Published in final edited form as:
Med Care. 2009 January ; 47(1): 121–124. doi:10.1097/MLR.0b013e31817d69c2.

Comparison of Two VA Laboratory Data Repositories Indicates That Missing Data Vary Despite Originating From the Same Source

Kathleen A. McGinnis, MS[*], **Melissa Skanderson, MSW**[*], **Forrest L. Levin, MS**[†], **Cynthia Brandt, MD**[†,‡], **Joseph Erdos, MD**[†,‡], and **Amy C. Justice, MD, PhD**[†,‡]

[*]Center for Health Equity Research and Promotion, VA Pittsburgh Healthcare System, Pittsburgh, Pennsylvania

[†]VA Connecticut Healthcare System, West Haven, Connecticut

[‡]Yale University School of Medicine, New Haven, Connecticut

Abstract

Background—Assessing accuracy and completeness of data is an important component of conducting research. VA Healthcare System benefits from a highly developed electronic medical information system. The Immunology Case Registry was designed to monitor costs and quality of HIV care. The Decision Support System was developed to monitor utilization and costs of veterans in care. Because these systems extract data from the same source using independent methods, they provide an opportunity to compare the accuracy and completeness of each.

Objective—To compare overlapping laboratory data from the Veterans Affairs Health Information System between 2 data repositories.

Research Design—For hemoglobin, $CD4^+$ lymphocyte counts (CD4), HIV RNA viral load, aspartate aminotransferase, alanine aminotransferase, glycosylated hemoglobin, creatinine, and white blood count, we calculated the percent of individuals with a value from each source. For results in both repositories, we calculated Pearson's correlation coefficients.

Subjects—A total of 22,647 HIV + veterans in the Virtual Cohort with a visit in fiscal year 2002.

Results—For 6 out of 9 tests, 68% to 72% of the observations overlapped. For CD4, viral load, and glycosylated hemoglobin less than 31% of observations overlapped. Overlapping results were nearly perfectly correlated except for CD4.

Conclusions—Six of the laboratory tests demonstrated remarkably similar amounts of overlap, though Immunology Case Registry and Decision Support System both have missing data. Findings indicate that validation of laboratory data should be conducted before its use in quality and efficiency projects. When 2 databases are not available for comparison, other methods of validation should be implemented.

Reprints: Amy C. Justice, MD, PhD, VA Connecticut Healthcare System, 950 Campbell Ave, 11-ACSLG, West Haven, CT. amy.justice2@va.go.

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans.

Keywords

laboratory; DSS; ICR; VA

Clinical quality and efficiency studies may be biased by inaccurate, incomplete, or inappropriately mapped data derived from electronic medical record systems. Quality of care studies may underreport patients receiving laboratory tests for prevention if data are incomplete. Studies reporting data on only patients with complete laboratory data available may be biased if data are not missing at random, underpowered if useful results are omitted, or reach false conclusions if the mapping of variables is inaccurate. Therefore, assessing the accuracy and completeness of datasets derived from electronic medical records systems is an important step.

The Veterans Affairs Healthcare System (VAHS) benefits from one of the most highly developed health information systems in the world.[1,2] Many quality and efficiency studies using Veterans Affairs Health Information System (VA HIS) data include the use of derived laboratory data available in clinical databases.[3–15] Data from previously extracted databases are used rather than that directly obtained from the VA HIS for several reasons. First, data extraction requires substantial computational time and can affect the performance of the VA HIS for patient care. Second, although the VA has a national HIS, laboratory names are locally determined, thus names must be correctly identified and mapped for each medical center (station). Ongoing monitoring of local test names is necessary as new tests are made available within the system and program mapping must be updated to reflect these changes.[16] Additionally, once data are extracted, they must be cleaned appropriately. Incorrect cleaning can result in loss of useful data or incorporation of incorrect data. Further, the correct laboratory date must be identified.

We have identified 2 national databases which extract overlapping laboratory data from the VA HIS using independent methods. Immunology Case Registry (ICR) was an automated electronic database designed to monitor the costs and quality of care to all veterans in care with human immunodeficiency virus (HIV).[16] Once in the registry, extraction of patients' medical record and laboratory data occurred automatically. ICR created a system that mapped laboratory names on a national level to local laboratory names.

Decision Support System (DSS) is a national database created from the standard Veterans Health Administration clinical and financial data sources by the VA Decision Support Office. DSS collects laboratory data from October 1, 1999 forward for all veterans in care for a subset of laboratory tests.[17] Datasets are created and made available for research.[18,19] Programming is maintained and run at the local level. By comparing data extracted from the VA HIS using 2 independent methods, we will evaluate the accuracy and completeness of these derivative databases.

Methods

We first identified a group of veterans in which to compare laboratory tests: HIV-positive veterans in the Veterans Aging Cohort Study Virtual Cohort with an in- or outpatient visit in fiscal year (FY) 2002. The Virtual Cohort of HIV-positive and HIV-negative veterans was created using administrative VAHS data to examine the independent effects of HIV, treatment, and comorbidities on various outcomes. We used International Classification of Diseases 9th Revision codes to identify veterans with an HIV diagnosis from October 1997 to September 2003.[14]

We chose 9 overlapping tests to compare between the 2 databases: hemoglobin, $CD4^+$ lymphocyte counts (CD4), HIV RNA viral load (VL), aspartate aminotransferase (AST), alanine aminotransferase, glycosylated hemoglobin (HB_{A1c}), creatinine, and white blood count.

Data Cleaning

To clean the data, we first removed noninformative text fields such as “pending,” “comment,” and “canc” from both sources. We removed text from the end of otherwise numeric fields, such as “*,” “>,” “<.” Additional cleaning was test specific and involved removing out of range values and taking the lowest or highest values if there were multiple values on the same day as shown in Table 1.

Analyses

For each test we combined the ICR and DSS data to determine the total number of individuals, stations, and observations available in the sources combined. Out of the individuals, stations, and observations in the sources combined, we calculated the percent of individuals, stations, and observations with a value from each source.

We merged ICR and DSS data on study id, date, and station to determine the percent of overlapping observations and to compare laboratory values occurring for the same person, on the same date, and at the same station. We calculated Pearson’s correlation coefficients to quantify the correlation between overlapping observations.

Multiple dates can be used for a laboratory date (date test was ordered, date patient underwent test, and date test result became available). We were concerned that ICR and DSS may have extracted different laboratory dates. For observations that did not link in the merge described above, we merged observations on study id, station, and laboratory value to evaluate the number of observations and difference in laboratory dates for observations merging in this way. All analyses were run using Stata version 9.2.

Results

The Virtual Cohort contains 22,647 HIV-positive veterans with at least 1 in or outpatient visit in FY 2002 at 127 different stations. Of the 22,647, 91% (20,641) had at least 1 of the 9 tests from either source. ICR contained 19,910 of the 22,647 in FY 2002 and their laboratory dataset contained 618,197 laboratory values on 17,545 individuals from 125 stations. DSS contained 556,282 laboratory values on 20,559 individuals from 127 stations. There were 17,463 overlapping individuals and 125 overlapping stations between the 2 datasets.

Table 2 summarizes the number of different local names that map to each of the 9 tests we evaluate in FY 2002 ICR. Each distinct test maps to over 50 and up to 167 different local names.

For hemoglobin, AST, HB_{A1c}, creatinine, white blood count, and glucose, DSS provided values for a greater percent of individuals than ICR. For alanine aminotransferase, the percents were slightly greater for ICR. For VL and CD4, ICR provided values for a greater percent of individuals than DSS. For each test, ICR provided data for more stations than DSS. For VL and CD4, ICR provided data on substantially more individuals and stations than DSS (Table 3).

For 7 of 9 tests, ICR provided more observations than DSS. For AST and HB_{A1c}, DSS provided more observations than ICR. For 6 of the 9 tests, from 68% to 72% of the observations overlapped (ie, occurred for the same person on the same date and at the same station). For HB_{A1c}, VL, and CD4, less than 31% of observations overlapped. For overlapping observations, correlation coefficients were high for all values (≥0.94), except CD4 (0.875) (Table 4). Of observations that did not overlap in the initial merge, few additional values merged on study id, station, and laboratory value.

Discussion

For 6 of the 9 laboratory tests, DSS and ICR are both good sources for data. However, there are substantially more VL and CD4 values in ICR than in DSS, based on number of observations, individuals, and stations represented in the dataset.

Of the 9 tests evaluated, 6 demonstrated similar amounts of overlap (between 68% and 72%) among the 2 datasets. In contrast, CD4, VL, and HB_{A1c} demonstrated much lower proportions of overlap (between 20% and 31%). We suspect that there are 2 major reasons for this lack of overlap. First, naming conventions may vary substantially by station.[20] This seems to be particularly true for CD4 and VL. Because ICR was focused on assessing HIV care, ICR spent a considerable amount of time focusing on mapping issues specifically for these measures. They have noted in the past “technical problems of inconsistent laboratory test names” for VL.16 The Center for Quality Management in Public Health, the national program office responsible for the ICR, has noted that some stations did not include CD4 or VL results in their local electronic medical records. This would explain why VL and CD4 are only provided on 122 stations in ICR whereas the other tests, except HB_{A1C}, are provided for 125 stations. We have noted in looking at the local station names that there are many names used for VL and CD4 that may not be obvious. Secondly, cleaning of these tests may be more difficult. For VL, many different tests have been used by station and over time. Additionally, there may be an issue with differential approaches to dating the test. In the VAHS Computerized Patient Record System, there are several dates and times recorded for laboratory specimens including the date the test was ordered, the date the test was drawn and the date the test was analyzed. For high volume tests like hemoglobin and creatinine, these 3 dates are likely to be on the same day. For lower volume tests like CD4, VL, and HB_{A1C}, these dates may be different. If the different data extraction mechanisms selected different laboratory dates, this may explain the lack of overlap. However, we did examine this date issue and did not find the date fields to explain the difference in overlap for these tests.

The correlation coefficient for CD4 is much lower than the correlation coefficients for the other tests. We believe this is likely a mapping issue. There are many tests that could be mapped to CD4 that do not represent CD4 counts, such as CD4 ratio and CD4/CD8 ratio. The group working on the ICR expended great effort to insure that the mapping for CD4 was accurate whereas the group developing and maintaining DSS had no reason to focus particularly on the accuracy of this test.

Results of quality and efficiency analyses may vary based on the laboratory data used. Sources containing more complete data may portray care based on number of tests performed as better than sources with less complete data. Additionally, less complete sources may be missing values that bias results in a certain way. For example, if data are more complete for stations treating sicker patients, then the population may seem to be sicker than it truly is. We can be reassured that in cases where observations are available in both sources, correlation is high indicating that the same values are being extracted for both datasets.

One limitation of this analysis is that our denominators are based on observations available in ICR or DSS. There are likely values that neither source is capturing, but we cannot evaluate this issue with these data. Additionally, we evaluated data collected in the past, so we cannot evaluate issues that could be related to using current or “real time” data.

The HIV Registry, part of the Clinical Case Registry (CCR), recently replaced the ICR and provides laboratory data. Laboratory datasets are created differently in CCR than in ICR according to the Center for Quality Management in Public Health. ICR created datasets by mapping tests to around 70,000 local laboratory names. For approximately the last 2 years, local facilities have been required to assign Logical Observation Identifiers Names and Codes (LOINC) to each local laboratory test name. The CCR contains the local name, the local national laboratory test code, and LOINC. Researchers using data from the CCR can use LOINC to extract the tests of interest and then use the local names and national laboratory test codes to verify they are using the correct LOINC.

Although ICR and DSS both draw data from the same electronic record system and correlate closely for our patient cohort, each contains observations that are not included in the other repository. This is likely because of the discrepancies in the mapping, downloading, and cleaning processes. Although our findings are based on VA datasets, they are important to researchers who use derivative datasets outside of the VA, as well. Often neither these processes nor their resulting data are validated. In the future, it may be beneficial for repositories to compare and collaborate on mapping and cleaning techniques. When it is not possible to use 2 different sources of data for validation, other methods of evaluation and quality assessment (eg, random audits of individual medical records, evaluating data completeness by sites of similar size) should be used. Validation of laboratory data, and other administrative electronic data, should be conducted when possible to ensure data quality, especially before using such data to determine quality or efficiency of care.

Acknowledgments

Supported by NIAAA U01-13566 and the Veterans Affairs Public Health Strategic Healthcare Group.

REFERENCES

1. Corrigan, JM.; Eden, J.; Smith, BM., editors. Leadership by Example: Coordinating Government Roles in Improving Healthcare Quality Committee on Enhancing Federal Healthcare Quality Programs. Washington, DC: National Academy Press; 2002.
2. McQueen L, Mittman BS, Demakis JG. Overview of the Veterans Health Administration (VHA) Quality Enhancement Research Initiative (QUERI). J Am Med Inform Assoc 2004;11:339–343. [PubMed: 15187071]
3. Backus LI, Phillips BR, Boothroyd DB, et al. Effects of hepatitis C virus coinfection on survival in veterans with HIV treated with highly active antiretroviral therapy. J Acquir Immune Defic Syndr 2005;39:613–619. [PubMed: 16044016]
4. Backus LI, Boothroyd DB, Phillips BR, et al. Pretreatment assessment and predictors of hepatitis C virus treatment in US veterans coinfected with HIV and hepatitis C virus. J Viral Hepatitis 2006;13:799–810.
5. McGinnis KA, Fine MJ, Skanderson M, et al. Understanding racial disparities using data from the Veterans Aging Cohort 3-Site Study and VA administrative data. Am J Public Health 2003;93:1728–1733. [PubMed: 14534229]
6. Fultz SL, McGinnis KA, Skanderson M, et al. Association of venous thromboembolism with human immunodeficiency virus and mortality in veterans. Am J Med 2004;116:420–423. [PubMed: 15006592]
7. McGinnis KA, Fultz SL, Skanderson M, et al. Hepatocellular carcinoma and non-Hodgkin’s lymphoma: the roles of HIV, hepatitis C infection, and alcohol abuse. J Clin Oncol 2006;24:5005–5009. [PubMed: 17075119]

8. Goulet JL, Fultz SL, McGinnis KA, et al. Relative prevalence of comorbidities and treatment contraindications in HIV-mono-infected and HIV/HCV-co-infected veterans. AIDS 2005;19 Suppl: 99–105.

9. Polgreen PM, Fultz SL, Justice AC, et al. Association of hypocholes-terolaemia with hepatitis C virus infection in HIV-infected people. HIV Med 2004;5:144–150. [PubMed: 15139979]

10. Fultz SL, Justice AC, Butt AA, et al. Project Team. Testing, referral, and treatment patterns for hepatitis C virus coinfection in a cohort of veterans with human immunodeficiency virus infection. Clin Infect Dis 2003;36:1039–1046. [PubMed: 12684917]

11. Kilbourne AM, Justice AC, Rollman BL, et al. Clinical importance of HIV and depressive symptoms among veterans with HIV infection. J Gen Intern Med 2002;17:512–520. [PubMed: 12133141]

12. McGinnis KA, Justice AC. Factors associated with dementia and cognitive impairments in veterans with human immunodeficiency virus. Arch Neurol 2002;59:490. [PubMed: 11890862]

13. Justice AC, Dombrowski E, Conigliaro J, et al. Veterans Aging Cohort Study (VACS): overview and description. Med Care 2006;44 Suppl:13–24.

14. Fultz SL, Skanderson M, Mole LA, et al. Development and verification of a “virtual” cohort using the National VA Health Information System. Med Care 2006;44 Suppl:25–30.

15. Gordon AJ, McGinnis KA, Conigliaro J, et al. Project Team. Associations between alcohol use and homelessness with healthcare utilization among human immunodeficiency virus-infected veterans. Med Care 2006;44 Suppl:37–43.

16. Backus L, Mole L, Chang S, et al. The Immunology Case Registry. J Clin Epidemiol 2001;54 Suppl:12–15.

17. Edward J, Hines, Jr. VIReC Research User Guide: VHA DSS Clinical National Data Extracts FY2000–FY2004. VA Hospital, Hines, IL: Veterans Affairs Information Resource Center; 2004 Aug. (Rev. August 2005).

18. Arnold, N.; Hynes, DM.; Stroupe, KT. Edward, Hines, Jr. VIReC Technical Report 1: Comparison of VA Outpatient Prescriptions in the DSS Datasets and the PBM Datasets. VA Hospital, Hines, IL: Veterans Affairs Information Resource Center; 2006 Jan 15.

19. Edward, Hines, Jr, editor. VIReC Research User Guide: Select Variable Frequencies From FY2002–FY2003 VHA Decision Support System Laboratory (LAB) Datasets. VA Hospital, Hines, IL: Veterans Affairs Information Resource Center; 2005.

20. Kannry JL, Wright L, Shifman M, et al. Portability issues for a structured clinical vocabulary: mapping from Yale to the Columbia Medical Entities Dictionary. J Am Med Inform Assoc 1996;3:66–78. [PubMed: 8750391]

TABLE 1

Description of Laboratory-Specific Data Cleaning

	Out of Range Values	If Multiple Values on Same Day, Used:
Hemoglobin	<4, >20	Lowest
Aspartate aminotransferase	>2000	Highest
Alanine aminotransferase	>1000	Highest
Glycosylated hemoglobin	None	Highest
Creatinine	>14	Highest
White blood count	>100	Lowest
Glucose	None	Highest
HIV RNA viral load	Changed fields containing “NO” or “NEG” to zero	Value without “<” or “>” sign highest
CD4$^+$ lymphocyte count	>20; any value with decimal point	Lowest

TABLE 2

Local Laboratory Names in ICR Data (125 Stations)

	ICR Mapped Local Laboratory Names
Hemoglobin	116
Aspartate aminotransferase	59
Alanine aminotransferase	135
Glycosylated hemoglobin	58
Creatinine	167
White blood count	25
Glucose	51
HIV RNA viral load	51
CD4^{+} lymphocyte count	66

TABLE 3

Summary of Data Completeness from ICR and DSS for Each Laboratory Value

	Individuals			Stations		
		% With Value From			% With Value From	
	# With Value From ICR or DSS	ICR	DSS	# With Value From ICR or DSS	ICR	DSS
Hemoglobin	19,848	87	91	127	98	91
Aspartate aminotransferase	18,985	86	99	127	98	94
Alanine aminotransferase	19,725	88	86	127	98	87
Glycosylated hemoglobin	4,008	80	95	121	95	93
Creatinine	19,470	86	92	127	98	91
White blood count	19,901	87	91	127	98	90
Glucose	19,535	85	96	127	98	98
HIV RNA viral load	16,654	98	63	122	98	75
CD4$^+$ lymphocyte count	16,381	98	64	122	98	59

TABLE 4

Summary of Observations in ICR and DSS Laboratory Data

	Observations					
		%				
	# Total ICR or DSS	ICR	DSS	# Overlapping	% Overlapping	Correlation Coefficient
Hemoglobin	114,200	87	82	79,112	69	0.997
Aspartate aminotransferase	76,500	85	87	55,097	72	0.997
Alanine aminotransferase	74,631	88	81	51,388	69	0.999
Glycosylated hemoglobin	9,298	60	70	2,848	31	0.991
Creatinine	103,413	87	82	71,400	69	0.990
White blood count	116,133	86	82	79,368	68	0.961
Glucose	110,196	86	82	75,531	69	0.940
HIV RNA viral load	62,873	79	45	14,799	24	0.989
$CD4^+$ lymphocyte count	63,625	76	45	12,821	20	0.875