

**THE ADAP FILES:
DATA QUALITY MANAGEMENT FROM A TO Z**

OCTOBER 2004



ARNOLD SCHWARZENEGGER
Governor
State of California

Kimberly Belshé
Secretary
Health and Human Services Agency

Sandra Shewry
Director
Department of Health Services

**THE ADAP FILES:
DATA QUALITY MANAGEMENT FROM A TO Z**

Prepared by:

Dennis T. Wong, Ph.D.

**Department of Health Services
Office of AIDS
HIV/AIDS Epidemiology Branch
<http://www.dhs.ca.gov/AIDS>**

**Kevin Reilly, D.V.M., M.P.V.M.
Deputy Director
Prevention Services**

**Michael Montgomery, Chief
Office of AIDS**

**Juan Ruiz, M.D., Dr.P.H., M.P.H., Chief
HIV/AIDS Epidemiology Branch
Office of AIDS**

October 2004

ACKNOWLEDGMENTS

I would like to thank Susan M. Sabatier and Kathleen Russell for providing valuable feedback on this report.

Correspondence

Please send any questions or comments to Dennis T. Wong: dwong2@dhs.ca.gov.

Suggested Citation

Wong, D.T., *The ADAP Files: Data Quality Management from A to Z*. California Department of Health Services, Office of AIDS, 2004.

TABLE OF CONTENTS

Executive Summary	1
Background	2
The Present Study	3
Method and Results	4
Phase 1	4
Client–Level Data	4
Prescription–Level Data	5
Phase 2	6
Phase 3	6
Discussion	7
References	8

TABLES

Table 1. Client–Level Data, FY 2002–03	9
Table 2. Prescription–Level Data, FY 2002–03	12
Table 3. QM Results for Client–Level Data, FY 2002–03	14
Table 4. QM Results for Prescription–Level Data, FY 2002–03	18
Table 5. Ramsell Corporation’s Weekly Client–Level Data QA Listing	21
Table 6. Ramsell Corporation’s Weekly Prescription–Level Data QA Listing	22

EXECUTIVE SUMMARY

Objectives. The purpose of this study was to examine and evaluate AIDS Drug Assistance Program (ADAP) data files and to develop a mechanism to continually improve the data collection process. Five phases were planned by Office of AIDS (OA) with the first three occurring within this study:

- Phase 1: Screen the entire array of individual client and prescription variables in fiscal year (FY) 2002–03 according to applicable Department of Defense (DoD), Total Data Quality Management (TDQM) standards (completeness, timeliness, uniqueness, and validity).¹
- Phase 2: Determine an acceptable error rate for the data and identify variables that fall below the criteria.
- Phase 3: Review and (if necessary) develop additional quality assurance (QA) listings for Ramsell Corporation, ADAP’s pharmacy benefits manager (PBM), to check when receiving data from enrollment sites and request Ramsell Corporation to incorporate and emphasize “common data mistakes to avoid” as part of their enrollment trainings.
- Phase 4: ADAP staff will identify data errors from their enrollment site visits.
- Phase 5: Future data will be screened on an annual basis to continually monitor error rates.

Design. Each client and prescription variable was screened according to a set of rules based on the definition of the field. The data values were classified as either *valid*, *maybe invalid*, *invalid*, or *missing data*. For each field, OA indicated the frequency and percentage of records that were classified as valid or one of the other groups.

Results and Conclusions. Phase 1: OA found that the average client variables were 95.14 percent *valid* and the average prescription variables were 99.90 percent *valid* indicating that the values were within an acceptable range in terms of completeness, timeliness, uniqueness, and validity. Such findings demonstrate that ADAP is collecting data at a very efficient level through its PBM, Ramsell Corporation. Phase 2: A five percent error rate was established for annual screenings of ADAP’s data. Phase 3: Two health indicators, CD4 counts and viral load and their test dates, and two co-pay fields for private insurance and Medi-Cal transactions require initial screening on Ramsell Corporation’s part.

¹ TDQM is based on Defense Information Systems Agency’s 2003 publication, “DoD Guidelines on Data Quality Management (Summary).”

BACKGROUND

Data quality management (QM) involves the collection, organization, maintenance, and usage of meaningful data. It is an essential part of the success of any program that relies on data to describe the program, monitor its activities, and to strategically plan for the program's future.

Despite the importance of data QM, an Internet search revealed limited public documentation on how to proceed with conducting such a task [Data Management Handbook (North American Research Strategy for Tropospheric Ozone (NARSTO), 2000), Data Quality Report for Client Demonstration Project (Health Resources and Services Administration (HRSA), 2002), and Data "Sanity": Statistical Thinking Applied to Everyday Thinking (Balestracci, 1998)].

DoD (2003), for example, identifies six components of TDQM to ensure data users are involved in the data improvement process, predetermined data standards of excellence are well-defined, and that data meet the following standards:

1. Accuracy – the degree to which data values are correct in comparison to the actual value (e.g., gender = male when the subject is a male).
2. Completeness – the degree to which data fields have values and no missing or unknown values.
3. Consistency – the degree to which matching values are the same across tables, files, or records.
4. Timeliness – the degree to which data values are current or up-to-date.
5. Uniqueness – the degree to which records have one primary key or no duplicates.
6. Validity – the degree to which data values conform to an acceptable classification system for all elements.

California's ADAP, established in 1987 under Title II of the Ryan White (RW) Comprehensive AIDS Resources Emergency (CARE) Act, provides Food and Drug Administration-approved drug therapies to HIV/AIDS individuals. The program is intended as a last resort for low-income individuals with limited or no coverage from private insurance or Medi-Cal to help pay for their medications. In FY 2002–03, 25,759 clients accessed 775,655 prescriptions through ADAP.

ADAP enrollment process begins at one of over 230 sites throughout the state, which are coordinated by 61 local health jurisdictions. At each site, enrollment workers screen applicants for ADAP eligibility in accordance with the following criteria. These applicants qualify for the program if they:

- are infected with HIV;
- have an annual federal adjusted gross income (FAGI) below \$50,000;²
- are not fully covered by nor eligible for Medi-Cal or other third-party payer;

² An individual is subject to a co-payment obligation if his/her annual FAGI is between 400 percent of federal poverty level (FPL) and \$50,000.

- are a resident of California;
- 18 years of age or older; and
- have a valid prescription from a California licensed physician.

Eligible clients are then entered into a database maintained by Ramsell Corporation. Ramsell Corporation, a PBM contractor with the California Department of Health Services (DHS) since July 1997 to present, oversees client enrollment and verification, maintains a pharmacy network, processes prescription claims, and coordinates reimbursement between the State and nearly 3,400 participating ADAP pharmacies.

Ramsell Corporation provides OA with weekly client and prescription files. For example, the weekly files for the invoice period from July 1, 2002 to July 7, 2002, contained 5,340 clients and 14,468 prescriptions. The client-level data file includes 43 variables with demographic (e.g., client's gender, date of birth, and race/ethnicity) and eligibility information (e.g., client's enrollment date, HIV/AIDS diagnosis, and income). The prescription-level data file includes another 22 variables with billing information such as the dispensing date of the drug, National Drug Code (NDC) of the medication, number of units dispensed by the pharmacy, days supply of the drug, and the drug's net cost without a dispensing fee. Table 1 (Client-Level Data) and Table 2 (Prescription-Level Data) show the entire list of fields or variables collected by ADAP, definitions of each variable, and a brief coding scheme for each variable.

THE PRESENT STUDY

ADAP is funded, in part, and administered at the federal level by HRSA. In FY 2001–02, HRSA issued a mandate requiring all programs that receive CARE Act funds “to develop, implement, and monitor QM programs.” In response to HRSA, the purpose of this data QM study was to examine and evaluate ADAP's data files and to develop a mechanism to continually improve the data collection process. Five phases were planned with the first three occurring within this study:

- **Phase 1:** Screen the entire array of individual client and prescription variables in FY 2002–03 according to applicable DoD TDQM standards (completeness, timeliness, uniqueness, and validity). Accuracy and consistency were not addressed, since OA cannot verify the data in-house and no variable appears in both files other than a unique identifier, which was used to match the client and prescription files.
- **Phase 2:** Determine an acceptable error rate for the data and identify the variables that fall below the criteria.
- **Phase 3:** Review and (if necessary) develop additional QA listings for Ramsell Corporation to check when receiving data from enrollment sites and request Ramsell Corporation to incorporate and emphasize “common data mistakes to avoid” as part of their enrollment worker trainings. As part of the original PBM contract, Ramsell Corporation currently has weekly QA listings for both client and prescription files, and these forms were reviewed.

- **Phase 4:** ADAP staff will identify data errors from their enrollment site visits.
- **Phase 5:** Future data will be screened on an annual basis to continually monitor error rates. This information will be reported back to Ramsell Corporation, ADAP staff, and ADAP enrollment workers for corrective action.

METHOD AND RESULTS

Phase 1

Client–Level Data. Each client and prescription variable was screened according to a set of rules based on the definition of the field. The data values were classified as either *valid*, *maybe invalid*, *invalid*, or *missing data*. For example, a client’s Federal ID string (idstring) is an 11-character field based on the first and third letter of the client’s first and last name followed by the client’s date of birth (mmddyy format) and gender (1 = male and 2 = female). The idstring for John Smith, who was born on April 21, 1973, would be JHSI0421731. Classification would be as follows:

- *Valid* idstring – Had letters in the first four fields as described above.
- *Maybe invalid* idstring – Had a number in the second or fourth position (third letter of first name or last name), which can be valid for individuals with two letters in their first or last name.
- *Invalid* idstring – Had a number in the first and second position (first and third letter of first name).

Date of birth and gender were not examined here and instead evaluated within their own field (i.e., no consistency check). Although up to four historic CD4 counts and viral load measures were recorded, only the most current values and test dates were examined. Also, undocumented clients, grace (period) status, and eligibility end dates were not examined because these fields were not collected at the beginning of the year. This reduced the total number of client variables from 43 to 28.

Table 3 shows the QM results for the Client–Level Data. For each field, OA indicated the frequency and percentage of records that were classified as *valid* or one of the other groups. For the 25,759 client records in FY 2002-03, an average of 95.14 percent were *valid* across the variables, 2.65 percent were *maybe invalid*, 0.79 percent were *invalid*, and 1.42 percent were *missing data*. Ten clients appeared to have *missing data* across all records.

Of the 28 client variables, only three fields had *valid* percentages less than 85 percent—CD4date (CD4 test date), vload (viral load), and vldate (viral load test date). The two issues with CD4date was that 9.96 percent of clients had *invalid* test dates prior to January 2001 or after June 2003 (less than .01 percent) and 7.10 percent had *missing data*. For vload, 40.09 percent of clients had *maybe invalid* measures of 0 (35.15 percent), between 1 – 49 (4.69 percent), or between 1 million – 9,999,998 (less than .01

percent). Because ADAP does not collect data on the viral load test type, OA could not confirm the sensitivity of extremely low values. Vldate had similar issues to CD4date with 6.89 percent with outdated or future test dates and 25.85 percent with *missing data*.

CD4 count also had a large percent of *maybe invalid* responses. That is, 10.98 percent of clients had CD4 counts of 0—which raised questions regarding the validity of the data. These data points reduced the percentage of *valid* CD4 counts to 87.87 percent.

For the remaining 25 client variables, possible data fields of concern include depends (number of dependents) in which 13.79 percent of clients indicated no dependents including themselves or 79 clients (less than 0.01 percent) those claiming 6-16 dependents. OA classified these records as *maybe invalid*. Also, HIVDXR (year when tested positive for HIV) had 2.54 percent with a test date between 1977–1984 before commercial HIV tests were available and less than 1 percent with only three digits in the test year (e.g., 198). All of these records were classified as *invalid* and probably attributable to data entry errors.

Prescription–Level Data. Using the same methodology as above for the client variables, OA screened the prescription variables. Three variables were not examined—idstring, backoutnum (backout numbers, which refer to a unique number for credits or refunds and do not appear in the prescription files once they are removed), and adapsoc (ADAP share-of-cost which was not collected at the beginning of the fiscal year)—thereby reducing the total number of prescription variables from 22 to 19. Idstring was evaluated with the client variables but not with the prescription variables.

Also, four variables were examined at the aggregate level rather than as individual records. Because of the sheer volume of prescription claims (775,655), evaluating these variables at the individual level would probably result in less than one percent *invalid*. For example, NABP is the National Association of Boards Pharmacy number for each ADAP participating pharmacy. One pharmacy was not identifiable and this particular pharmacy dispensed 895 prescriptions in FY 2002–03. Thus, 1 out of 1,766 NABPs were *invalid* (0.06 percent) at the aggregate level. In contrast, there would have to be 7,757 invalid records at the individual level for the variable to have a one percent error rate. In this case, 895 out of 775,655 were *invalid* (0 percent) and for the same error in pharmacy number. By aggregating the variable, OA can use a more sensitive test to quantify the error more representatively. The other aggregate variables were ndc (NDC for each individual drug), jurisite (local health jurisdiction code), and rxnum (pharmacy's assigned prescription number).

The results of the QM check for the prescription variables are shown in Table 4. For the 19 variables of interest, the average number of *valid* records was a near perfect 99.90 percent. The number of *maybe invalid*, *invalid*, and *missing records* were all less than one percent. Three data fields of concern were days (days supply of a prescription drug), mdccopy (costs for Medi-Cal transactions without a dispensing fee), and othcopy (costs for private insurance co-payments without a dispensing fee). For days, 1 record had a 0 day's supply, 127 had 91–120 day's supply, and 47 had 121–330 day's

supply. Most prescriptions were 30–day supplies with 90–day supplies as the limit. Mdcopay and othcopay had a small number of records (10 for mdcopay and 29 for othcopay) with the amount over \$2,000 that were considered invalid. These were believed to be data entry errors and require further investigation.

Phase 2

The average results of 95.14 percent *valid* records for client variables and 99.90 percent *valid* records for prescription variables exceeded all expectations. While ADAP's contract with Ramsell Corporation does not specify an error rate for its data collection process, a reasonable rate could be set at five percent since most researchers consider research findings to be statistically significant at the five percent level. By this criterion, only the client variables aidstat (HIV/AIDS diagnosis), CD4date, and vldate did not meet this data quality standard. Also, vload had a large percentage of *maybe invalid* responses, which would put the variable on an ADAP “watch list.” All prescription variables were well below the five percent error rate.

Phase 3

Tables 5 and 6 show the Client and Prescription Data QA Listing that Ramsell Corporation generates every week. There are 26 potential error listings for 16 of the client variables, and 12 listings for 10 prescription variables.³ Since OA found *invalid* test dates for CD4 counts and viral load and *invalid* viral load measures, OA will recommend that Ramsell Corporation do the following:

- For CD4 dates, change the error listing from “CD4DATE predates ELGSTART (eligibility start date) by 24 mos (months)” to “CD4DATE predates ELGSTART by 12 months” to ensure more recent tests;
- Include a similar listing for viral load dates;
- Include listings that specify an acceptable range for both CD4 counts (0 or greater than 1,500) and viral load measures (0 or greater than 10,000,000); and
- Include listings for mdcopay, and othcopay that exceed \$2,000.

At present, OA meets with Ramsell Corporation on a quarterly basis in the Joint Data Policy Meeting to discuss such issues.

³The original PBM contract included QA criteria for most fields available at that time. Because the PBM contract is up for renewal in FY 2005-06, those listings are currently being updated and reviewed and will be presented in a follow-up study.

DISCUSSION

This study was conducted primarily to examine and evaluate the client and prescription variables collected by ADAP. OA found that the client variables were 95.14 percent valid and the prescription variables were 99.90 percent valid indicating that the values were within an acceptable range in terms of completeness, timeliness, uniqueness, and validity. Such findings demonstrate that ADAP is collecting data at a very efficient level through its PBM, Ramsell Corporation. Ten clients had no identifiable client demographic information, and such occurrences must be checked with Ramsell Corporation.

A five percent error rate was established for annual screenings of ADAP's data. OA will attempt to continue to exceed this rate and bring up the more "difficult" variables to this standard (e.g., CD4 counts and test dates for both CD4 counts and viral load). Efforts will be made with Ramsell Corporation and local ADAP enrollment workers to emphasize the importance of obtaining accurate and recent data for such fields.

Because of the importance of indicators such as CD4 counts and viral load in assessing the health status of ADAP clients, OA has already begun examining these measures along with test dates and eligibility start dates using stricter criteria than in this study (Wong and Fairgrievies, 2004). For example, a recent test date should be no more than six months prior to the client's eligibility start date. Also, two co-pay fields for private insurance and Medi-Cal transactions require initial screening on Ramsell Corporation's part.

The next two phases of this QM effort will be to incorporate the findings from ADAP staff site visits with this data and to coordinate these results with Ramsell Corporation, ADAP staff, and ADAP enrollment workers to ensure the highest degree of accuracy for the data collected. Further research will also begin examining more comprehensive bivariate and multivariate relationships between two and three variables. For example, a client with an income over 400 percent FPL should have an ADAP share-of-cost, or client enrollment dates must precede client's eligibility start date. Also, OA will transition to the stricter criteria for CD4 and viral load test dates in relationship to client eligibility start dates.

REFERENCES

Balestracci D. *Data "Sanity": Statistical Thinking Applied to Everyday Data*. Special Publication of the American Society for Quality Statistics Division, Summer 1998.

Department of Defense (DoD), *DOD Guidelines on Data Quality Management (Summary)*. Defense Information Systems Agency, 2003.

Health Resources and Services Administration, *Data Quality Report for Client Demonstration Project (CDP): CDP Site Virginia*, 2002.

North American Research Strategy for Tropospheric Ozone (NARSTO), *Data Management Handbook*, 2000.

Wong, D.T., Fairgrievies, K.S., *CD4 Counts and Viral Load Measurements in AIDS Drug Assistance Program (ADAP): Quality (Data) Management and Health Indicators*. California Department of Health Services, Office of AIDS, 2004.

The ADAP Files: Data Quality Management from A to Z

TABLE 1: CLIENT-LEVEL DATA, FY 2002-03					
#	NAME OF FIELD	DEFINITION OF FIELD	FIELD FORMAT	FIELD LENGTH	VALUES/COMMENTS
1	IDSTRING	Federal ID string	Character	11	(All fields are non-blank unless noted)
2	ZIPCODE	Zip Code where client resides	Character	5	90001 – 96162
3	ELGSTART	Current annual eligibility start date	Date	10	mm/dd/yyyy
4	NRLDATE	Date client first enrolled in ADAP	Character	10	mm/dd/yyyy
5	BDATE	Client's date of birth	Character	10	mm/dd/yyyy
6	GENDER	Client's biological gender	Numeric	1	1 = male, 2 = female, 3 = male-to-female, and 4 = female-to-male
7	RACE1	Spanish, Hispanic, Latino ethnic heritage (= 300 with sub-categories), or (000 = Non-Spanish, non-Hispanic, or non-Latino)	Character	3	i.e., 320 = Cuban, 350 = South American, 380 = Other Hispanic
8	RACE2	Racial/Ethnic heritage (= 100, 200, 400, 500, 600 sub-categories, or 999)	Character	3	Blank ok/Non-Duplicate of RACE 1 – 4
9	RACE3	Client's other ethnic heritage	Character	3	Blank ok/Non-Duplicate of RACE 1 – 4
10	RACE4	Client's other ethnic heritage	Character	3	Blank ok/Non-Duplicate of RACE 1 – 4
11	JURIS	Local health jurisdiction where client was enrolled or was last re-certified	Character	2	Codes 01 – 61
12	SITE	Enrollment site where client enrolled or was last re-certified	Character	2	Codes 01 – 99
13	INCOME	Annual adjusted gross income reported during the most recent eligibility	Numeric	6	Value <= \$50,000 and 999999 for Unknown
14	DEPENDS	Number of dependents	Numeric	2	
15	MEDICAL	Medi-Cal status	Numeric	1	1 = yes, 2 = no, 3 = Medi-Cal approval pending, and 9 = unknown
16	PVTINS	Private insurance coverage	Numeric	1	1 = yes, 2 = no, and 9 = unknown
17	PUBINS	Client's other public insurance coverage	Numeric	1	1 = yes, 2 = no, and 9 = unknown
18	SHAMT	Client's monthly determined ADAP share-of-cost	Numeric	9	
19	MEDSHCST	Client's monthly Medi-Cal share-of-cost	Numeric	9	

The ADAP Files: Data Quality Management from A to Z

TABLE 1: CLIENT-LEVEL DATA, FY 2002-03 CONTINUED

#	NAME OF FIELD	DEFINITION OF FIELD	FIELD FORMAT	FIELD LENGTH	VALUES/COMMENTS
20	HIVDXYR	Reported year when client tested seropositive for HIV	Numeric	4	yyyy = year and 9999 = unknown
21	AIDSTAT	Most recent AIDS diagnosis code	Numeric	1	1 = HIV asymptomatic, 2 = HIV symptomatic, 3 = AIDS diagnosed, and 9 = unknown
22	AIDSDXYR	Month and year associated with most recent HIV/AIDS diagnosis	Character	7	mm/yyyy
23	CD4COUNT	Most recent CD4 count	Numeric	4	9998 = CD4 test not performed and 9999 = CD4 test results unknown; if CD4count < 200 then AIDSTAT = 3
24	CD4DATE	Month and year associated with most recent CD4 count reported	Character	7	99/9999; if > (ELGSTART = 2) then the date must be verified
25	CD4CNT1	Previous CD4 count	Numeric	4	9998 = CD4 test not performed and 9999 = CD4 test results unknown
26	CD4DATE1	Month and year associated with previous CD4 count reported	Character	7	mmm/yyyy
27	CD4CNT2	Previous CD4 count1 (history)	Numeric	4	9998 = CD4 test not performed and 9999 = CD4 test results unknown
28	CD4DATE2	Month and year associated with previous CD4 count1 reported (history)	Character	7	mm/yyyy
29	CD4CNT3	Previous CD4 count2 (history)	Numeric	4	9998 = CD4 test not performed and 9999 = CD4 test results unknown
30	CD4DATE3	Month and year associated with previous CD4 count2 reported (history)	Character	7	mm/yyyy
31	VLOAD	Most recent viral load	Numeric	9	
32	VLDATE	Month and year associated with most recent viral load reported	Character	7	mm/yyyy
33	VLOAD1	Previous viral load (history)	Numeric	9	
34	VLDATE1	Month and year associated with previous viral load reported (history)	Character	7	mm/yyyy
35	VLOAD2	Previous viral load1 (history)	Numeric	9	
36	VLDATE2	Month and year associated with previous viral load1 reported (history)	Character	7	mm/yyyy
37	VLOAD3	Previous viral load2 (history)	Numeric	9	
38	VLDATE3	Month and year associated with previous viral load2 reported (history)	Character	7	mm/yyyy

The ADAP Files: Data Quality Management from A to Z

TABLE 1: CLIENT-LEVEL DATA, FY 2002-03 CONTINUED

#	NAME OF FIELD	DEFINITION OF FIELD	FIELD FORMAT	FIELD LENGTH	VALUES/COMMENTS
39	CONSENT	Indicates client has signed a consent form	Numeric	1	1 = yes or 2 = no
40	LANG	Language client prefers to receive printed materials	Numeric	1	1 = English, 2 = Spanish, 3 = Tagalog, 4 = Cantonese/Mandarin, and 5 = other
41	UNDOC	Client's undocumented status	Numeric	1	1 = yes and 2 = no
42	ELIGEND	Client's annual eligibility end date	Character	10	mm/dd/yyyy
43	GRSTATUS	Client's 30-day grace status	Numeric	1	0 = no and 1 = yes

The ADAP Files: Data Quality Management from A to Z

TABLE 2: PRESCRIPTION-LEVEL DATA, FY 2002-03					
#	NAME OF FIELD	DEFINITION OF FIELD	FIELD FORMAT	FIELD LENGTH*	VALUES/COMMENTS
1	IDSTRING	Federal ID string	Character	11	(All fields are non-blank unless noted)
2	NABP	Pharmacy NABP number	Character	7	
3	NDC	Drug NDC	Character	11	
4	GPI	GPI (Generic Price Indicator) indicates whether or not the drug is priced as a generic drug or a brand drug	Numeric	1	1 = generic priced and 2 = no generic priced or non-drug item
5	PHSFLAG	Identifies drugs as a Public Health Service (PHS) drug or a Non-PHS drug	Numeric	1	1 = PHS drug and 2 = non-PHS drug
6	NETCOST	State's financial share for drug dispensed; it does not include dispense fee	Numeric	12(4)	
7	NETUNITS	Number of drug units associated with the netcost of the drug	Numeric	8(2)	
8	UNITS	Number of units billed by pharmacy	Numeric	7(2)	
9	DAYS	The prescription days supply	Numeric	5(1)	
10	MRA	Contracted reimbursement rate associated with the transaction date	Numeric	12(4)	
11	D_FEE	Dispense fee invoiced to State	Numeric	4(2)	
12	DISPDATE	Date prescription was dispensed	Date	10	mm/dd/yyyy
13	INVDATE	Date prescription was invoiced	Date	10	mm/dd/yyyy
14	JURISITE	Local health jurisdiction code	Character	4	Codes 01 – 99
15	ADPCOPAY	Represents the member's ADAP share-of-cost payment towards the claim	Numeric	12(4)	
16	MDCOPAY	Represents the State's net cost for Medi-Cal transactions	Numeric	12(4)	
17	OTHCOPAY	Represents all transactions processed as a copayment amount; this includes insurance copayments	Numeric	12(4)	
18	RXNUM	Pharmacy's assigned prescription number	Character	10	

The ADAP Files: Data Quality Management from A to Z

TABLE 2: PRESCRIPTION-LEVEL DATA, FY 2002-03 CONTINUED					
#	NAME OF FIELD	DEFINITION OF FIELD	FIELD FORMAT	FIELD LENGTH*	VALUES/COMMENTS
19	GRPNUM	Client's assigned five digit group number; first two digits represent the enrollment county and the last three digits represent the client coverage (i.e., Medi-Cal, private insurance, etc.)	Character	5	
20	CLAIMNUM	Uniquely assigned claim number	Character	18	
21	BACKOUTNUM	Uniquely assigned backout number	Character	18	
22	ADAPSOC	ADAP share-of-cost amount	Numeric	12(4)	
* Numbers in parentheses () indicate decimal places.					

The ADAP Files: Data Quality Management from A to Z

TABLE 3: QM RESULTS FOR CLIENT-LEVEL DATA, FY 2002-03						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
IDSTRING	FREQ	25,641	117	1	0	25,759
	PCT	99.54%	0.45%	0.00%	0.00%	100.00%
Valid	25,641	Had letters in 1st and 3rd position of first and last name				
Maybe Invalid	20	Had number 9 in 2nd position (3rd letter of first name)				
	1	Had number 9 in 2nd and 4th position (3rd letter of first and last name)				
	96	Had number 9 in 4th position (3rd letter of last name)				
Invalid	1	Had number 9 in 1st and 2nd position (1st and 3rd letter of first name)				
Note		Only checked 1st 4 characters; others to be checked with bdate and gender				
ZIPCODE	FREQ	25,666	0	76	17	25,759
	PCT	99.64%	0.00%	0.30%	0.07%	100.00%
Valid	25,666	Had zipcode between 90001 – 96162 with identifiable geographical area				
Invalid	76	Had zipcode between 90001 – 96162 without identifiable geographical area				
ELGSTART	FREQ	25,433	130	186	10	25,759
	PCT	98.73%	0.50%	0.72%	0.04%	100.00%
Valid	21,274	Had eligibility start date between July 1, 2002 – June 30, 2003				
	4,159	Had eligibility start date between July 1, 2001 – June 30, 2002				
Maybe Invalid	73	Had eligibility start date between June 1, 2001 – June 30, 2001 (1 grace period)				
	57	Had eligibility start date between May 1, 2001 – May 31, 2001 (2 grace periods)				
Invalid	181	Had eligibility start date before May 1, 2001				
	5	Had eligibility start date after June 30, 2003				
NRLDATE	FREQ	25,745	0	3	11	25,759
	PCT	99.95%	0.00%	0.01%	0.04%	100.00%
Valid	25,745	Had enrollment date between July 1, 1987 – June 30, 2003				
Maybe Invalid	2	Had enrollment date before October 1, 1987				
	1	Had enrollment date after June 30, 2003				
BDATE	FREQ	25,678	71	0	10	25,759
	PCT	99.69%	0.28%	0.00%	0.04%	100.00%
Valid	25,678	Had birth date between July 1, 1928 and June 30, 1985 (18 – 74 years old)				
Maybe Invalid	69	Had birth date before July 1, 1928 (75 years old or above)				
	1	Had birth date May 24, 1992 (11 years old)				
	1	Had birth date September 10, 2002 (less than 1 year old)				
GENDER	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Coded as male, female, or transgender				
RACE1	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Coded as Spanish/Hispanic/Latino or not				
RACE2	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	24,893	Coded for racial/ethnic heritage				
	856	Coded as unknown				

The ADAP Files: Data Quality Management from A to Z

TABLE 3: QM RESULTS FOR CLIENT-LEVEL DATA, FY 2002-03 CONTINUED						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
RACE3	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	305	Coded for other racial/ethnic heritage				
	2	Coded as unknown				
	25,442	Did not have a response				
RACE4	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	21	Coded for other racial/ethnic heritage				
	25,728	Did not have a response				
JURIS	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,739	Had local health jurisdiction between 01 – 61				
	1	Coded as 62 for County Medical Service Program clients				
	9	Coded as 99 for Ramsell Corporation (client confidentiality)				
SITE	FREQ	25,746	0	3	10	25,759
	PCT	99.95%	0.00%	0.00%	0.05%	100.00%
Valid	25,737	Coded for site name				
	9	Coded as 9901 for Ramsell Corporation (client confidentiality)				
Invalid	3	Did not have site name from FY 2001-02 Q3 to FY 2002-03 files; 895 scripts				
Note		Site name obtained from matching with Ramsell Corporation quarterly files				
INCOME	FREQ	25,674	33	0	52	25,759
	PCT	99.67%	0.13%	0.00%	0.20%	100.00%
Valid	25,674	Had income of \$0 or between \$101 – \$50,000				
Maybe Invalid	33	Had income between \$1 – \$100				
DEPENDS	FREQ	22,117	3,632	0	10	25,759
	PCT	85.86%	14.10%	0.00%	0.04%	100.00%
Valid	22,117	Had 1 – 5 dependents, including client				
Maybe Invalid	3,553	Had 0 dependents, including client				
	79	Had 6 – 16 dependents, including client				
MEDICAL	FREQ	25,731	0	8	10	25,749
	PCT	99.93%	0.00%	0.03%	0.04%	100.00%
Valid	25,731	Coded as yes, no, or maybe				
Invalid	8	Coded as 0				
PVTINS	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Coded as yes, no, or maybe				
PUBINS	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Coded as yes, no, or maybe				
SHAMT	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Had ADAP share-of-cost between \$0 – \$483				

The ADAP Files: Data Quality Management from A to Z

TABLE 3: QM RESULTS FOR CLIENT-LEVEL DATA, FY 2002-03						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
MEDSHCST	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Had Medi-Cal share-of-cost between \$0 – \$4,201				
HIVDXYR	FREQ	24,835	0	914	10	25,759
	PCT	96.41%	0.00%	3.55%	0.04%	100.00%
Valid	23,581	Had year of HIV positive test date between 1985 – 2003				
	1,254	Coded as unknown				
Invalid	655	Had year of HIV positive test date between 1977 – 1984				
	253	Had 1 number in year of HIV positive test date (0 or 4)				
	6	Had 3 numbers in year of HIV positive test date (198, 199, or 200)				
AIDSTAT	FREQ	24,355	1,392	2	10	25,759
	PCT	94.55%	5.40%	0.01%	0.04%	100.00%
Valid	24,355	Coded as HIV asymptomatic, HIV symptomatic, or AIDS diagnosed				
Maybe Invalid	1,392	Coded as unknown				
Invalid	2	Coded as 0				
AIDSDXYR	FREQ	24,363	0	70	1,326	25,759
	PCT	94.58%	0.00%	0.27%	5.15%	100.00%
Valid	24,363	Had HIV diagnosis date between March 1985 – June 2003				
Invalid	17	Had HIV diagnosis date before March 1985 (FDA approval of ELISA test)				
	53	Had HIV diagnosis date after June 2003				
CD4COUNT	FREQ	22,635	2,829	137	158	25,759
	PCT	87.87%	10.98%	0.53%	0.61%	100.00%
Valid	22,635	Had CD4 count between 1 – 1,500				
Maybe invalid	2,829	Had CD4 count of 0				
Invalid	137	Had CD4 count above 1,500				
CD4DATE	FREQ	21,435	0	2,496	1,828	25,759
	PCT	83.21%	0.00%	9.69%	7.10%	100.00%
Valid	13,321	Had CD4 test date between July 2002 – June 2003				
	8,114	Had CD4 test date between January 2001 – June 2002				
Invalid	2,421	Had CD4 test date between June 1988 – December 2000				
	75	Had CD4 test date after June 2003				
VLOAD	FREQ	15,415	10,327	7	10	25,759
	PCT	59.84%	40.09%	0.03%	0.04%	100.00%
Valid	15,415	Had viral load between 50 - 999,998				
Maybe Invalid	9,055	Had viral load of 0				
	1,208	Had viral load between 1 – 49				
	64	Had viral load between 1,000,000 – 9,999,998				
Invalid	3	Had viral load with negative value				
	3	Had viral load above 10,000,000				
	1	Had viral load of 99,999,999				

The ADAP Files: Data Quality Management from A to Z

TABLE 3: QM RESULTS FOR CLIENT-LEVEL DATA, FY 2002-03 CONTINUED						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
VLDATE	FREQ	17,324	0	1,776	6,659	25,759
	PCT	67.25%	0.00%	6.89%	25.85%	100.00%
Valid	6,356	Had viral load test date between July 2002 – June 2003				
	10,968	Had viral load test date between January 2001 – June 2002				
Invalid	1,703	Had viral load test date between June 1968 – December 2000				
	73	Had viral load test date after June 2003				
CONSENT	FREQ	25,171	578	0	10	25,759
	PCT	97.72%	2.24%	0.00%	0.04%	100.00%
Valid	25,171	Coded as yes				
Maybe Invalid	578	Coded as no				
LANG	FREQ	25,749	0	0	10	25,759
	PCT	99.96%	0.00%	0.00%	0.04%	100.00%
Valid	25,749	Coded as yes				
AVERAGE		95.14%	2.65%	0.79%	1.42%	100.00%

The ADAP Files: Data Quality Management from A to Z

TABLE 4: QM RESULTS FOR PRESCRIPTION-LEVEL DATA, FY 2002-03						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
IDSTRING	FREQ	N/A	N/A	N/A	N/A	N/A
	PCT	N/A	N/A	N/A	N/A	N/A
Note		Not applicable because idstring in client file is aggregate version				
NAPB (Agg)	FREQ	1,766	0	1	0	1,767
	PCT	99.94%	0.00%	0.06%	0.00%	100.00%
Valid	1,766	Had pharmacy name				
Invalid	1	Did not have pharmacy name from FY 2001-02 Q3 to FY 2002-03 files; 895 scripts				
Note		Pharmacy name obtained from matching with Ramsell Corporation quarterly files				
NDC (Agg)	FREQ	1,929	0	0	0	1,929
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	1,929	Had 11-digit National Drug Code code.				
Note		Brand and label description obtained from matching with ADAP drug files				
GPI	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,655	Coded as generic or brand drug				
PHSFLAG	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,655	Coded as PHS or non-PHS drug				
NETCOST	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,655	Had cost between \$0.03 – \$7,455.42				
NETUNITS	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,654	Had net units between 0.01 – 6,938.75				
	1	Had net units of 11250.00				
UNITS	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,654	Had net units between 0.50 – 4,803.00				
DAYS	FREQ	775,480	0	175	0	775,655
	PCT	99.98%	0.00%	0.02%	0.00%	100.00%
Valid	104,039	Had 1 – 29 day supply				
	660,854	Had 30 day supply				
	7,942	Had 31 – 60 day supply				
	2,645	Had 61 – 90 day supply				
Invalid	1	Had 0 day supply				
	127	Had 91 – 120 day supply				
	47	Had 121 – 330 day supply				
MRA	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,654	Had Medical Reimbursement Amount between \$0.0080 – \$864.3242				
	1	Had Medical Reimbursement Amount of \$3,550.00				

The ADAP Files: Data Quality Management from A to Z

TABLE 4: QM RESULTS FOR PRESCRIPTION-LEVEL DATA, FY 2002-03 CONTINUED						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
D_FEE	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,655	Had dispense fee of \$4.05				
DISPDATE	FREQ	768,438	7,199	18	0	775,655
	PCT	99.07%	0.93%	0.00%	0.00%	100.00%
Valid	768,438	Had dispense date between July 1, 2002 – June 30, 2003				
Maybe Invalid	74	Had dispense date between August 16, 2000 – June 30, 2001				
	7,125	Had dispense date between July 1, 2001 – June 30, 2002				
Invalid	18	Had dispense date of July 1, 2003				
INVDATE	FREQ	775,655	0	0	0	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,655	Had dispense date between July 9, 2002 – July 1, 2003				
JURISITE (Agg)	FREQ	213	0	1	1	215
	PCT	99.07%	0.00%	0.47%	0.47%	100.00%
Valid	213	Coded for site name				
	1	Coded as 9901 for Ramsell Corporation (client confidentiality); 139 scripts				
Invalid	1	Did not have site name from FY 2001-02 Q3 to FY 2002-03 files; 50 scripts				
Note		Site name obtained from matching with Ramsell Corporation quarterly files				
ADPCOPAY	FREQ	775,646	0	9	0	775,665
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,009	Had ADAP copay of \$0.00				
	637	Had ADAP copay between \$4.61 – \$1,000				
Invalid	4	Had ADAP copay between -\$207.00 – -\$323.00				
	5	Had ADAP copay between over \$1,400.00				
MDCOPAY	FREQ	775,611	34	10	0	775,655
	PCT	99.99%	0.00%	0.00%	0.00%	100.00%
Valid	741,052	Had ADAP copay of \$0.00				
	34,559	Had ADAP copay between \$0.14 – \$1,000				
Maybe Invalid	34	Had ADAP copay between \$1,001 – \$2,000				
Invalid	10	Had ADAP copay between \$2,241.00 – \$7,159.29				
OTHCOPIAY	FREQ	775,530	96	26	0	775,652
	PCT	99.98%	0.01%	0.00%	0.00%	100.00%
Valid	612,262	Had ADAP copay of \$0.00				
	163,268	Had ADAP copay between \$0.27 – \$1,000				
Maybe Invalid	96	Had ADAP copay between \$1,001 – \$2,000				
Invalid	29	Had ADAP copay between \$2,068.78 – \$4,588.99				
RXNUM (Agg)	FREQ	299,583	0	0	0	299,583
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid		Had ten-digit pharmacy prescription number				
GRPNUM	FREQ	775,650	0	0	5	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	475,557	Had ADAP group number 010 – ADAP with no other payer				
	300,093	Had other valid ADAP group number				
Missing	5	Had missing group code				

The ADAP Files: Data Quality Management from A to Z

TABLE 4: QM RESULTS FOR CLIENT-LEVEL DATA, FY 2002-03 CONTINUED						
NAME OF FIELD	STAT	VALID	MAYBE INVALID	INVALID	MISSING	TOTAL
CLAIMNUM	FREQ	775,654	0	0	1	775,655
	PCT	100.00%	0.00%	0.00%	0.00%	100.00%
Valid	775,654	Had valid (unique) claim number				
Invalid	1	Had duplicate claim number				
AVERAGE		99.90%	0.05%	0.03%	0.02%	100.00%

The ADAP Files: Data Quality Management from A to Z

TABLE 5: RAMSELLCORPORATION'S WEEKLY CLIENT-LEVEL DATA QA LISTING		
#	NAME OF FIELD	ERROR DESCRIPTION
1	IDSTRING	Invalid name characters 1 – 4
2	ZIPCODE	Field must not be blank
3	ELGSTART	Date is in future
4	NRLDATE	NRLDATE is later than ELGSTART
4	NRLDATE	Predates 10/1/87
5	BDATE	Age is over 75
5	BDATE	Age is under 18
6	BDATE	Does not match birthdate in IDSTRING
6	GENDER	Does not match gender in IDSTRING
6	GENDER	Invalid code
8	RACE2	Invalid RACE2 code
9	RACE3	RACE3 code duplicates RACE2
11	JURIS	Invalid code
13	INCOME	Missing or unknown
15	MEDICAL	Coded 1 but MEDSHCST = 0
15	MEDICAL	Invalid code number
15	MEDICAL	Not coded 1 but MEDSHCST >0
18	SHAMT	Greater than zero but MEDICAL = 1
20	HIVDXYR	Later than NRLDATE
20	HIVDXYR	Missing or unknown
20	HIVDXYR	Predates Jan 1982
21	AIDSTAT	Invalid code number
21	AIDSTAT	Not coded 3 but CD4COUNT less than 200
24	CD4DATE	CD4DATE predates ELGSTART by > 24 mos
24	CD4DATE	Valid CD4DATE but missing/unknown CD4COUNT
39	CONSENT	Coded 2 but 30 days past ELGSTART

TABLE 6: RAMSELL CORPORATION'S WEEKLY PRESCRIPTION-LEVEL DATA QA LISTING		
#	NAME OF FIELD	ERROR DESCRIPTION
1	IDSTRING	Invalid name characters 1 – 4
6	NETCOST	Invalid negative number
6	NETCOST	Net cost exceeds MRA x NETUNITS by >5%
7	NETUNITS	Net units less than .5
8	UNITS	Units less than .5
8	UNITS	Units less than NETUNITS
9	DAYS	DAYS greater than 90
12	DISPDATE	Date is more than 6 months in past
13	INVDATE	Date is in future
14	JURISITE	Field must not be blank
15	ADPCOPAY	Invalid negative number
18	RXNUM	Invalid non-numeric characters