05/20/2010

Examining the Effect of Previously Missing Blood Lead Level (BPb) Surveillance Data on Results Reported in the *MMWR (April, 2, 2004/53(12):268-270)*

Mary Jean Brown ScD, RN[1]; Curtis Blanton MS[1]; Thomas Sinks PhD[1]

National Center for Environmental Health; Centers for Disease Control and Prevention; Atlanta, GA

Introduction

Between 2000 and 2003, the District of Columbia (DC) detected very high lead concentrations in its drinking water. In February 2004, DC DOH requested that CDC assess the health effects of these elevated lead levels in DC residential tap water. DC DOH supplied to CDC available BPb test result surveillance data for 1998–2003. CDC's review found that between 2000 and 2003, BPb values ≥ 5 µg/dL declined in homes without lead service lines, while the percent of BPb test results ≥ 5 µg/dL did not decline in homes with lead water service lines. CDC's findings indicated that lead in tap water contributed to a small increase in BPb levels in DC. In 2004, those findings appeared in the Morbidity and Mortality Weekly Reports (MMWR).[1]

But the MMWR did not include a substantial number of test results from blood specimens collected in 2003. These results were missing from the surveillance data DC DOH provided to CDC in February of 2004 and consequently missing from the findings published in the April 2004 MMWR article. Either the clinical laboratory did not supply the test results to the DC Childhood Lead Poisoning Prevention Program (DC CLPPP), or the DC CLPPP did not enter the results into their surveillance database. Recently outside CDC, lead poisoning prevention advocates, and Members of Congress have raised concerns that the missing BPb test results might have resulted in an underestimation of the effect elevated drinking water lead levels had on BPb in 2003. And the 2000– 2003 longitudinal analysis findings in the MMWR article could have been likewise affected (See Figure 1). The missing data have, however, been located and CDC has acquired all known 2003 BPb test results for DC residents. To reevaluate any potential

[1] Stokes L, Onwuche NC, Thomas P, et al., Blood Lead Levels in Residents of Homes with Elevated Lead in Tap Water – District of Columbia, 2004; MMWR Weekly, April 2, 2004, 53(12); 268-270.

bias caused by under-reporting 2003 BPb tests, we compare here 1) the data used in the 2004 MMWR analysis with all currently available data, including those data available in 2004 and those reported by the clinical laboratories in 2009; and 2) the data used in the 2004 MMWR analysis with the data reported by the clinical laboratories in 2009.

FIGURE. Percentage of tests with elevated blood lead levels, by year and water-line type — District of Columbia, January 1998–September 2003

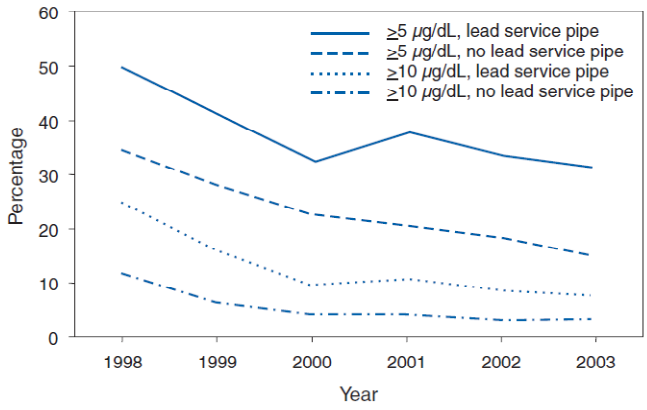


Figure 1: Test results by percent of tests above 5 and above 10 µg/dL by year and water service line type as published MMWR 2004

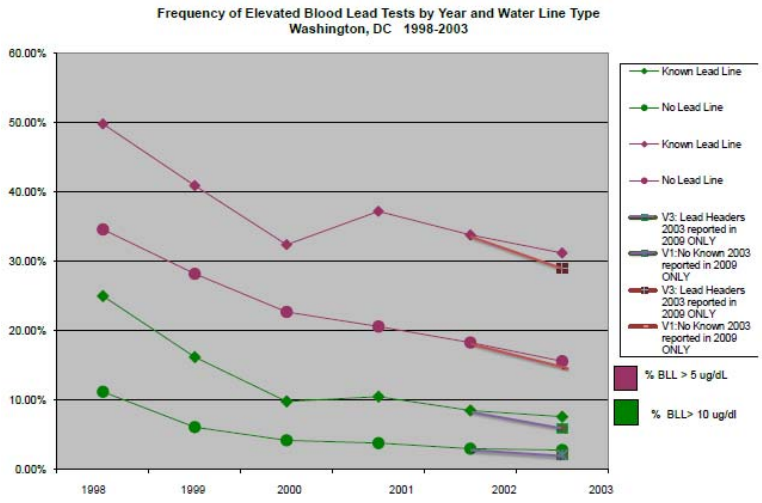


Figure 2: Test results by percent of tests above 5 and above 10 µg/dL by year and water service line type with 2003 data shown as published in 2004 and as calculated from data reported in 2009 by the clinical laboratories.

Methods

Through cooperative agreements with lead prevention programs at the state and local level, CDC provides funding for childhood lead poisoning prevention activities. In 2003, CDC funded 42 state and local health departments, including Washington, DC. In 2003, DC CLPPP was within the DC DOH; however, in 2009 it relocated to the DC Department of the Environment (DC

DOE). Under the cooperative agreement, the state or local childhood lead poisoning prevention program

- Develops and implements a strategic plan to eliminate childhood lead poisoning,
- Provides case management and environmental assessment for children identified with elevated BPb
- Implements and maintains a BPb surveillance system and requires laboratories to report BPb, and
- Develops and supports strategic partnerships that ensure the development of legislative policies that preemptively control or eliminate sources of lead in childrens' environment.

CDC staff are substantially involved in cooperative agreement programmatic activities beyond routine grant monitoring, including technical assistance and advice on surveillance and data systems, implementation of major programmatic activities, and program evaluation. CDC also approves key personnel.

On three separate occasions, CDC requested 2003 BPb test result data from DC authorities. The requested test result datasets included 1) Surveillance Datasets 1 and 2, which represented 2003 surveillance data routinely collected by the DC CLPPP and entered into the blood lead tracking database system (STELLAR), and 2) Clinical Dataset 3, which represented 2003 data provided to DC DOE by laboratories in response to a 2009 request to find missing data. The data in all three datasets represented individual test results rather than individual persons, and included results from venous blood, capillary blood, and unknown blood sample types. Although capillary samples are subject to ambient lead contamination, they have been demonstrated to provide an accurate measure of the prevalence of elevated BPb within communities. [2]

DC DOH collects childhood BPb surveillance data as required by city ordinance and reports those data to CDC, which has Institutional Review Board (IRB) and Office of Management and Budget approval to collect identifiable data for public health surveillance and response.

[2] Schlenker T, Fritz C, Mark D, Layde M, Linke G, Murphy A, Matte' T. (1994) JAMA, 271, 1346–348.

Surveillance Dataset 1

In 2003, the private laboratories sent to DC CLPPP paper records of BPb test results via facsimile or U.S. Postal Service. DC CLPPP staff manually entered into STELLAR these test results, as well as those from the DC public health laboratory. Results below the thresholds for intervention were entered as time and resources allowed. The entered data were primarily used for case management of lead- poisoned children, and priority was given to entering elevated results—defined at different times, again depending on resources, as either ≥10 µg/dL or 15 µg/dL. In March 2004, CDC requested that DC CLPPP provide all BPb tests results recorded in the STELLAR database from January 1998 through December 2003. Of the 84,929 test results that were received and used in the MMWR report, only 9,765 were from 2003.

Surveillance Dataset 2

After publication of the March 2004 MMWR, CDC became concerned about the lag time between testing and DC CLPPP's entry of screening data in STELLAR. CDC requested that DC CLPPP provide all recorded screening test results from October to December 2003. CDC received the last of these data in July 2006.

Clinical Dataset 3

A CDC programmatic review of DC CLPPP indicated that the number of test results reported by year from DC CLPPP was significantly lower in 2003 than the number reported in either 2001 or 2002. In September 2009, CDC made its third request for 2003 BPb test results. CDC requested that the DC DOE collect all BPb test results for 2003 from each laboratory known to have tested DC children for lead in 2003 and provide these data to CDC. DC DOE contacted the eight laboratories known to have analyzed blood lead levels for DC residents. The laboratories sent electronic or paper files of test results to DC DOE, which forwarded the laboratory data to CDC. CDC staff entered the paper records into the data system. CDC staff also appended the laboratory addresses to the electronic test results data provided by the DC DOH. Because the personnel responsible for lead in the laboratories and at DC DOH had completely changed, we were unable to compare laboratory 2003 reporting and data entry protocols for the data that appeared in the 2004 MMWR with protocols for the 2009 data collection.

We did nonetheless establish a chain of custody for these data. First, for security reasons we scanned the data. Then we made a copy of each laboratory's file. The original data were

migrated to a secure server, and the copies were delivered to analysts at CDC's Division of Emergency and Environmental Health Services' (DEEHS) Office the Director and to the DEEHS Healthy Homes Lead Poisoning Prevention Branch (HHLPPB).

Then the datasets were "cleaned" to remove multiple entries of the same test as well as tests on persons living outside DC. Figure 3 graphically presents the data cleaning and de-duplication process described below. (See Figure 3)

CDC cleaned each of the three datasets thusly:

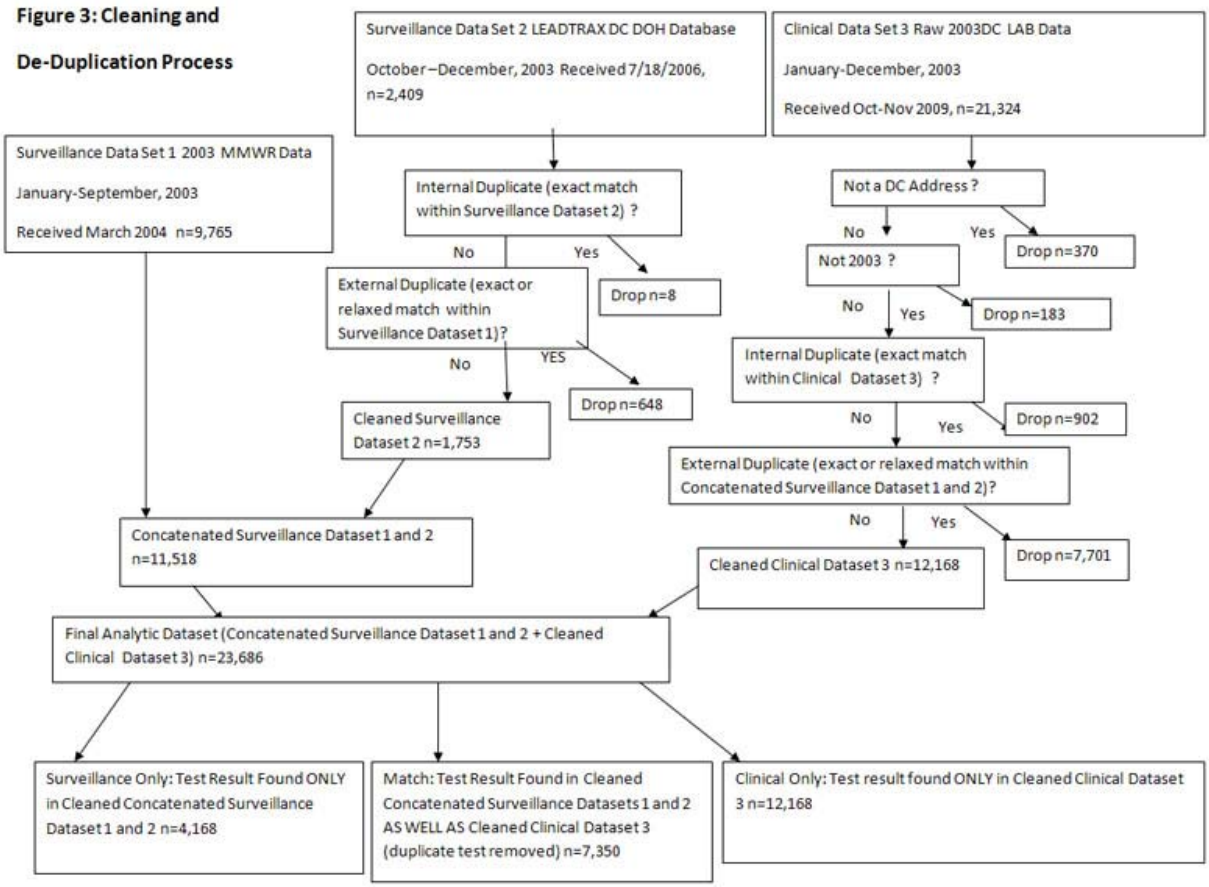
1. In March 2004, Surveillance Dataset 1 was cleaned to remove tests reported twice and to remove tests for persons who lived outside DC. The entire dataset (1998 – 2003) was cleaned at that time. No duplicate tests or non-DC residents occur in the 2003 data set used in the analyses presented here.
2. In Surveillance Dataset 2, a total of 8 tests were removed because they were reported in the dataset twice.
3. Clinical Dataset 3 included 370 tests from persons whose addresses were recorded as not living in DC and included 183 tests from years other than 2003. In addition, 902 tests were reported twice. All of these tests were removed.

The three datasets were then reviewed to identify test results reported in multiple datasets. Multiple entries of the same test were removed so that only one test result remained in the combined datasets. These were identified using matching criteria in a hierarchical process across the three datasets. This involved establishing exact and relaxed matching criteria. First, test results were identified that matched 6 fields exactly (last name, first name, date of birth, date of sample, blood lead level, and address). Then a hierarchical relaxed matching process was performed using exact matches of each combination of 5 of the 6 data fields. Duplicate test results were deleted as described below:

1. Matching criteria were first applied to match Surveillance Dataset 2 with Surveillance Dataset 1; 648 test results (593 exact matches and 55 relaxed matches) in Surveillance Dataset 2 were dropped, resulting in 1,753 unique test results in Cleaned Surveillance Dataset 2.

2. Next, the same algorithm was applied to match Clinical Dataset 3 to the Concatenated Surveillance Datasets 1 and 2 (n=11,518; 9,765 test results from Dataset 1, and 1,753 from Dataset 2). This resulted in elimination of 7, 701 Clinical

Dataset 3 test results (4, 978 exact matches and 2,723 relaxed matches). Thus 12,168 unique test results remained in Cleaned Clinical Dataset 3.



Final Analytic Dataset

The Final Analytic Dataset consists of 23,686 unique test results. Although all the tests were conducted in 2003, they were reported to CDC at three different times as described above. To evaluate whether the analytical results varied by reporting time period, we analyzed the test results by the dates when they were reported to CDC. March 2004: Surveillance Dataset 1, July 2006: Surveillance Dataset 2, and October–November 2009: Clinical Dataset 3.

We also evaluated whether the analytic results varied because the reported tests were among the data entered in the DC DOH surveillance data system. We classified each test as surveillance data only, as reported from the clinical laboratory reports only, or as reported in

both the surveillance data and the clinical laboratory reports. Each test included in the final analytic file was, without duplication, placed into one of the following three categories:

1. The "Surveillance Only" category included the 4,168 test results in the Concatenated Surveillance Dataset 1 and 2 that were **not** found in Cleaned Clinical Dataset 3.

2. The "Match" category included the 7, 350 test results found in both Concatenated Surveillance Dataset 1 and 2 and in Cleaned Clinical Dataset 3. Only one entry per test was included in the "Match" category.

3. The "Clinical Only" category included 12, 168 test results from Clinical Dataset 3 that were not found in Concatenated Surveillance Dataset 1 and 2 (note that the "Clinical Only" Category is the same as the Cleaned Clinical Dataset 3.)

The DC Water and Sewer Authority WASA provided CDC with a list of 26,155 homes presumed by WASA to have an lead water service line identified using the criteria established by, the Lead and Copper Rule. The street addresses from blood lead tests reported to CLPPP and the WASA address data were standardized using Centrus Desktop™ software version 4.02 (Sagent Technology, Mountain View, CA) and matched to the complete street address.

A CDC statistician independent of HHLPPB reviewed each step of data management and analysis.

Data Analysis

We generated frequency tables to examine the distribution of tests within the Final Analytic Dataset using the following criteria:

- When the tests were reported to CDC;
- If the tests were reported as surveillance data, or clinical laboratory data, or both;
- The percent of tests ≥ 5µg/dl or ≥ 10 µg/dl; and
- The type of water service line.

The threshold value of 5 ug/dL was selected because it represented the 95th percentile of BPb for U.S. children aged 1 to 5 in 2003–2004.[3] The threshold value of 10 µg/dL was selected

[3] Fourth National Report on Human Exposure to Environmental Chemicals. Department of Health and Human Services. Centers for Disease Control and Prevention. 2009 (page 212).

because CDC recommends individualized case management for children with BPb ≥10 µg/dL. In the analyses comparing the data reported in the MMWR (Surveillance Dataset 1) with the Final Analytic File, unknown addresses or water service line types were deleted—n=82 for Surveillance Dataset 1 and n=1571 for the Final Analytic File.

In addition, we used a chi-square test of proportions to examine whether the percent of BPb ≥ 5 or 10 µg/dL in Surveillance Dataset 1 was different from the data in the Clinical Only Dataset. We compared the distribution of BPb test results by age, year, and season for 1999–2003 in Surveillance Dataset 1 to the Clinical Only Dataset by water service line type. Tests were excluded from these analyses if they could not be linked to water service line type.

Results

Comparison of Surveillance Dataset 1 with Final Analytic Dataset

Most of the tests were from children 5 years of age or younger. About 6% of tests were for persons more than 5 years of age in any given year. Table 1 contains by dataset the distribution of the 2003 DC childhood blood lead level test results.

Table 1. DC BPb tests for 2003: Summary of datasets, data cleaning, and matching

Name of dataset	Total 2003 tests received	Duplicates within dataset	Address Not in DC	Wrong Year (not 2003)	Number of External Duplicates	Final number of Unique Tests*
Surveillance Dataset 1: Original MMWR	9,765	0	0	0	NA	9,765
Surveillance Dataset 2: Post MMWR	2,409	8	0	0	648	1,753
Clinical d Dataset 3: Raw 2003 DC lab data	21,324	902	370	183	7,701	12,168

* Tests without a match in previous datasets following the hierarchical matching process described above.

Table 2a shows the comparisons between the Final Analytic Dataset and Surveillance Dataset 1. The results reported in the 2004 MMWR were based on Surveillance Dataset 1. Because in 2003 and 2004 as much as a 90-day delay occurred in reporting and entering test results into the STELLAR database, many October–December 2003 test results were not included in Surveillance Dataset 1 and therefore not included in the 2004 MMWR longitudinal analysis. To

account for this, Table 2b compares Surveillance Dataset 1 with Final Analytic Dataset truncated at September 30, 2003.

In the Final Analytic Dataset, the percent of 2003 BPb tests ≥ 5 and ≥ 10 µg/dL in homes with lead water service lines are lower compared with Surveillance Dataset 1, which, again, we used in the 2004 MMWR report. Nevertheless, whether the entire year or the dataset truncated at the end of September 2003 is used, the findings are the same (See Tables 2a and 2b).

Table 2a. Comparison between Surveillance Dataset 1 reported in the MMWR and the Final Analytic Dataset January–December 2003

Water Service Line Type	Surveillance Dataset 1 (2004 MMWR)*	Final Analytic File **	Surveillance Dataset 1 (2004 MMWR)*	Final Analytic File**
	% ≥ 10 µg/dL	% ≥ 10 µg/dL	% ≥ 5 µg/dL	% ≥ 5 µg/dL
Lead Service Line	7.6	6.8[1.]	31.2	30.15[3.]
No Lead Service Line	2.8	2.3[2.]	15.6	14.9[4.]

* n= 9,683; ** n=21,016

[1.] p=0.351; [2.] p=0.0227; [3.] p=0.4404; [4.] p=0.184

Table 2b. Comparison between Surveillance Dataset 1 reported in the MMWR and Final Analytic Dataset truncated at September 31, 2003 January-September 2003

Service Line Type	Surveillance Dataset 1 (2004 MMWR)*	Final Analytic File truncated to September 31, 2003**	Surveillance Dataset 1 (2004 MMWR)*	Final Analytic File truncated to September 31, 2003**
	% ≥ 10 µg/dL	% ≥ 10 µg/dL	% ≥ 5 µg/dL	% ≥ 5 µg/dL
Lead Service Line	7.6	6.2[1.]	31.2	28.2[3.]
No Lead Service Line	2.7	2.2[2.]	15.6	14.3[4.]

* n=9,173; ** n=16,937

[1.] p=0.1006; [2.] p=0.0045; p=[3.]0.0502; [4.] p= 0.0089

Table 3 shows by reporting laboratory how the tests were distributed across the Surveillance Only, Match, and Clinical Only datasets. Five laboratories analyzed 95% (3,951/4,168) of the Surveillance Only test results. Four of these five laboratories contributed 7,130 tests (97%) to the tests included in the Match category. But the five that contributed most heavily to the surveillance datasets contributed only 4,978 (41%) of the tests in the Clinical Only category. Two laboratories (Laboratory 12 and UNK_2) contributed 6,900 test results (57%) to the Clinical Only dataset and no tests to either surveillance data file (Surveillance Only and Match). We also noted that Laboratory 2 reported 720 test results into surveillance data files (Surveillance Only and Match), but in 2009 only 41 of these tests were also reported to CDC. Two laboratories, Laboratory 12 and UNK_2, had no tests reported or entered in the Surveillance datasets but over 6,000 tests reported in the Clinical Laboratory dataset. Other than reporting laboratory, however, we did not find any variable that systematically predicted whether a test reported and entered into the DC surveillance datasets was also reported in the Clinical Only dataset (Table 3)

Table 3. Distribution of BPb tests by Laboratory and Match Status

	Surveillance Only		Match		Clinical Only		TOTAL
Laboratory	n	%	n	%	n	%	
Missing	53	1.3	24	0.3	0	0.0	77
1	2	0.0	1	0.0	0	0.0	3
2	679	16.3	41	0.6	0	0.0	720
3	739	17.7	1,068	14.5	1,921	15.8	3728
4	2	0.0	0	0.0	0	0.0	2
5	1,050	25.2	3,037	41.3	2,647	21.8	6734
6	606	14.5	1,905	25.9	0	0.0	2511
7	49	1.2	56	0.8	63	0.5	168
8	877	21.0	1,120	15.2	410	3.4	2407
9	1	0.0	0	0.0	0	0.0	1
10	1	0.0	1	0.0	0	0.0	2
11	0	0.0	0	0.0	95	0.8	95
12	0	0.0	0	0.0	5,562	45.7	5562
13	65	1.6	91	1.2	0	0.0	156
14	7	0.2	1	0.0	0	0.0	8
UNK LAB_1	0	0.0	0	0.0	132	1.1	132
UNK LAB_2	0	0.0	0	0.0	1,338	11.0	1338
15	37	0.9	5	0.1	0	0.0	42
Total	4,168	100.0	7,350	100.0	12,168	100.0	23,686

Comparison between Surveillance Dataset 1 and the Clinical Only Dataset

Table 4 compares the tests in Surveillance Dataset 1 (used in the 2004 MMWR) with the tests in the Clinical Only Dataset. In this comparison, the percent of elevated tests was lower in the Clinical Only Dataset, regardless of service line type. For 3 of the 4 comparisons, these differences are statistically significant.

Table 4: Comparison: Surveillance Dataset 1 as Reported in the MMWR and Clinical Dataset 3

Service Line Type	Surveillance Dataset 1 (2004 MMWR)*	Clinical Dataset 3 (Reported in 2009)**	Surveillance Dataset 1 (2004 MMWR)*	Clinical Dataset 3 (Reported in 2009)**
	% ≥ 10 µg/dL	% ≥ 10 µg/dL	% ≥ 5 µg/dL	% ≥ 5 µg/dL
Lead Service Line	7.6	6.0 [1.]	31.2	26.5 [3.]
No Lead Service Line	2.8	2.0 [2.]	15.6	13.4 [4.]

* n=9,683; ** n=10,637

[1.] p=0.09; [2.] p< 0.001; [3.] p=0.007; [4.] p< 0.001

Conclusion

In 2003, 48.6% of all known 2003 BPb test results for DC children were reported and entered into the DC CLPPP surveillance database. But more than half (12, 168) of the BPb tests conducted on DC children in 2003 were not included in the surveillance data. The vast majority of tests not included in the DC BPb surveillance system came from four laboratories, two of which contributed no tests to the surveillance system or to the 2004 MMWR analysis. We could not determine whether 1) the tests had not been reported to DC CLPPP, or 2) they had been reported and the DC CLPPP had not entered the data into STELLAR. In any event, the missing test results did not alter the direction or magnitude of CDC's previously reported 2004 MMWR findings; in fact, the percent of elevated BPb tests in the final analytic file was lower than the percent of elevated BPb tests originally reported. Thus previously missing but now-available 2003 data did not cause an underestimation for 2003 of the association between elevated blood lead levels and lead water service lines. Again, when the original 2004 MMWR data are compared with the data that had not been provided earlier, the percentages of elevated BPb values (those ≥ 5 or 10µg/dL) are lower in the previously unreported data.