouping or selecting of rows that belong to multiple subjects. The AEGRPID is valued to allow easy selection, collapsing, or grouping of records within a subject. Think of - -CAT as "Category across Subjects" and - -SCAT as "Sub-Category within Category which is across Subjects".

There are characteristics for these variables that help in remembering the use of these variables. Three variables are being used to establish two kinds of groupings, within versus between subject groupings. - -CAT and - -SCAT, by the similarity in their names, are easy to remember as a pair of variables that work together, leaving GRPID to be used in establishing the other type of grouping.

It is unlikely that any one subject will have so many AEs as to need a very complicated, grouping and sub-grouping, scheme. But if we wanted to create groupings for *all* the AEs for *all* the subjects in the whole domain, we might want a complicated and powerful grouping scheme. - -GRPID, being only one variable, will allow the creation of simple grouping and is sufficient for grouping rows within a subject. The two variables - -CAT and - -SCAT together will allow the creation of more complex grouping schemes and are used to establish the complex groupings needed for across subject analysis.

In Figure 4 we can see examples of - -CAT, - -SCAT, and - -GRPID variables used in two data sets, AE and CM. We will use Figure 4 to examine the use of - -GRPID, - -CAT, and - -SCAT in both domains. We should remember that due to the trips to the Schutzen Hall, hangovers would be recorded by many subjects.

In AE, if an investigator wanted to pull all records associated with headaches for all subjects, s/he could code where upcase(AECAT) = 'HEADACHE' ; likewise, if an investigator wanted to pull all records associated with collisions for all subjects, s/he could code where upcase(AESCAT) = 'COLLISION' ;.

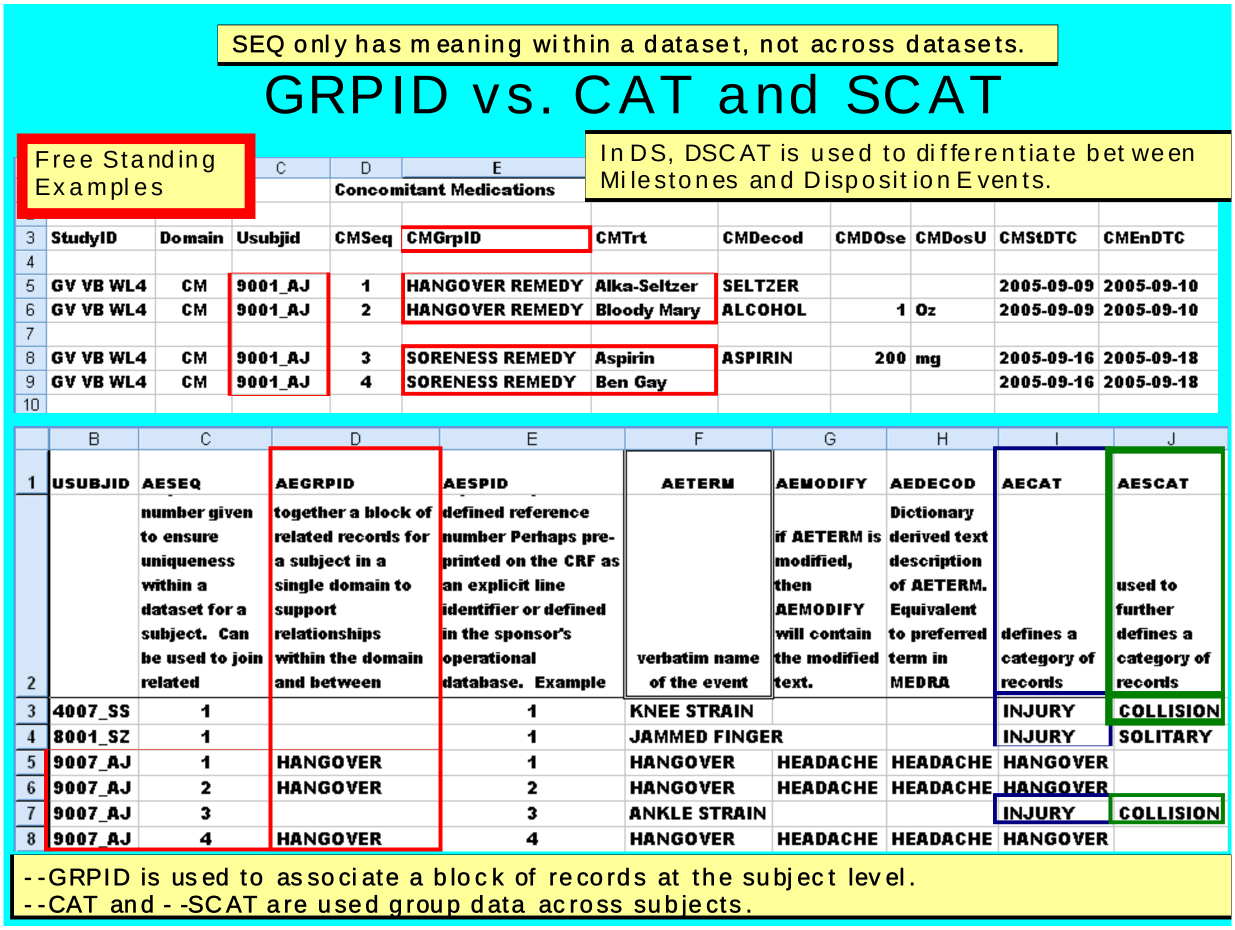
In a cursory examination of Figure 4 it appears that the variable AECAT is valued in a manner very similar to that of AEGRPID, but we must remember AEGRPID groups within subjects and AECAT groups across subjects. Subject Russ recorded four AEs and not all four were alcohol related. In the AE domain, if an investigator wanted to pull all records associated with hangovers for subject Russ, s/he could code

```
        where upcase(AEGRPID) = 'HANGOVER' and USUBJID="8007_RL" ;.
```

- -CAT and - -SCAT are used to group and sub-group rows in a manner that is consistent **across** subjects. AECAT is valued as INJURY for subject Russ and also for subject Susan. This identifies these AEs as injuries. The coding of "INJURY" in AECAT for subjects Russ (injury=ankle strain) and Susan (injury=knee strain) would allow an investigator to easily pull all AEs that were injuries. It would have been nice, if space had permitted, to include an observation that demonstrated an injury being linked to another subject. Inside the category of injury we can see AESCAT identifies the injuries to Susan and Russ to be a collision.

While poor practice, values of - -GRPID can be used with different meanings if subjects differ. This means that gout, for subject Alexander, could be coded as “toe pain” and a broken toe, for subject Sam, could also be coded as “toe pain”. - -GRPID is for within subject groupings and only meaningful within one subject.

In the bottom half of Figure 4, we see that Russ recorded four concomitant medications. An error was intentionally included in this part of the graphic to facilitate discussion. The CMGRPID for the fourth medication is recorded as soreness. Since there are no other rows coded as soreness, this value does not group rows and should be blank. The fact that this aspirin use was for soreness is recorded by the value “SORE” in CMINDC (concomitant medication indication). The CM Domain has CMCAT, CMSCAT and CMGRPID variables and, as did the AE Domain, they are used to group values of the topic variable. They are not shown in this graphic due to lack of space.



SEQ only has meaning within a dataset, not across datasets.

GRPID vs. CAT and SCAT

Free Standing Examples

In DS, DSCAT is used to differentiate between Milestones and Disposition Events.

Concomitant Medications

	StudyID	Domain	Usubjid	CMSeq	CMGrpID	CMTrt	CMDecod	CMDOse	CMDosU	CMStDTC	CMEnDTC
4											
5	GV VB WL4	CM	9001_AJ	1	HANGOVER REMEDY	Alka-Seltzer	SELTZER			2005-09-09	2005-09-10
6	GV VB WL4	CM	9001_AJ	2	HANGOVER REMEDY	Bloody Mary	ALCOHOL	1	Oz	2005-09-09	2005-09-10
7											
8	GV VB WL4	CM	9001_AJ	3	SORENESS REMEDY	Aspirin	ASPIRIN	200	mg	2005-09-16	2005-09-18
9	GV VB WL4	CM	9001_AJ	4	SORENESS REMEDY	Ben Gay				2005-09-16	2005-09-18
10											

	B	C	D	E	F	G	H	I	J
1	USUBJID	AESEQ	AEGRPID	AESPID	AETERM	AEMODIFY	AEDECOD	AECAT	AESCAT
2		number given to ensure uniqueness within a dataset for a subject. Can be used to join related	together a block of related records for a subject in a single domain to support relationships within the domain and between	defined reference number Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database. Example	verbatim name of the event	if AETERM is modified, then AEMODIFY will contain the modified text.	Dictionary derived text description of AETERM. Equivalent to preferred term in MEDRA	defines a category of records	used to further defines a category of records
3	4007_SS	1		1	KNEE STRAIN			INJURY	COLLISION
4	8001_SZ	1		1	JAMMED FINGER			INJURY	SOLITARY
5	9007_AJ	1	HANGOVER	1	HANGOVER	HEADACHE	HEADACHE	HANGOVER	
6	9007_AJ	2	HANGOVER	2	HANGOVER	HEADACHE	HEADACHE	HANGOVER	
7	9007_AJ	3		3	ANKLE STRAIN			INJURY	COLLISION
8	9007_AJ	4	HANGOVER	4	HANGOVER	HEADACHE	HEADACHE	HANGOVER	

--GRPID is used to associate a block of records at the subject level.
--CAT and - -SCAT are used group data across subjects.

Figure 5

Figure 5 is a “free standing example”, not part of the study we have been discussing and was inserted to illustrate the variables currently under discussion. The relationship among - -CAT, - -SCAT, and - -GRPID can be seen in Figure 5.

CM shows that on 10-09-2005, subject AJ took two medications for a hangover. On 18-09-2005 the CM Domain shows he took two medications for soreness. In Figure 5 we see how CMGRPID was used in the CM Domain to collect the differing medications into groups that are or might be of interest to an investigator.

In the AE Domain, we see subject A.J. has four AEs. Subject AJ had three hangovers and they are grouped in AEGRPID. In the AE Domain numbers one, two, and four were grouped as hangovers in AEGRPID. The third AE was an ankle sprain.

Cross subject grouping is done using the variables - -CAT and - -SCAT. AECAT creates subdivisions of the topic variable. There were three sports injuries and a programmer can “pull” all injuries by coding “where upcase(AECAT) =’INJURY’;” AECAT is valued as “INJURY” for three subjects, 4007_SS, 8001_SZ and 9007_AJ. Using the variable AESCAT, injuries were subdivided into collisions between players and solitary injuries.

Coding where USUBJID = 9007_AJ and AEGRPID = ”HANGOVER” will pull all records for hangovers for subject AJ While poor practice, values of - -GRPID can be re-used *with different meanings* if subjects differ. This means that gout, for subject Alexander, could be coded as “toe pain” and a broken toe, for subject Sam, could also be coded as “toe pain”. - -GRPID is valued within subject and only meaningful within records for one subject.

CMINDC (not shown) stores the value for the indication of the medicine or why the medicine was taken. If the medicine was taken for an AE, the value of CMINDC should match the value of AETERM, AEMODIFY, or AEDECOD.

In Figure 5 we see that subject AJ had AEs for hangovers and took medications for hangovers. It would be useful be able to link the AE and the medication taken for the AE. This will be done with the RELREC domain.

AE Domain: Topic Variables
The **AE domain** has the typical relationship among the variables - -TERM, - -MODIFY and –DECOD(e). AETERM records the verbatim for the event. If the sponsor decides to recode the verbatim term to a set of values that are particular to that sponsor, the sponsor’s decode would be entered into AEMODIFY. The final variable in this triple relationship is AEDECOD. This variable contains a decode from a standard dictionary. It is equivalent to the preferred tem in MedDRA. The dictionary used for decoding and the version of the dictionary, should be entered in the Data Definition Document in the AE Domain.

Other relationships
If the variable AESER (“AE Serious”) is valued as “Y”, there should be a record of why the AE was classified as serious. If AESER is “Y” there should be a “Y” value in one or more of the following variables: AESCONG, AESDISAB, AESDTH, AESHOSP, AESLIFE, et al.

When the variable Other Medically Important Serious Medical Event (AESMIE) is valued “Y”, it means that other categories are needed to describe why the event was judged serious. That reason should be included in SUPPQUAL or SUPPAE.

Adverse Events can be collected in two ways:

1) by recording freely reported (unprompted) verbatim comments
2) by checking boxes from a pre-printed list on the CRF (reporting can be prompted or not). If the investigator reads the pre-printed list TO the subject and asks if the subject has experienced a particular AE, AEOCCUR should be valued as “Y”. Asking if an event occurs increases the frequency of reporting and that fact must be registered.

In the CDISC STDM IG v3.1.2 we see a new variable AEPRESP (pre-specified). This variable is valued as “Y” if the AE was pre-printed on the CRF. Values of this variable will be related with AEOCCUR (which records *should* information about the event *solicited*), AESTAT (was the question asked), and AEREASND (reason why the question was not asked). These variables are logically related.

AESEV (severity or intensity of the event) and AETOXGR (a toxicity grade according to a standard toxicity scale, such as CTCAE) should not both have values. If a standard toxicity grade is specified, then the study documentation should include the scale and version used. This information should be entered in the Data Definition Document in the AE Domain.

CM Domain: Overall Variables – Key variables for linking files
The CDISC group created the CM (Concomitant Medications) Domain to hold information about non-study medications or therapies. This domain often contains information about a subject’s background treatment for non-study related medial conditions. This domain may also be used to record medications for intermittent medical problems.

STUDYID identifies the study to which this data set belongs. A submission will likely have many studies and all studies will likely have AE and CM Domains. The value “CM” in the DOMAIN variable identifies this file as the CM file, in a manner that is internal to the file (as opposed to identifying the file via the filename). This means that the information (that this is a row in CM) can easily be used in computerized filtering and merging of files. The USUBJID variable establishes that a row contains data for *that particular* subject. CMSEQ is a variable that is a unique identifier *within a subject* (and not within the domain – all subjects with a concomitant medication will have a CMSEQ of 1). The combination of STUDYID, DOMAIN, USUBJID, and CMSEQ uniquely identifies a row in a particular study’s CM Domain ***and creates a key***. This key identifies a unique piece of information (row of data) for the submission. Within the study, in the CM Domain, USUBJID=8007_RL, CMSEQ=1 is unique.

USUBJID 8007_RL can have a CMSEQ=3 and USUBJID 4005_SF can also have a CMSEQ=3. It is likely, but not required, that the CMSEQ variable would order observations for a subject by ascending date and time.

CM Domain: Timing Variables

Timing variables in many domains are related to two timing variables in the DM (Demographic) Domain. The two timing variables are RFSTDTC and RFENDTC in the DM Domain hold the date and time when a subject enters and leaves the study.

The variable CMSTDTC holds an ISO 8601 formatted value that indicates the date-time when the subject started taking the medication. The variable CMENDTC holds an ISO 8601 formatted value that indicates the date-time when the medication was discontinued. CMENDTC should be greater or equal to CMSTDTC. ISO 8601 can hold differing levels of precision if the subject does not remember the month, week, day, or time of the event (see paper by Shi Tao Yeh for more coverage). CMDUR holds the duration of the event, ONLY if it is recorded on the CRF. If the duration of the event is calculated from the start and end dates for the event, CMDUR should not have a value. If the CRF has values for CMSTDTC and CMENDTC, then CMDUR should not have a value on the CRF.

CMSTDY and CMENDY hold the study day for the start and end of the concomitant medication. Variables ending with - -DY are sometimes called “study day variables”. A study day is a measure of how long the subject has been participating in the study and CANNOT be valued as zero. Study day variables (for CMs) can be calculated with the following formulas. All study day variables are calculated reference to the RFSTDTC variable in the DM Domain and have the same logic. The formulas for CM are slight modifications of the formulas for AEs. All study day calculations use the same logic.

CMSTRF and CMENRF are for recording imprecise memories. CMSTRF should not be used if the investigator was able to assign a start to the CM and record that start date-time in CMSTDTC. CMENRF should not be used if the investigator was able to assign an end to the CM and record that end date-time in CMENDTC. (remember ISO 8601 can record varying levels of precision, so the subject need not be able to recall a day, or month). CMSTRF and CMENRF variables simply record if the CM happened BEFORE, DURING, or AFTER the subject’s period of participation in the study, as defined by RFSTDTC and RFENDTC in the DM Domain.

There are new timing variables in the IG v3.1.2. They are an expansion of the concept above. It can be useful to have more than the one reference period established by the variables RFSTDTC and RFENDTC in the DM Domain. CMSTTPT (Concomitant Medications reference period STart Time PoinT) and CMENTPT (Concomitant Medications reference period ENd Time PoinT) allow the creation of additional reference periods. CMSTRTPT (Concomitant Medication Start relative to Starting Time PoinT) and CMENRTPT (Concomitant Medication ENd Relative to ending Time PoinT). Valuing these variables uses the same logic as was described in Table 1. Please the Table1.

Dosing Relationships

Several variables, CMDOSE, CMDOSTXT, CMDOSU, CMDOSFRM, CMDOSFRQ, CMDOSTOT, CMDOSRGM, CMROUTE, are required to provide sufficient detail concerning a dose. Only by considering the combination of several variables can one understand how much medication was taken.

- CMDOSE is the numeric, unit-less amount of the medication taken. This variable might have a value of 200.
- CMDOSTXT is the character version of the amount or range of dosing. It can be a range like 200-400.
- CMDOSU contains the units for the dose. If CMDOSE is 200, CMDOSU might be mg.
- CMDOSFRM is the form of the dose. The value might be TAB or GELTAB.
- CMDOSFRQ is the ACTUAL frequency of dosing. It might be “BID”. The value should include the frequency and the time period.
- CMDOSTOT is the total daily dose, using the units in CMDOSU. CMDOSTOT is just a number without units. The units for the number are found in CMDOSU.
- CMROUTE contains the route of administration. A value could be “ORAL”.
- CMDOSRGM is valued as the intended schedule for the medication. The values are similar to those recorded in CMDOSFRQ.

CM Domain: Topic Variables

The CM Domain has the typical relationship among the variables - -TRT, - -MODIFY, and - -DECOD(e). CMTRT records the verbatim term for the medication, either pre-printed or recorded on the CRF. If the sponsor decides to recode the verbatim terms to a set of values that are particular to that sponsor, the decode would be entered into CMMODIFY. The final variable, in this triple relationship, is CMDECOD. This variable contains a decode from a standard dictionary, such as a generic medication name from WHO dictionary.

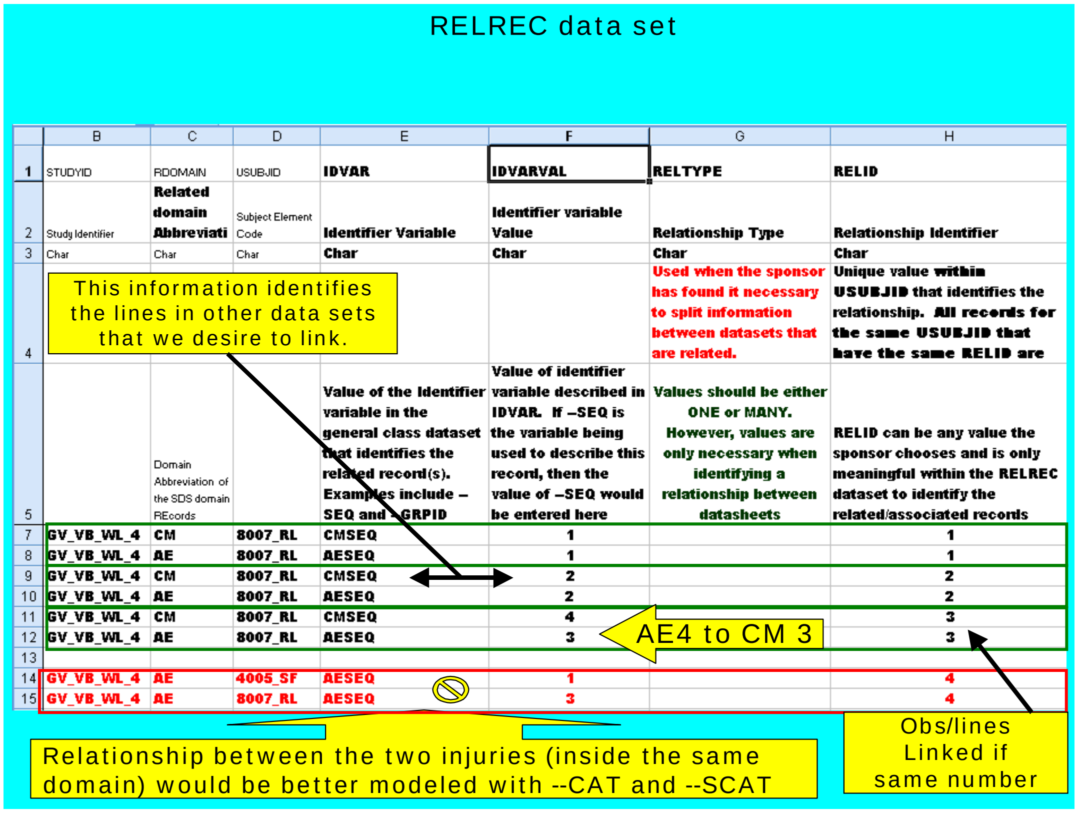
Information about concomitant medications may be collected in two ways:

1) recording freely reported (unprompted) comments verbatim
2) checking boxes for from a list pre-printed on the CRF (reporting can be prompted or not)

If the investigator reads the list TO the subject and asks if the subject has taken a particular medication, CMOCCUR should be valued as “Y”. Asking if a medication was taken increases the frequency of reporting and that fact must be registered.

In the CSISC STDM IG v3.1.2 we see a new variable CMPRESP (pre-specified). This variable is valued as “Y” if the CM was pre-printed on the CRF. Values of this variable will be related with CMOCCUR (was information about the event solicited) and CMSTAT (was the question asked) and CMREASND (reason why the question was not asked).

CM Domain



RELREC data set

	B	C	D	E	F	G	H
1	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID
2	Study Identifier	Related domain Abbreviati	Subject Element Code	Identifier Variable	Identifier variable Value	Relationship Type	Relationship Identifier
3	Char	Char	Char	Char	Char	Char	Char
4						Used when the sponsor has found it necessary to split information between datasets that are related.	Unique value within USUBJID that identifies the relationship. All records for the same USUBJID that have the same RELID are
5		Domain Abbreviation of the SDS domain REcords		Value of the Identifier variable in the general class dataset that identifies the related record(s). Examples include –SEQ and –GRPID	Value of identifier variable described in IDVAR. If –SEQ is the variable being used to describe this record, then the value of –SEQ would be entered here	Values should be either ONE or MANY. However, values are only necessary when identifying a relationship between datasheets	RELID can be any value the sponsor chooses and is only meaningful within the RELREC dataset to identify the related/associated records
7	GV_VB_WL_4	CM	8007_RL	CMSEQ	1		1
8	GV_VB_WL_4	AE	8007_RL	AESEQ	1		1
9	GV_VB_WL_4	CM	8007_RL	CMSEQ	2		2
10	GV_VB_WL_4	AE	8007_RL	AESEQ	2		2
11	GV_VB_WL_4	CM	8007_RL	CMSEQ	4		3
12	GV_VB_WL_4	AE	8007_RL	AESEQ	3		3
13							
14	GV_VB_WL_4	AE	4005_SF	AESEQ	1		4
15	GV_VB_WL_4	AE	8007_RL	AESEQ	3		4

This information identifies the lines in other data sets that we desire to link.

AE4 to CM 3

Relationship between the two injuries (inside the same domain) would be better modeled with --CAT and --SCAT

Obs/lines Linked if same number

Figure 6 – This establishes links to records shown in Figure 4

The important variables in the RELREC domain are RDOMAIN, USUBJID, IDVAR, IDVARVAL, RELTYPE, and RELID. Together these variables make a key that identifies unique rows in other domains. Because of the way values of - -SEQ are allowed to repeat inside domains because they are only unique within a USUBJID and domain. RELREC is used to create relationships between rows for a subject in a domain.

Establishing a relationship between two data sets requires entering at least two rows of data into RELREC. Establishing a relationship among three domains requires entering at least three rows of data in RELREC. The phrase “at least” is used because the relationship being established might be among one row of data from one domain out to several rows of data in another domain. It is easy to imagine this happening. As an example, one AE might “trigger” several lab tests or medications.

Since the number of domains linked and observations linked in a particular domain can vary, a flexible method is required to establish relationships. Entering values into RELID (or relationship ID) is the flexible mechanism by which relationships/links are established. Rows in RELREC with the same value of RELID are considered to be related.

The example Figure 6 shows RELREC being used to establish three relationships between the domains in Figure 4. Figure 6 deliberately includes a counter-example to help clarify the logic of using RELREC.

We will use Figure 6, the RELREC Domain, to illustrate three relationships and we will discuss the relationships in increasing order of RELID.

When RELID is valued as 1, it establishes a relationship between subject Russ’s AE number 1 (on 09-09-2005) and the CM entry valued as 1 (aspirin use on the same day). If the first relationship were translated into words it might be expressed as:

> In the AE Domain (DOMAIN = AE) of study GV_VB_WL_4, Subject 8007RL’s (USUBJID = 8007RL) record that has a value of 1 (IDVARVAL = 1) in the variable AESEQ (IDVAR = AESEQ) is related to a value that is

in the CM Domain (DOMAIN = CM) of study of study GV_VB_WL_4, for Subject 8007RL's (USUBJID = 8007RL) record that has a value of 1 (IDVARVAL = 1) in the variable CMSEQ (IDVAR = CMSEQ)

Remembering that the variables USUBJID and - -SEQ are used to create a unique identifier for a row in a domain. In "data base talk" STUDYID, DOMAIN, USUBJID, and - -SEQ create a unique key for the data set. Since STUDYID, DOMAIN, USUBJID, and - -SEQ create a key the authors expect that the value in IDVAR will most likely be one of the - -SEQ variables.

Note that the dates for related records in the related domains are likely to be 'the same" but this is not a required relationship.

In Figure 6, when RELID is valued as 2, it establishes a relationship between subject Russ's AE number 1 (on 16-09-2005) and the aspirin use on the same day. Note that the dates are likely to be 'the same" but this is not a required relationship.

When RELID is valued as 3, it establishes a relationship that is a bit different from the examples above. In the AE Domain (DOMAIN = AE) Subject 8007RL's (USUBJID = 8007_RL) record that has a value of 4 (IDVARVAL = 4) in the variable AESEQ (IDVAR= AESEQ) is related to a row in the CM Domain (DOMAIN=CM) for Subject 8007_RL (USUBJID = 8007_RL) record that has a value of 3 (IDVARVAL = 3) in the variable CMSEQ (IDVAR = CMSEQ)

The final row in CM records information about the final use of aspirin by Subject Russ (on 10-07-2005). This aspirin use is not related to any recorded AE. As the record shows, subject Russ was asked about his aspirin use (CMOCCUR=Y) and reported both the aspirin use and the muscle soreness that prompted the use. Presumably, because of the mildness of the discomfort, the subject did not report an AE.

The rows in RELREC where RELID is valued as 4 are a deliberate mistake to allow a compare-and-contrast between proper and improper CDISC coding. These rows seem to "establish" a relationship between the two AE records that contain information about subject Russ crashing into and injuring subject Susan. The coding of values is very similar to the coding of the first three examples and *seems* proper until we remember that there is another mechanism for linking observations within a single data set.

It is suggested that the relationship between these two observations in the AE Domain (DOMAIN = AE, USUBJID = 4005_SF, AESEQ = 1, and (DOMAIN = AE, USUBJID = 8007_RL, AESEQ = 3) is better established using AECAT and AESCAT.

SUMMARY

RELREC provides a flexible, if complicated, technique of linking records in different data sets.

CONTACT INFORMATION

Susan Fehrer
BioClin, Inc.
smfehrer@bioclin.net

Russ Lavery
russ@russ-lavery.com