Development of an automated consolidation algorithm to resolve inconsistent dates of diagnosis from multiple sources

Xiuling Zhang, Amy R. Kahn, Francis P. Boscoe and Paul M. Buckley. New York State Cancer Registry, Albany, NY

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, 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.

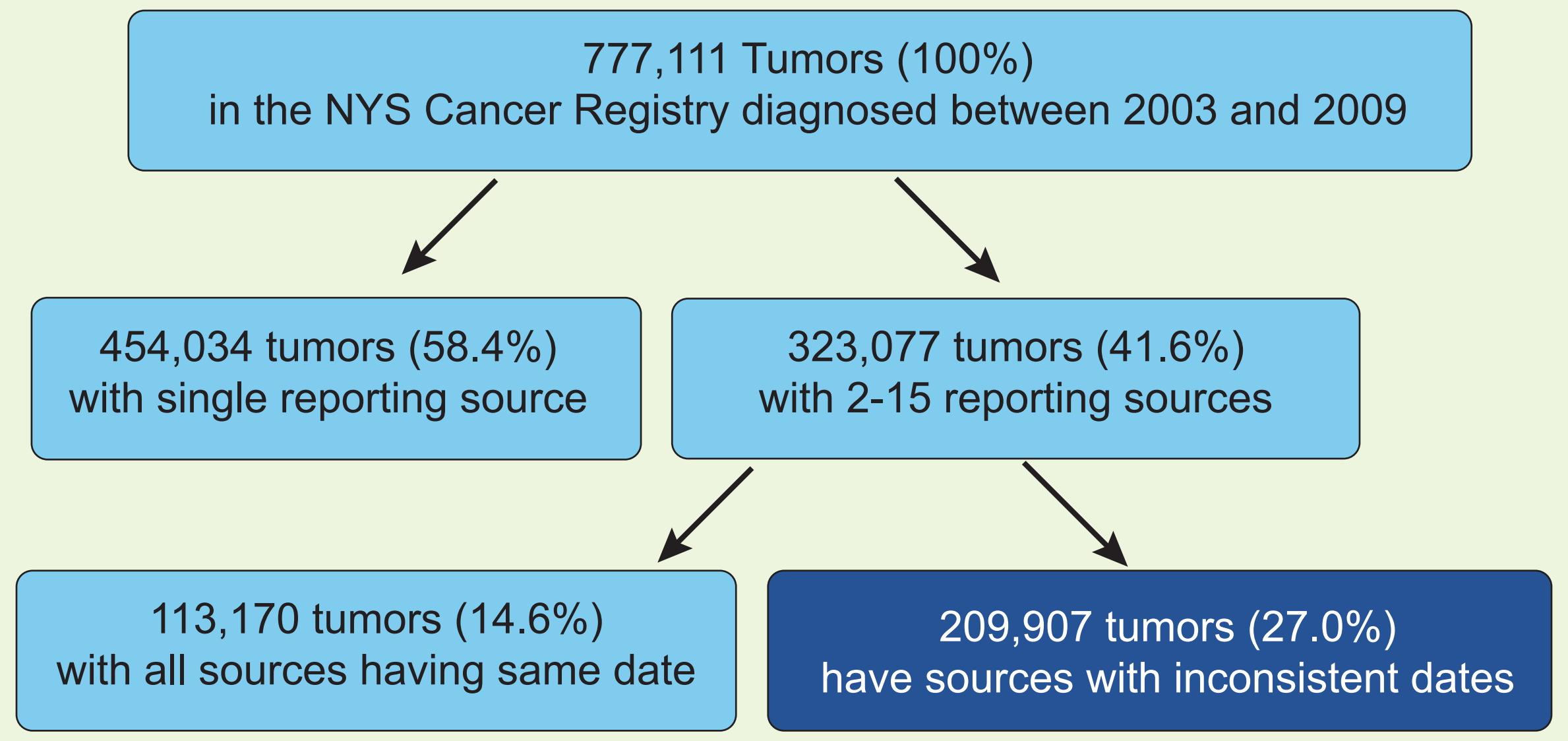
Among 209,907 tumors with inconsistent dates from >=2 sources in the NYSCR, the algorithm resolved the inconsistent dates for 95% of the tumors, leaving 5% of the tumors for manual review. Of the resolved tumors, there was 98% agreement between the algorithm-derived diagnosis year and the original consolidated 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%.

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.

INTRODUCTION

Each tumor should have only one valid date of diagnosis; however, in practice, we often receive inconsistent dates of diagnosis from different reporting sources. In the NYSCR, inconsistent dates of diagnosis were received on 27% of the tumors diagnosed between 2003 and 2009 (see below). 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.

Inconsistency of Dates of Diagnosis from Reporting Sources in the NYSCR



777,111 Tumors (100%)
in the NYS Cancer Registry diagnosed between 2003 and 2009

454,034 tumors (58.4%)
with single reporting source

323,077 tumors (41.6%)
with 2-15 reporting sources

113,170 tumors (14.6%)
with all sources having same date

209,907 tumors (27.0%)
have sources with inconsistent dates

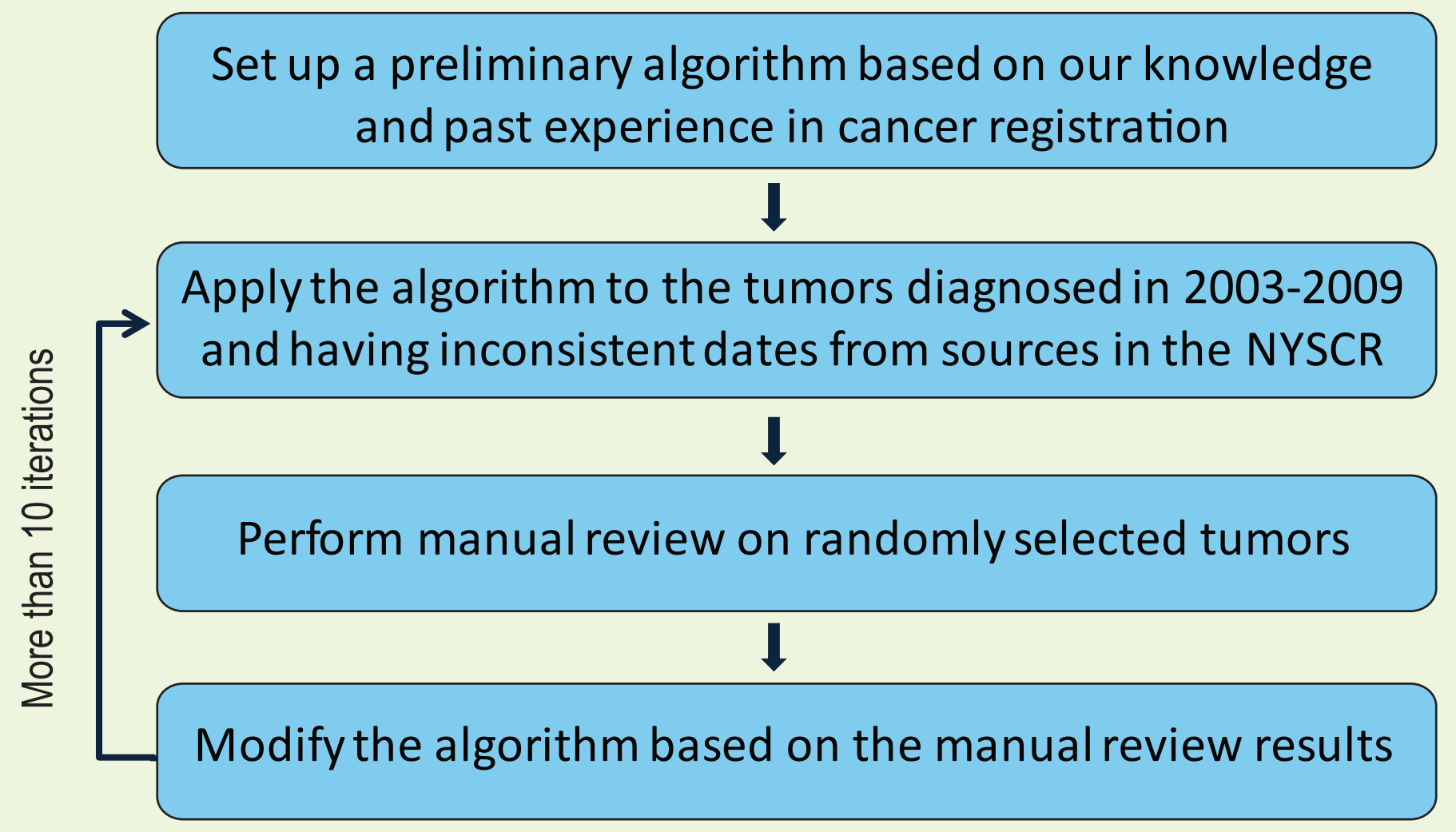
Characteristics of Diagnosis Date Inconsistencies

Inconsistency status	Tumor counts	%
Incomplete dates* on all sources	270	0.13
One or more incomplete dates and one complete date	24,173	11.52
One or more incomplete dates and more than one complete date	12,420	5.92
Complete dates on all sources	173,044	82.44
Total	209,907	100.00
Different year of diagnosis for at least two sources	36,932	17.59
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

* Incomplete date refers to a date with an unknown diagnosis month

METHODS

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



Set up a preliminary algorithm based on our knowledge and past experience in cancer registration

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

Verification of Algorithm-Derived Dates of Diagnosis

An experienced coding supervisor was asked to review randomly selected tumors from different conditions to verify the correctness of the algorithm-derived dates of diagnosis. Another experienced coding supervisor and a research scientist were also asked to review some of these random samples to see how consistent the manually picked dates were among different reviewers.

Notes for Algorithm Steps 4a & 4b

Priority Ranking of Reporting Sources Based on Class of Case and Service Type Information

Rank	Class of case code	Service Type (when class of case code is missing)
1	00, 10-14, 34,35 (old: 0, 1, 4)	Inpatient, Non-NY case, Private medical practitioner (office visit), Laboratory followback
2	20-22, 36, 37, 40-42, 32* (old: 2, 6, 3*)	Outpatient, Clinic (within Facility), Ambulatory Care Center, Radiation treatment only, DCO/followback
3	43, 30, 99, 38, 49, 31, 33, 32** (old: 7, 9, 5, 8, 3**)	Laboratory - within Facility, Consult only, Port/Cath, Unknown, Death Certificate Only, Autopsy Only (Diagnosed During), Hospice

* If the first contact date is within (including) 60 days following the date of diagnosis
** If the first contact date is beyond 60 days following the date of diagnosis

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 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 once. This would yield four independent sources: date 1 from A, date 2 from B, and date 1 and date 3 from C.

Acknowledgements: We acknowledge Beth Bernard and Laura Soloway of the NYSCR for their assistance in reviewing cases. This work was supported in part by the Centers for Disease Control and Prevention's Cooperative Agreement U58/DP000783, awarded to the New York State Department of Health through the National Program of Cancer Registries. It would not have been possible without the diligence of the dedicated Certified Tumor Registrars of the hospitals of New York or the coding of the New York State Cancer Registry.

RESULTS - Algorithm

209,907 tested tumors with inconsistent dates of diagnosis from reporting sources were identified from the tumors diagnosed during 2003-2009 in the NYSCR.

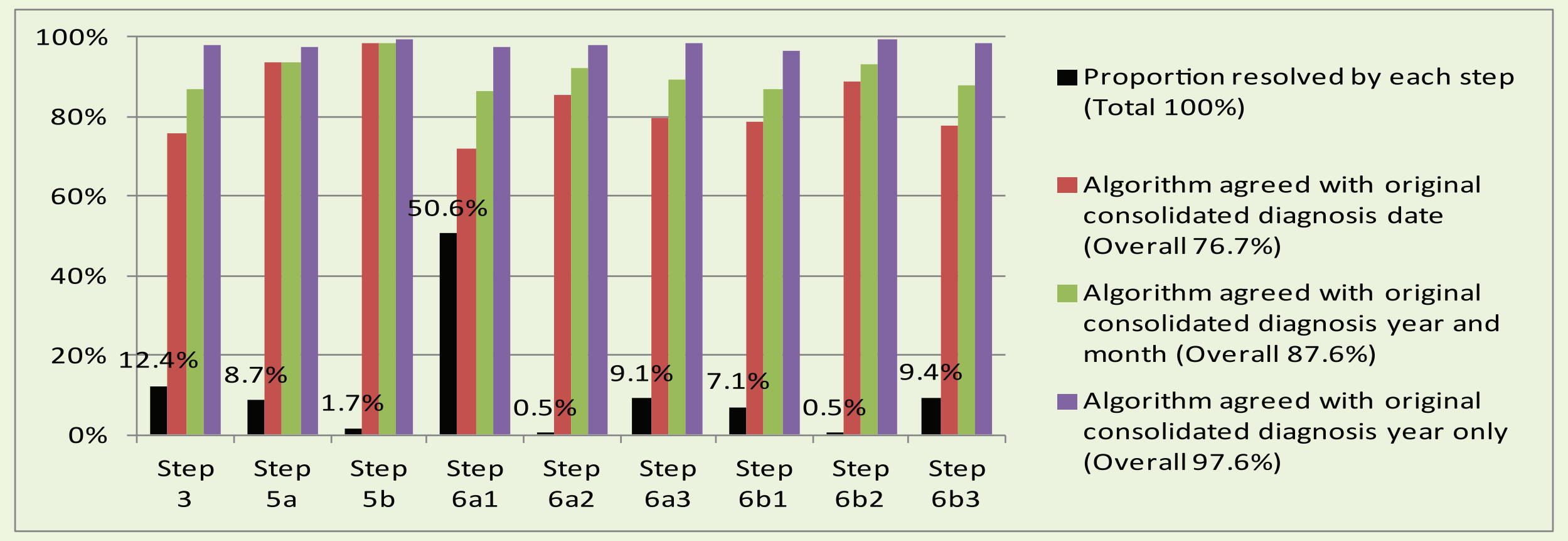
Step	Description	Number of Tumors	Percent of Tumors
1	When a tumor meets any of the following three criteria, do not perform automatic consolidation:		
1a	When a tumor has a consolidated date of diagnosis that does not equal any of the dates from the sources, do not perform consolidation. Leave it as it is.	1,611	0.77%
1b	When a tumor has a 5 or more years difference between the earliest and the latest year of diagnosis across all the sources, do not perform consolidation. Leave it for the coders to consolidate manually.	1,713	0.82%
1c	When all sources of a tumor have an unknown month of diagnosis, do not perform consolidation. Leave it for the coders to consolidate manually.	233	0.11%
2	When a tumor does not meet criteria 1a - 1c, remove the sources that have an unknown month of diagnosis on the tumor.		
3	If one or more of the remaining sources of the tumor are class of case "43", "30", or class of case missing and with the service type of "Laboratory - within Facility", "Consult only", and "Port/Cath", and the date of diagnosis from one or more of these sources is the earliest one across all the sources on the tumor, then use this earliest date as the consolidated date.	24,857	11.84%
4	When the date of diagnosis of the tumor cannot be resolved in the above steps, perform the following actions:		
4a	Assign priority ranks for the remaining sources of the tumor based on the class of case and service type information (see Notes for Algorithm).		
4b	Identify independent reporting sources from the remaining sources of the tumor according to the definition of independent reporting sources (see Notes for Algorithm).		
5	For the tumors that have one independent source, or more than one independent source but with the same date of diagnosis:		
5a	If the tumor has only one independent source, use the date of diagnosis from this source as the consolidated date.	17,376	8.28%
5b	If the tumor has more than one independent source and the different sources give the same date of diagnosis, use that date of diagnosis as the consolidated date.	3,300	1.57%
6	For the tumor that has more than one independent source and the different sources provide inconsistent dates of diagnosis, perform consolidation in the following order:		
6a	When a tumor has rank 1 independent sources, use only rank 1 independent sources:		
	a1. When there is only one rank 1 independent source on a tumor, use that date as the consolidated date.	101,096	48.16%
	a2. When there are multiple rank 1 independent sources that have the same date of diagnosis on a tumor, use that date as the consolidated date.	906	0.43%
	a3. When there are two rank 1 independent sources that have inconsistent dates of diagnosis on a tumor, the earlier date is chosen (even if the earlier date has unknown components). If the dates are consistent, the more complete date is chosen.	18,155	8.65%
	a4. When there are more than two rank 1 independent sources that have inconsistent dates of diagnosis on a tumor, do not consolidate. Leave it for the coders to consolidate manually.	666	0.32%
6b	When a tumor has no rank 1 independent sources, but has rank 2 independent sources, use only rank 2 independent sources:		
	b1. When there is only one rank 2 independent source, use that date of diagnosis as the consolidated date.	14,272	6.80%
	b2. When there are multiple rank 2 independent sources that have the same date of diagnosis on a tumor, use that date as the consolidated date.	1,009	0.48%
	b3. When there are two rank 2 independent sources that have inconsistent dates of diagnosis on a tumor, the earlier date is chosen (even if the earlier date has unknown components). If the dates are consistent, the more complete date is chosen.	18,825	8.97%
	b4. When there are more than two rank 2 independent sources that have inconsistent dates of diagnosis on a tumor, do not consolidate. Leave it for the coders to consolidate manually.	3,752	1.79%
6c	When a tumor has neither rank 1 nor rank 2 sources, but does have rank 3 independent sources, do not perform consolidation; leave it for the coders to consolidate manually.	2,136	1.02%

Steps in red indicate that the consolidated dates of diagnosis were assigned by the Algorithm.

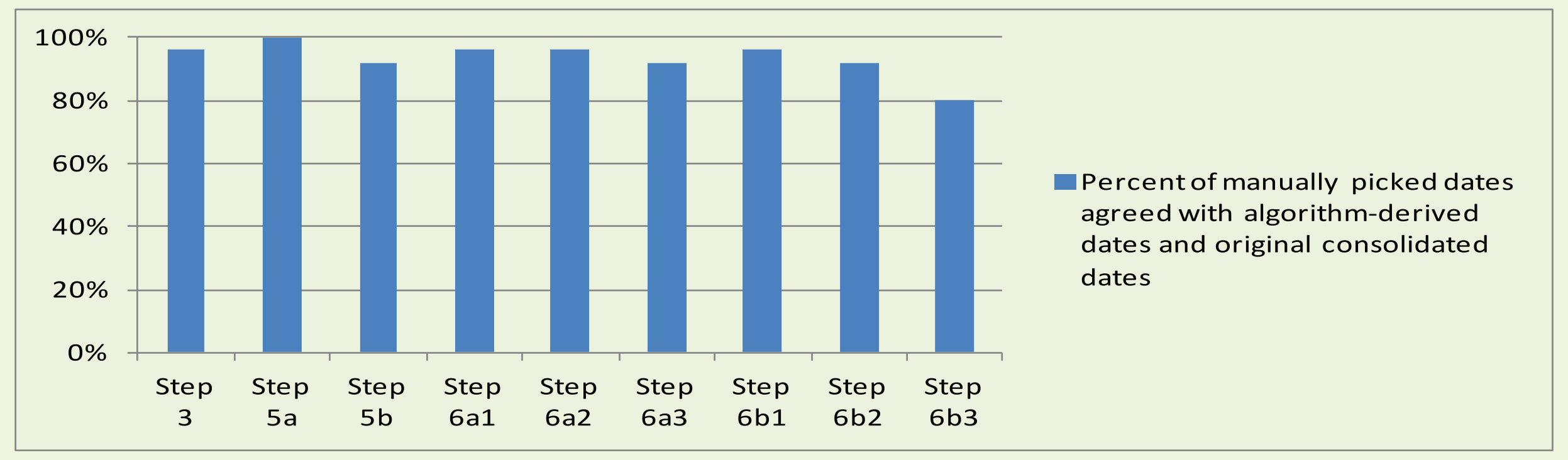
RESULTS - Evaluation of Algorithm

Out of 209,907 tumors tested, the newly developed algorithm has resolved the inconsistent dates for 95% of the tumors (199,796), leaving 5% of the tumors for manual review.

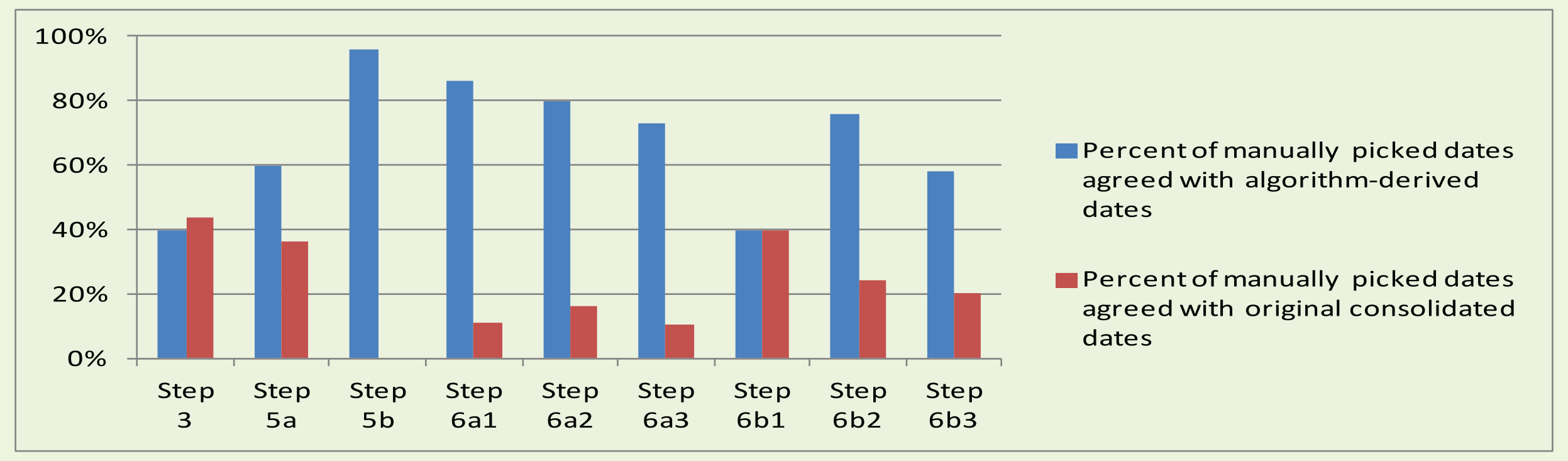
Agreement between Algorithm-derived and Original Consolidated Dates of Diagnosis



Among the tumors where the algorithm-derived dates agreed with original consolidated dates, 25 tumors for each step were randomly sampled and manually reviewed. The overall agreement was 93.3%.



Among the tumors where the algorithm-derived dates disagreed with original consolidated dates, a minimum of 25 tumors for each step were randomly sampled and manually reviewed. The overall agreement between manually picked dates and the algorithm-derived dates was 74.1% and between manually picked dates and the original consolidated dates was 17.1%.



Another coding supervisor was also asked to review 297 tumors from the above randomly sampled tumors. The overall agreement between the two coding supervisors was 81%. For the 56 tumors with discrepant dates between two coding supervisors, a research scientist was asked to review these cases again. The manually picked dates by the third person were half in agreement with the first person and half in agreement with the second person.

CONCLUSIONS

- The application of the newly developed automated algorithm will greatly increase the efficiency of diagnosis date 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

Develop
Method

Verific