ABSTRACT

Although each tumor should have one valid date of diagnosis, multiple dates are often received from different reporting sources. Resolving these inconsistencies can be a labor-intensive task. To our knowledge, no algorithms for the consolidation of diagnosis dates have been published. The New York State Cancer Registry (NYSCR) has developed such an algorithm and would like to share it with other registries.

The algorithm was developed through many iterations of a trial and error process. The preliminary algorithm was designed based on our knowledge and past experience, tested using the tumors diagnosed during 2003-2009, refined based on the results of manual review from a random sample of tumors, and tested again. The reported date of diagnosis, class of case, service type, and the date of diagnosis are used by the algorithm to verify the source and the date of diagnosis. Some of the clinical codes were denormalized in the algorithm. Manual review of selected tumors by an experienced coding supervisor was performed to verify the algorithm-derived dates of diagnosis. Among 209,007 tumors with inconsistent dates from 21 sources in the NYSCR, the algorithm resolved the inconsistent dates for 95% of the tumors, leaving 5% of the tumors for manual review. Of the resolved tumors, 85% had agreement between the algorithm-derived date and the original consolidated diagnosis date, 74% of the time, the originally consolidated date was correct 17% of the time, and neither was correct 9%.

Inconsistent Diagnoses from Reporting Sources in the NYSCR

INTRODUCTION

Each tumor should have only one valid date of diagnosis; however, in practice, we often receive inconsistent diagnoses from different reporting sources. In the NYSCR, inconsistent diagnoses were received on 27% of the tumors diagnosed between 2003 and 2009 (see Table 1). Resolving these inconsistencies has been a labor-intensive and time-consuming task since it has required clerical review. The aim of this study was to develop an algorithm to consolidate automatically these inconsistent source level dates of diagnosis.

METHODS

Development of the Automated Consolidation Algorithm (A Trial and Error Method)

Setup a preliminary algorithm based on our knowledge and past experience in cancer registry

Apply the algorithm to the tumors diagnosed in 2003-2009 and having inconsistent dates from sources in the NYSCR

Perform manual review on randomly selected tumors

Modify the algorithm based on the manual review results

RESULTS - Algorithm

209,007 tested tumors with inconsistent dates of diagnosis from reporting sources were identified from the tumors diagnosed during 2003-2009 in the NYSCR.

RESULTS - Evaluation of Algorithm

Out of 209,007 tumors tested, the newly developed algorithm has resolved the inconsistent dates for 95% of the tumors (199,766), leaving 5% of the tumors for manual review.

Agreement between Algorithm-derived and Original Consolidated Dates of Diagnosis

Step | Description | Number of Tumors | Percent of Tumors
--- | --- | --- | ---
1 | When a tumor has no rank of the following three criteria, do not perform automatic consolidation | 1,009 | 0.48%
2 | When a tumor has a consolidated date of diagnosis that does not equal any of the three criteria, or both, do not perform consolidation. | 1,013 | 0.49%
3 | When a tumor has or more years difference between the earliest and the latest year of diagnosis across all the sources, do not perform. | 1,713 | 0.82%
4 | When all sources provide inconsistent diagnosis month, do not perform consolidation. Leave for the codes to consolidate manually. | 233 | 0.11%
5 | When all sources provide different year of diagnosis, do not perform consolidation. Leave for the codes to consolidate manually. | 233 | 0.11%

Note: Among those where the algorithm-derived date agreed with original consolidated dates, 25 tumors for which step 1 was randomly sampled and manually reviewed. The overall agreement was 83.3%.

RESULTS - Notes for Algorithm

Definition of Independent Reporting Sources

1. A single report from a single facility.
2. Where there are multiple reports from a single facility, all with the same date of diagnosis, count as one independent source. Choose the most recent source with the highest rank.
3. Where there are multiple reports from a single facility with different dates of diagnosis, group the sources by the earliest date of diagnosis. Within each group, choose the most recent source with the highest rank.

Example

A tumor has been reported to the registry six times. Facility A reported date 1 once; Facility B reported date 2 twice; and facility C reported date 1 twice and date 3 from C.

Apply Algorithm:

1. Rank Class of case code Service Type (when class of case code is missing)
   1. 00, 10-11, 14-15, 24 (0, 1, 4)
   2. 00, 10, 26, 31, 34-42, 42 (0, 0, 5)
   2. Duplicated, Clin, Specialty Lab, Ambulatory Care Center, 6, 8, 10
   2. Laboratory - Path, Oncology, Path, Cardiac, 7, 8, 10
   2. Laboratory - Path, Oncology, Path, Cardiac, 7, 8, 10
2. If the first consolidated date is within (including) 30 days following the date of diagnosis
3. If the first consolidated date is beyond 30 days following the date of diagnosis

Characteristics of Diagnosis Date Inconsistencies

<table>
<thead>
<tr>
<th>Characteristics of Diagnosis Date Inconsistencies</th>
<th>Tumor counts</th>
<th>Tumors affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconsistent dates in all sources</td>
<td>270</td>
<td>0.13%</td>
</tr>
<tr>
<td>One or more inconsistent dates and complete date</td>
<td>24,173</td>
<td>11.52%</td>
</tr>
<tr>
<td>One or more inconsistent dates and more than one complete date</td>
<td>12,420</td>
<td>5.92%</td>
</tr>
<tr>
<td>Complete date on all sources</td>
<td>173,044</td>
<td>82.44%</td>
</tr>
<tr>
<td>Total</td>
<td>209,907</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Different year of diagnosis for at least two sources | 36,032 | 17.39% |

Same year, but different month for at least two sources | 101,118 | 48.17% |

Same year & month, but different day for at least two sources | 71,857 | 34.23% |

Total | 209,907 | 100.00% |

Steps in Table do not indicate that the consolidated dates of diagnosis were assigned by the Algorithm.

CONCLUSIONS

- The application of the newly developed automated algorithm will greatly increase the efficiency of diagnosis data consolidation, without sacrificing the data quality.

- The application of the new consolidation algorithm improves the quality of the date of diagnosis in the NYSCR.

- There are ambiguous dates of diagnosis on some tumors due to poor diagnosis information from reporting sources.
Although each tumor should have one valid date of diagnosis, multiple dates are often received from different reporting sources. Resolving these inconsistencies can be a labor-intensive task. To our knowledge, no algorithms for the consolidation of diagnosis dates have been published. The New York State Cancer Registry (NYSCR) has developed such an algorithm and would like to share it with other registries.

The algorithm was developed through many iterations of a trial and error process. The preliminary algorithm was designed based on our knowledge and past experience, tested using the tumors diagnosed during 2003-2009, modified based on the results of manual review from a random sample of tumors, and tested again. The reported date of diagnosis, class of case, service type (a NY-specific item similar to Type of Reporting Source), date of first contact and the previously consolidated date of diagnosis were considered in the algorithm. Manual review of randomly selected tumors by an experienced coding supervisor was performed to verify the algorithm-derived dates of diagnosis.

Among 209,907 tumors with inconsistent dates from >=2 sources in the NYSCR, the algorithm resolved the inconsistent dates for 95% of the tumors, leaving 5% of the tumors for manual review. Of the resolved tumors, there was 98% agreement between the algorithm-derived diagnosis year and the original consolidated diagnosis year, 88% agreement for diagnosis year and month, and 77% agreement for diagnosis year, month, and day. For the tumors where there was agreement between the algorithm-derived dates and the original consolidated dates, manual review of a total of 225 randomly selected tumors revealed that the algorithm-derived date was correct 93% of the time. For the tumors where there was disagreement between the algorithm-derived dates and the original consolidated dates, manual review of a total of 451 randomly selected tumors revealed that the algorithm-derived date was correct 74% of the time, the originally consolidated date was correct 17% of the time, and neither was correct 9% of the time.

These results suggest that the application of an automated algorithm not only saves time and labor but also improves the quality of tumor date of diagnosis.