IMM/Scrub: A Domain-Specific Tool for the Deduplication of Vaccination History Records in Childhood Immunization Registries

Perry L. Miller, Sandra J. Frawley, and Frederick G. Sayward

Center for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut

Received April 22, 1999

INTRODUCTION

IMM/Scrub is a pilot tool developed to assist in the deduplication of vaccination history records in childhood immunization registries. This problem is complicated by a number of factors including that fact that: (1) some doses are numbered and some are not, (2) doses may have different dose numbers, (3) doses may specify different preparations within a vaccine series, (4) one dose may indicate a combination vaccine and the other dose may specify one component of that combination, (5) two doses may have slightly different dates, and (6) combinations of any of these problems may occur together.

IMM/Scrub is designed to help detect 10 different types of vaccination dose duplicates and also allows the user to specify flexibly the conditions in which a duplicate dose might be automatically eliminated. In addition, IMM/Scrub is linked to the IMM/Serve immunization forecasting program, which can provide additional assistance in the data cleaning process. The paper describes (1) the design of the current pilot implementation of IMM/Scrub, (2) the lessons learned during its implementation, and (3) our preliminary experience applying it to data from three immunization databases, from a state, a metropolitan area, and an academic medical center.

Dealing appropriately with issues of data quality is central to the mission of an immunization registry. Only if a registry’s data is sufficiently complete and accurate, can it
be used effectively to guide patient care and to assess the adequacy of immunization levels among the registry’s patients.

One common data quality issue that arises is the presence of duplicate records. In an immunization registry, duplicate records may occur at two general levels. First, duplicates occur at the demographic level where two immunization records as a whole describe the same patient and need to be combined. This problem is encountered in many medical databases (not just those involving immunization) and in many databases outside of medicine as well. A second form of duplicate record arises within the vaccination history of a single patient. In its simplest form, this type of duplication could result from an identical vaccination dose being recorded more than once in the database, for example, when the history indicates that a vaccine was given twice on the same day.

Duplicate records can occur for many reasons. A common reason is that a patient has moved, for example, from one geographic location to another, from one provider to another, or from insurer to another. As a result, the patient may have records in several different immunization databases, which are then merged, for example, in a statewide registry. Other sources of duplication are encountered within a single registry. For example, a vaccination might be inadvertently entered twice, several days apart, once when it was administered and once when it was billed.

Duplicate records within a patient’s vaccination history, of course, are a problem specific to the immunization domain. Whereas a variety of software tools are available to assist in the detection and correction of demographic duplicate records, no general tools are available, to our knowledge, focused on the problem of vaccination history record duplicates.

As part of a cooperative research project supported by the Centers for Disease Control, we have been exploring the problem both of duplicate demographic records in an immunization database, and the problem of duplicate vaccination history records. This paper explores the latter problem in the context of three real-world immunization registries: a state registry containing over 430,000 patient records, a metropolitan area registry containing over 180,000 patient records, and an academic medical center clinic database containing roughly 7,500 patient records.

The goal of the paper is (1) to describe IMM/Scrub, a pilot domain-specific tool built to detect and help correct vaccination history errors, (2) to describe our preliminary experience using IMM/Scrub to assess the prevalence of 12 types of error and duplication in the immunization histories, (3) to discuss the lessons learned in building this tool, and (4) to describe how the approach could be extended in the future.

BACKGROUND

Duplicate records are a potential problem that must be addressed in virtually any transaction database that contains demographic data. The difficulties of deduplicating records based on demographic information is compounded by many factors, for example: (1) many people have common names, (2) records involving the same person may have the person’s name spelled differently or otherwise expressed differently, (3) names may change when women marry or for other reasons, (4) people frequently change their address, (5) for privacy reasons some registries choose not to record social security number and in any case many children do not have one, and (6) demographic data fields
may be empty or may contain error. In health care, the problem of duplicate records is particularly topical today as health care systems move to integrate patient information from many institutions to create an enterprise master patient index (EMPI). A number of computer-based tools have been developed to assist in the deduplication of records at a demographic level (4, 5).

The deduplication of records in an immunization registry has been identified as a major problem to be solved (6). A primary purpose of most registries is to contribute to the national goal of ensuring that at least 90% of 12- to 23-month-old children are fully up-to-date with recommended immunizations by the year 2000. As a registry grows in size, it receives data from increasing numbers of sources. The uncertainty resulting from unknown levels of missing, duplicate, incorrect, or otherwise problematic data also increases. Problems associated with auditing data quality and verifying potentially ambiguous data have proven to be greater than originally anticipated, and operational registries have learned to make data quality assurance a high priority and to allocate substantial personnel resources to the task (7–9). To our knowledge there has not been work reported in the literature which has focused on the deduplication of immunization history data in immunization registries.

OVERVIEW OF IMM/SCRUB’S OPERATION

The goal of IMM/Scrub is to help in the process of deduplicating a patient’s vaccination history. Figure 1 provides a schematic overview of IMM/Scrub’s operation. IMM/Scrub is currently implemented to operate in “batch” mode. It takes as input a file of immunization histories and produces several files as its output. These output files are then available for manual inspection and for further computer-based analysis. In the future, as described in a later section, we plan to build a graphical-user-interface (GUI) extension to IMM/Scrub, which will take the output files and allow the user to inspect and query them in an interactive fashion.

![Diagram of IMM/Scrub's operation]

FIG. 1. A schematic overview of IMM/Scrub’s operation as described in the text.
This section describes the major steps in IMM/Scrub’s current batch operation.

*Input to IMM/Scrub.* IMM/Scrub accepts as input a file of immunization histories. Each immunization history has (1) a patient identifier, (2) the patient’s date of birth, (3) a limited amount of coded clinical information about the patient, such as any vaccine contraindications and the mother’s HBsAg status, and (4) the vaccination history. The vaccination history consists of a set of doses, each of which indicates (1) a vaccine name (expressed as an HL7 code), (2) the date of administration, and (3) (optionally) a dose number in the vaccine’s series, for example, indicating that a particular vaccination is dose 2 in the Polio vaccine series. IMM/Scrub is also given a “forecast date,” which it will in turn pass to IMM/Serve as described below. The forecast date might be the last date that the immunization database being analyzed was updated.

*Preprocessing.* The first step in IMM/Scrub’s analysis consists of a preprocessing phase which looks for certain straightforward types of data errors. These include (1) dates which do not make sense because the day, month, and/or year are nonsensical, e.g., 11/35/98, (2) doses given before the birth date, and (3) doses given in the future (that is, after the forecast date). IMM/Scrub also looks for any problems with the referential integrity of the data, such as doses that are not linked to a patient identifier.

*IMM/Scrub’s processing of the immunization histories.* IMM/Serve then analyzes each patient’s vaccination history in turn looking for 10 different types of duplicate record, as described below in more detail. All anomalies identified are recorded in a set of error files. This analysis is guided by a set of user-defined processing parameters, which IMM/Scrub reads from a file at the beginning of each run. As described later, this file indicates (1) which types of duplicates IMM/Serve should look for in that run, (2) in what circumstances one of the duplicate dose records should be eliminated from the history, and (3) which record should be eliminated. In response to these instructions, IMM/Scrub produces a “scrubbed” history file, which contains all the doses not eliminated.

*The role of IMM/Serve.* After each patient history has been scrubbed as described above, the user may also request (via the processing parameters) that it be passed to IMM/Serve (10). IMM/Serve is an immunization forecasting program that contains the current childhood immunization recommendations of the CDC’s Advisory Committee on Immunization Practices (ACIP). IMM/Serve accepts the immunization history and produces recommendations as to which vaccinations are due (as of the forecast date) and which vaccinations should be scheduled next.

In the process of performing its analysis, IMM/Serve itself may identify a further set of errors and potential problems with the data, which are also listed in an error file. IMM/Serve’s forecast is also produced. In certain circumstances, the examination of this forecast can alert registry staff to additional problems with the data. For example, if a dose is screened as invalid by IMM/Serve because it has been given too soon after the previous dose in its vaccine series, it may be a duplicate dose not identified by IMM/Scrub.

### THE TEN TYPES OF DUPLICATION HANDLED BY IMM/SCRUB

This section describes the ten types of duplication that IMM/Scrub is currently designed to detect and help correct. In the process of describing these different types of duplications, the section also discusses several issues that make the process of correcting
duplicates considerably more complicated than one might initially expect. These complications are caused by a number of factors.

1. Some doses may have dose numbers and others may not.
2. If two doses have dose numbers, the numbers may be different.
3. Different doses may specify different preparations within a vaccine series (e.g., DT, DTP, DTaP, and Td within the D/T vaccine series).
4. One dose may indicate a combination vaccine (DTP-Hib) and the other dose may contain one of the component vaccines (Hib).
5. Two doses may indicate slightly different dates.

In addition, even though each of these factors can individually cause problems in correcting a potential duplicate dose, several of these problems may apply to a given pair of potential duplicate doses, further complicating the problem of determining how best to correct them.

1. In some cases it may be clear that the potential duplication can be resolved in a completely automated fashion. For example, if two doses specify the same vaccine (the same HL7 code), the same date, and the same dose number (or no dose number), then one of the doses can be randomly selected for automatic elimination.
2. In other circumstances, it may be clear that the problem can never be resolved automatically, and that human judgment will be required. For example, to take an extreme case, if (1) one dose specifies a combination vaccine (e.g., DTP-Hib), (2) the second dose specifies a different preparation of one of the components (e.g., DTaP), (3) the two doses have a different dose number, and (4) the two doses have dates that differ by one day, then the situation is sufficiently confused that human judgment (and possibly referring to the patient’s chart) is likely required to correct the duplication.
3. In intermediate circumstances, registry staff may be comfortable allowing a particular type of duplication be resolved automatically because they understand how the error was introduced. For example, a provider may often inadvertently enter a dose both when it is administered and a few days later when it is billed. If two doses from that provider are identical except that one is three days later, registry staff may be comfortable automatically eliminating the later dose.

The remainder of this section describes in turn each of the ten types of vaccination record duplications IMM/Scrub is currently designed to help detect and correct. The first three types of duplicates involve pairs of vaccination records both of which have the exact same date.

Test 1: Identical Vaccines on the Same Date

Here both dose records indicate that the same vaccine (same HL7 code) was given on the same date. As mentioned previously, however, the vaccine series dose number may not be same.

1. Both may have no dose number or both may have the same dose number.
2. One dose may have a dose number and the other may not.
3. Both may have a dose numbers, but the numbers may disagree.

When IMM/Scrub is run, user-defined processing parameters indicate for each duplicate test (1) whether that test should be performed in that run, and (2) exactly how each
set of possible conditions that might arise when performing the test (e.g., as outlined above for Test 1) should be handled.

If the test is run, then all potential duplicates identified will be listed in IMM/Scrub’s error files. The user also specifies for each set of conditions relevant to the test, (1) whether one of the records should be eliminated from further processing and, if so, which record, or (2) whether both records should be left in the history. If one of the records is eliminated, it is not removed from the input file. Rather, (1) it is not included in any further analysis of the history in the current run, and (2) it is not included in the “scrubbed history” file produced by IMM/Scrub.

For this particular test, it is probably reasonable to have IMM/Scrub perform as follows:
1. If both doses have no dose number or the same dose number, then eliminate one dose at random.
2. If one dose has a dose number and the other does not, then eliminate the dose with no number.
3. If both doses have a dose number and the numbers disagree, then do not eliminate either and let registry staff inspect the record to decide what to do.

On the other hand, when one is working with quite dirty data trying to identify an initial set of errors as efficiently as possible, one might wish to set the processing parameters so that in all conditions one of the records is eliminated. The advantage of this approach is that the duplication does not remain in the history to influence the later tests that are performed, including IMM/Serve’s analysis.

Test 2: Subsumed Doses on the Same Date

Here one dose record indicates a combined vaccine (e.g., DTaP-Hib) and the other indicates one of the components (e.g., Hib). With this test, several conditions arise.
1. Neither dose is numbered or both are numbered identically. Here it is presumably reasonable to eliminate the single series dose.
2. Both doses are numbered but the numbers disagree. Here it is presumably reasonable to bypass dose elimination.
3. The combination dose is numbered and the single series dose is not. Here it is presumably reasonable to eliminate the single series dose.
4. The single series dose is numbered and the combination dose is not. This condition raises several additional issues as discussed below.

IMM/Scrub’s goal is to help detect and correct duplicate doses in the vaccination history. One way in which it can help correct the duplication is by automatically eliminating one of the doses. An additional capability, not currently implemented, would be to allow IMM/Scrub to automatically synthesize a new dose record from the information provided in the two dose records being analyzed. For example, in condition (4) above where the single series dose is numbered and the combination dose is not, it might be reasonable for IMM/Scrub to synthesize a new combination dose record that includes the number from the single series dose, and then to eliminate both original dose records. In this case, the Scrubbed History file output by IMM/Scrub would contain a new dose record not included in the input files.
An additional issue raised by condition (4) is that most immunization registries allow each vaccine dose to have at most a single dose number. As more and more combination vaccines are used, however, it will increasingly be the case that such a combination may involve different dose numbers for each component vaccine series. For example, a HepB-Hib combination dose might involve HepB dose 2 and Hib dose 1. There are a variety of ways to handle this problem when designing a registry database. The database approach taken might dictate how IMM/Scrub should handle this condition for a particular registry.

Test 3: Coincident (Same Series) Doses on the Same Date

Here the two dose records indicate the same vaccine series (e.g., the D/T series) but different HL7 codes are used, indicating different vaccines in that series (e.g., DTaP, DTP, DT, or Td). With this test, the following conditions arise.

1. Neither dose is numbered or both have the same dose number. Here IMM/Scrub allows the user to specify (via the processing parameters) a preference order among the various possible HL7 codes for each series for selecting which dose to keep and which to eliminate.

2. Both doses are numbered with different dose numbers. Here it is presumably reasonable to bypass dose elimination.

3. One dose is numbered and one is not. Here it is probably best to bypass dose elimination with the current version of IMM/Scrub, since the most reasonable action might well be to synthesize a new dose containing the preferred HL7 code and the dose number.

Test 4: Identical Vaccines within a Date Window

Test 5: Subsumed Doses within a Date Window

Test 6: Coincident Doses within a Date Window

The next three duplicate tests are similar to the three outlined above, but do not involve the same date. Instead, the user may use the processing parameters to set a “date window” of 30 days or less. If the user chooses 7 days as the window, for example, then all the duplicate doses identified by these “date window” tests will have dates that differ by 1–7 days. (This date window may extend across the beginning and end of a month.)

These three tests involve similar considerations as those discussed for tests 1–3. The issues are compounded, however, by the fact that the dates are slightly different. As a result, the user may not feel comfortable using IMM/Scrub’s dose elimination to prepare a Scrubbed History file automatically, and may feel that human inspection of the record is required to resolve the duplication. As described previously, however, the user may be aware of certain systematic sources of error which make the user comfortable, for example, to eliminate the later date automatically in certain circumstances.

Test 7: Identical Vaccines within a Fixed Day Neighborhood

Test 8: Subsumed Doses within a Fixed Day Neighborhood

Test 9: Coincident Doses within a Fixed Day Neighborhood
The records of some immunization databases do not always include the day of the month when the vaccination was administered (just the month and year). When these records are merged with other immunization records (for example, in a statewide registry), an arbitrary day of the month is typically inserted into each such dose record. This day of the month might be, for example, 1 or 15. (Day 15 might be chosen since it is presumably, on average, closest to the actual date of administration.)

IMM/Scrub allows the user to use the processing parameters to set a “fixed day of the month” (such as 1 or 15) and then a “neighborhood” around that day (such as 14 days). These three tests will then identify duplicates, where one dose specifies the fixed day of the month, and the other dose specifies a different date within the same month and year, inside the specified neighborhood. (In these tests, the neighborhood does not extend beyond the beginning or end of the month.)

1. These three tests involve similar considerations as those discussed with tests 1–3, compounded by the fact that the dates are slightly different.

2. It should be noted that the duplicates identified by these tests may overlap with the duplicates identified by the “date window” tests. (If, however, the date-window duplicates were eliminated, they would not be counted here.)

3. A registry may know that only doses from certain sources have an arbitrary fixed day-of-month inserted. In addition, the registry may be able to record the source of each dose in its database. If so, it would be possible to have IMM/Scrub look for the specified fixed day-of-month only in doses that originate from that source. IMM/Scrub does not currently allow the user to specify this type of “source-specific” logic in its processing parameters, but might be extended to allow this capability in the future.

**Test 10: Identical Dose Numbers in a Vaccine Series**

This test will identify any dose pairs in the same vaccine series which have the same dose number and a different date, irrespective of how far apart the dates are. As a result, the duplicates identified by this test may partially overlap with the duplicates found by tests 4 and 7 above (if those tests did not eliminate one of the records when the dose number was the same). When inspecting the record as a whole, it may become apparent that one of the dose records identified by this test is either misnumbered or contains a data entry error involving the date of administration of the dose.

**Additional Problems Identified by IMM/Serve**

After IMM/Scrub has analyzed a vaccination history as described above, the user may specify (using the processing parameters) that the IMM/Serve forecasting program be invoked. Since IMM/Serve is quite a complex program, its use will significantly increase the time required to analyzed the input file. For example, analyzing an input file involving roughly 430,000 patient records requires roughly 15 min without IMM/Serve on a 300 MHz Pentium II machine, and roughly 3 h if IMM/Serve is invoked.

If IMM/Serve is invoked, it is called twice for each vaccination history, once with the original history and once with the scrubbed history (after duplicate doses have been eliminated as specified by the user). As described previously, IMM/Serve takes the history as
input and produces immunization recommendations. In the process, it also analyzes the history and may identify a variety of errors. These include:

1. A dose number is too high for its series.
2. Too many prior doses are specified before a numbered dose.
3. Numbered doses are out of chronological order with respect to the dates specified.
4. Two doses are too close together. (This may occur because of the presence of a duplicate dose or because a dose was administered in the real world sooner than the guidelines recommend.)

In addition, IMM/Serve will indicate, for each vaccine series and for the history as a whole, whether the data provided was “clean” enough to allow the successful generation of a forecast for that series. This information can be used in two ways.

1. The user can inspect the doses in that series to see if any additional data problems exist.
2. By comparing how many series IMM/Serve was able to analyze successfully before and after IMM/Scrub’s dose elimination, the user can get a rough sense about the level of error in the database. We anticipate that this capability could be used incrementally as a registry’s database is being cleaned to help assess how the process is progressing.

In addition, the forecast generated by IMM/Serve can also be inspected by the user to see if it makes sense or whether it suggests the possible presence of further error. We anticipate that it will be significantly easier for the user to make this type of assessment from looking at the forecast (which analyzes the data according to national guidelines and, in so doing, organizes it in a way that makes it easier to understand) than by looking at the raw history data alone.

RESULTS OF APPLYING IMM/SCRUB TO THREE IMMUNIZATION DATABASES

This section describes the results of using IMM/Scrub on a pilot basis to analyze data from three immunization databases.

A State Registry

One database, containing 431,024 patient records, is from a state that is in the process of building a statewide immunization registry. This database was in the early stages of combining data from seven different sources and was known to contain a significant level of errors and duplicate records.

A City Registry

The second database, containing 186,661 patient records, is from a city registry. This registry has been in successful operation for several years. (The city registry does not record dose numbers. The other two database do.)

An Academic Medical Center

The third database, containing 7,479 patient records, is from the pediatric clinic of an academic medical center. This immunization data was extracted from a somewhat broader
patient record system, which has been operational for roughly 2 years and which collects a variety of health maintenance data on the clinic’s patients.

Table 1 provides certain baseline data about these databases, including the number of patients and the total number of vaccination dose records. The table also indicates the number of certain types of error identified by IMM/Scrub that do not involve dose record duplication.

### Doses with Data Errors

The figure combines several types of error identified by IMM/Scrub: doses which are “orphans” in the database in the sense that they are not linked to a patient, doses with illegal dates (e.g., 11/35/98), and doses with an invalid HL7 code entered for the vaccine.

### Doses before the Date of Birth

These doses presumably involve a data entry error either for the dose date or for the patient’s date of birth.

### Doses in the Future

These are doses with dates that are later than the last update to the database.

IMM/Scrub identifies each of these errors in its error files so that registry staff can inspect the records and make corrections.

### Duplicate Dose Tests 1–3 and 10

Table 2 presents the results from running duplicate dose tests 1–3 and 10, involving identical (test 1), subsumed (test 2), and coincident (test 3) doses on the same date, and duplicate dose numbers (test 10). In performing these duplicate dose tests, we used two different dose elimination strategies.

### Conservative Dose Elimination

We made one set of IMM/Scrub runs with a “conservative” dose elimination strategy. Here we eliminated doses only in conditions where we felt that automatic dose elimination would be quite clearly appropriate.

### TABLE 1

<table>
<thead>
<tr>
<th>Baseline Data about the Three Immunization Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Patient records</td>
</tr>
<tr>
<td>Total doses</td>
</tr>
<tr>
<td>Doses with data errors</td>
</tr>
<tr>
<td>Doses before date of birth</td>
</tr>
<tr>
<td>Doses in the future</td>
</tr>
</tbody>
</table>
We made a second set of IMM/Scrub runs with a “liberal” dose elimination strategy.

Here we eliminated doses in all conditions except those in which we felt automatic dose elimination would always be inappropriate.

The purpose of using these two strategies was to assess, for each test, (1) the minimum number of duplicate dose records that could probably be eliminated automatically by a registry, and (2) the maximum number that could potentially be eliminated automatically. Exactly where along this spectrum a registry decides to allow automated dose elimination would depend on the registry’s deduplication policies, their understanding of how different types of errors were introduced and how comfortable they are with doses being eliminated without any human assessment of the circumstances.

Table 2 indicates the duplicate doses detected and those eliminated by both strategies. One interesting observation is that the number of duplicates detected by the conservative elimination strategy is occasionally slightly higher than by the liberal strategy. This is not an error, but reflects the fact that IMM/Scrub identifies pairs of duplicates and occasionally three or more duplicates of the same dose exist in the data-

### Table 2

<table>
<thead>
<tr>
<th>Duplicate doses detected (conservative</th>
<th>liberal)</th>
<th>Duplicate doses eliminated (conservative</th>
<th>liberal)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State registry (431,024 patients)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identical doses, same date</td>
<td>71,849</td>
<td>71,848</td>
<td>70,823 (99%)</td>
</tr>
<tr>
<td>Subsumed doses, same date</td>
<td>2,464</td>
<td>2,464</td>
<td>688 (28%)</td>
</tr>
<tr>
<td>Coincident doses, same date</td>
<td>5,947</td>
<td>5,933</td>
<td>5,228 (88%)</td>
</tr>
<tr>
<td>Duplicate dose numbers</td>
<td>16,600</td>
<td>15,978</td>
<td>0</td>
</tr>
<tr>
<td>Total duplicate doses above</td>
<td>96,860</td>
<td>96,223</td>
<td>76,739 (79%)</td>
</tr>
<tr>
<td>Patients involved</td>
<td>46,224</td>
<td>35,299</td>
<td>46,213 (100%)</td>
</tr>
<tr>
<td>Percentage of patients in database</td>
<td>11%</td>
<td>8%</td>
<td>11%</td>
</tr>
<tr>
<td><strong>City registry (186,661 patients)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identical doses, same date</td>
<td>1,653</td>
<td>1,653</td>
<td>1,653 (100%)</td>
</tr>
<tr>
<td>Subsumed doses, same date</td>
<td>1,931</td>
<td>1,931</td>
<td>1,931 (100%)</td>
</tr>
<tr>
<td>Coincident doses, same date</td>
<td>3,345</td>
<td>3,345</td>
<td>3,345 (100%)</td>
</tr>
<tr>
<td>Duplicate dose numbers</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Total duplicate doses above</td>
<td>6,929</td>
<td>6,629</td>
<td>6,629 (100%)</td>
</tr>
<tr>
<td>Patients involved</td>
<td>3,596</td>
<td>3,596</td>
<td>3,596 (100%)</td>
</tr>
<tr>
<td>Percentage of patients in database</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Medical Center Database (7,479 patients)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identical doses, same date</td>
<td>22</td>
<td>20 (91%)</td>
<td>22 (100%)</td>
</tr>
<tr>
<td>Subsumed doses, same date</td>
<td>28</td>
<td>14 (50%)</td>
<td>20 (71%)</td>
</tr>
<tr>
<td>Coincident doses, same date</td>
<td>31</td>
<td>5 (16%)</td>
<td>5 (16%)</td>
</tr>
<tr>
<td>Duplicate dose numbers</td>
<td>225</td>
<td>224</td>
<td>0</td>
</tr>
<tr>
<td>Total duplicate doses above</td>
<td>306</td>
<td>305</td>
<td>39 (13%)</td>
</tr>
<tr>
<td>Patients involved</td>
<td>234</td>
<td>28 (12%)</td>
<td>207 (88%)</td>
</tr>
<tr>
<td>Percentage of patients in database</td>
<td>3%</td>
<td>0.4%</td>
<td>3%</td>
</tr>
</tbody>
</table>
base. When the first pair is identified, the liberal strategy may eliminate one of the pair and the conservative strategy may not. The dose eliminated by the liberal strategy will not be available to match any further duplicates of that dose, but will be available in the conservative strategy.

As Table 2 indicates, all three registries have significant numbers of dose duplicates identified. The conservative elimination figures suggest that registries would probably be comfortable correcting a significant portion of the duplicates automatically. The level of duplicate doses is much higher in the state registry, reflecting the fact (mentioned previously) that this data set was obtained prior to the registry’s deduplication activities. Table 2 suggests that IMM/Scrub could be a valuable tool in the initial cleaning of such a database. In addition, even though the other two registries have been fully operational for some time, duplicate doses still exist, and a tool like IMM/Scrub could be of significant value in helping to detect and correct them.

**Duplicate Dose Tests 4–6**

Table 3 presents the results for running duplicate dose tests 4–6, involving identical (test 4), subsumed (test 5), and coincident (test 6) doses with dates that are not the same, but are within a user-defined number of days, the “date window” (DW). To derive these figures, we performed multiple runs with the DW set at different values. The numbers shown are cumulative, in the sense that, for example, the identical doses found with DW = 7 include those found with DW = 3.

Here again, duplicates were found in all three registries. The percentage of such doses in the city registry was higher than at the other two registries, suggesting perhaps that their data cleaning efforts had not focused heavily on this particular type of error. Even though the medical center database had very few of these duplicates, a tool like IMM/Scrub that identified these “needles in a haystack” could have definite value.

**Duplicate Dose Tests 7–9**

This section discusses duplicate dose tests 7–9, involving identical (test 7), subsumed (test 8), and coincident (test 9) duplicates where (1) one dose is on a fixed user-defined day of the month, and (2) the other dose is on a different day in a user-defined number-of-days neighborhood (N) within that month. As discussed previously, these tests were included in IMM/Scrub because some provider databases only record the month and year of a dose, and when such doses are merged into a broader registry, an arbitrary day of month (e.g., 1 or 15) is typically inserted.

To assess whether these tests might be useful for any of the three registries, we first profiled the databases to see if any day of the month had a disproportionate number of doses. As shown in Table 4, one example of this phenomenon was observed. In the state registry, day 1 of the month has roughly twice as many doses as the other days. In the other two registries, no day stood out in this fashion.

Further analysis revealed that the state database received data from seven sources. In the data from one of those sources (a large private health maintenance organization),
roughly 113,000 doses were recorded as having been administered on day 1 of the month, whereas only roughly 30,000 doses from that source were recorded for each of the other days of the month. This strongly suggests a dose from that source indicating day 1 of a month may be suspect.

As a result, we focused our analysis of tests 7–9 on the state registry (see Table 5). In running these tests, we focused only on error detection, not correction. It is interesting that Table 5 does not show a disproportionate number of duplicate doses with a fixed day-of-month of 1. This suggests that the database does not contain many duplicate records specifying other days of the month for these patients. In other databases, however, this problem may result in duplicate records that tests 7–9 could help resolve.
Incomplete Histories

Table 6 shows another analysis that IMM/Scrub performs, identifying all “incomplete histories” for a vaccine series. A series is considered incomplete under the following conditions. (1) The series contains at least one numbered dose. (It may also contain unnumbered doses.) (2) At least one numbered dose has an insufficient number of prior doses. Since the city registry database did not contain dose numbers, it is not included in this analysis.

This analysis could be useful for several purposes.

1. It allows a registry to understand the amount of this type of error in its database.
2. By running this analysis before and after IMM/Serve’s dose elimination, one can determine to what extent the dose elimination is increasing the number of incomplete histories.
3. This analysis also allows the user to see how helpful it would be to use an immunization forecasting program, such as IMM/Serve, which can accommodate incomplete histories of this sort (11).

Using the IMM/Serve Immunization Forecasting Program

The final component of IMM/Scrub’s analysis involves the option of passing the patient’s history to IMM/Serve for its forecasting analysis, which also includes identifica-
tion of several types of error. If this option is requested, the initial and the scrubbed vaccination histories for each patient are both passed to IMM/Serve. Based on IMM/Serve’s analysis, IMM/Scrub compiles statistics as to how many cases, as well as how many series, were able to be successfully run by IMM/Serve, and the reasons why series failed.

Table 7 shows certain of the additional errors that IMM/Serve is able to identify. In the IMM/Scrub runs that generated these figures, we requested dose elimination for all of IMM/Scrub’s tests except for the fixed-date tests and used a date window of 21 days. The purpose was to eliminate as many errors already identified by IMM/Scrub as possible, so that any errors identified by IMM/Serve would not be redundant.

Viewing IMM/Serve’s forecast can also be useful in data cleaning. When analyzing a large database in batch mode, it is not realistic to inspect all the forecasts. In the future, however, when we build an interactive GUI to let the user inspect individual cases, the ability to view IMM/Serve’s forecast could provide significant help in the data cleaning process.

### CURRENT STATUS AND FUTURE DIRECTIONS

The current version of IMM/Scrub is a pilot implementation built to help assess the level and types of duplicate records within several real-world immunization registries, and to explore the design issues involved in assisting to detect and correct such duplications. A next step will be to work closely with our collaborating registries to use IMM/Scrub operationally and to get their feedback as to how the tool might best be extended and refined. We currently anticipate making several such refinements.

#### TABLE 6

<table>
<thead>
<tr>
<th>Incomplete Series Identified by IMM/Scrub</th>
<th>State registry</th>
<th>Med center database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incomplete D/T series</td>
<td>31,839</td>
<td>1,522</td>
</tr>
<tr>
<td>Incomplete HepB series</td>
<td>16,885</td>
<td>1,379</td>
</tr>
<tr>
<td>Incomplete Hib series</td>
<td>12,089</td>
<td>740</td>
</tr>
<tr>
<td>Incomplete MMR series</td>
<td>5,747</td>
<td>823</td>
</tr>
<tr>
<td>Incomplete Polio series</td>
<td>28,261</td>
<td>1,290</td>
</tr>
<tr>
<td>Incomplete Var series</td>
<td>39</td>
<td>5</td>
</tr>
<tr>
<td>Patients with 1+ incomplete series</td>
<td>58,411</td>
<td>3,210</td>
</tr>
<tr>
<td>Percentage of all patients in database</td>
<td>14%</td>
<td>43%</td>
</tr>
</tbody>
</table>

#### TABLE 7

<table>
<thead>
<tr>
<th>Errors Identified by IMM/Serve</th>
<th>State registry</th>
<th>City registry</th>
<th>Med center database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose screened (given too early)</td>
<td>45,459</td>
<td>24,020</td>
<td>1,172</td>
</tr>
<tr>
<td>Dose number too big for series</td>
<td>18,043</td>
<td>NA</td>
<td>111</td>
</tr>
<tr>
<td>Too many doses prior to a numbered dose</td>
<td>9</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>Numbered doses not in chronological order</td>
<td>9,645</td>
<td>NA</td>
<td>14</td>
</tr>
</tbody>
</table>
1. There will likely be additional types of duplicate records that IMM/Scrub can help detect.

2. As described previously, it may well prove useful to allow IMM/Scrub to synthesize a new dose record, using information from two duplicate dose records, in certain circumstances when merely eliminating one of the existing dose records is not appropriate.

3. It may also be useful to allow “source-specific” processing parameters, i.e., to allow certain tests to apply only to vaccination records that originate from specified sources rather than to all the vaccination records in the database. This could be appropriate when it is known that certain systematic errors are source-specific.

4. We would also like to add an interactive GUI extension to IMM/Scrub. After the batch analysis has been performed, the user could then inspect and query the error files, examine selected records in detail, and correct those errors manually when appropriate.

A further extension of the overall approach would be to adapt IMM/Scrub so that it could take as input the vaccination histories from two EPRs and analyze the combined history for potential duplicate vaccination records. This analysis could assist in the more global process of assessing whether two EPRs involved the same patient, and could therefore supplement the record level deduplication process normally performed using demographic information. When two records looked similar based on demographic information, and automatic processing of the two vaccination histories in a combined fashion could often lend strong support as to whether the two records involved the same patient or not.

DISCUSSION AND LESSONS LEARNED

This section discusses several of the issues that arose and lessons learned during the implementation and testing of IMM/Scrub.

Domain-Specific Deduplication of the EPR

Even though previous work on the EPR has tended to emphasize the deduplication of patient records as a whole, based largely on demographic information, it is clear that there are a range of clinical domains within an individual EPR where deduplication is also an issue. The present project illustrates that domain-specific computer-based tools can assist in the deduplication of such clinical subareas within the ERP as a whole. An interesting observation, discussed previously, is that the domain-specific tool could also be used to analyze the combined records of two EPRs and thereby also assist in the more global deduplication process.

Detection vs Correction, Elimination vs Synthesis

The goal of IMM/Scrub is to help in the process of deduplicating a vaccination history. One role IMM/Scrub performs is that of detection of potential duplicates. When dealing with a large registry database, this role alone can be of significant value. In addition, IMM/Scrub is currently able to correct a duplicate by eliminating one of the duplicate records if the user so desires. A future extension would be to allow IMM/Scrub to correct a duplicate by synthesizing a new record using, for example, the HL7 code from one existing dose record and the dose number from a different existing dose record.
It is important to point out that the automatic elimination of a dose record can be performed for different reasons. It can be done to produce a partially cleaned version of the registry database automatically. It can also be done to better assess the level of error present in the data. Removing as many doses as possible allows subsequent steps (including analysis by the IMM/Serve forecasting program) to proceed in a less cluttered, more focused fashion. (Otherwise duplicates and data errors may tend to be identified several times.)

**Automatic vs Manual Deduplication of Vaccination Histories**

As described previously, the implementation of IMM/Scrub illustrates that a range of different conditions may apply to a potential duplicate identified by one of the ten duplicate record tests. For certain of those conditions, the user may be completely comfortable allowing automatic dose elimination. For other conditions, the user will never be comfortable. For a third set of conditions, the user may be comfortable with automatic dose elimination depending on the nature of the data collection process as performed by a specific registry.

**Use of an Immunization Forecasting Program to Assist in Deduplication and Data Cleaning**

A final observation is that an immunization forecasting program (such as IMM/Serve) can play several useful roles in the process of deduplication and data cleaning.

1. IMM/Serve can detect certain specific types of error beyond those currently built into IMM/Scrub.
2. IMM/Serve can also help the user find other errors by analyzing the data and creating a forecast based on national guidelines, making the data easier to view and understand.
3. Finally, by examining IMM/Serve’s internal statistics which indicate, for example, how many histories (and series within those histories) are able to be successfully processed for forecasting before and after IMM/Scrub’s processing, the user can help assess the overall level of error in the registry data.

**SUMMARY**

As the development of the EPR continues at an enterprise, regional, and national level, the problem of record deduplication will become increasingly significant. Deduplication will be necessary both using demographic information at the level of a patient record as a whole, and also at the level of the various clinical domains within an individual EPR. In fact, as discussed above, these two levels can be related because, for example, an examination of the number of duplicates in the vaccination component of two EPRs may assist in determining whether the two records as a whole might be the same.

A variety of computer-based domain-specific tools may significantly facilitate the deduplication of records within clinical domains of an EPR. This paper explores the implementation of one such tool focused on the domain of childhood immunization.

**ACKNOWLEDGMENTS**

This research was supported in part by Cooperative Grant U1W/CCU114707 from the Centers for Disease Control and by NIH Grant R01 LM06682 from the National Library of Medicine.
REFERENCES


