Using Multiple Years of National Program of Cancer Registries (NPCR) Submission Data to Monitor Data Quality for Survival Analysis

[2]X Dong, [2]Y Ren, [2]K Zhang, [1]R Wilson

[1]Division of Cancer Prevention and Control, CDC, Atlanta, GA [2]ICF International, Fairfax, VA

Introduction:

1. Survival analysis is an essential application of cancer surveillance data. Presumed-alive assumption can be used for relative survival analyses.
2. The assumption makes the methodology sensitive to the date quality in incidence reporting, date of diagnosis (DX), date of last contact (DLC), and vital status. Pre-analysis data QC is important.
3. The study describes an informative way to use readily available historical incidence data to visualize unusual data patterns systematically and quickly, potentially before the data being submitted to NPCR.

Study Data:

1. NPCR data submissions from November 2016, 2017 and 2018 are used.
2. States with consistent patient IDs between two or more adjacent submissions

Method:

1. The study is based on the findings of two previous internal studies of NPCR: death status reporting delay causing overestimation of survival, and incidence reporting delay causing underestimation of survival.
2. The analyses are divided into 2 sets: "historical baseline", 2016 vs. 2017, and "current patterns", 2017 vs 2018.
3. In each set, cases are separated into 2 groups: existing – cases in prior and current submissions; new – cases in current but not prior submission. The groups are further categorized by vital status.
4. Cases in two adjacent submission are merged by patient ID and state.
5. Major reports for each set of analyses:
   - Existing deceased cases without DLC,
   - Existing deceased cases that changed DLC,
   - Existing deceased cases that became alive,
   - Existing deceased cases that became missing,
   - Existing alive cases that became dead,
   - New alive cases reporting patterns,
   - New deceased cases reporting patterns,
   - Corresponding patterns by site for each report
6. Patterns are assessed to establish acceptable normal pattern of each report.
7. Unusual patterns are evaluated with historical baseline to determine if the patterns are unique or persistent across submissions. Cases in the unusual patterns are extracted for further investigation
8. The result section demonstrates two pairs of patterns, normal vs. unusual. The ZONE FOR CHECKING defines a targeted area for detecting unusual patterns.

Results and Discussions (State level patterns for Submission 2017 vs. 2018):

Table 1a: Existing Deceased Cases Changed DLC in Submission 2018 of A State: Normal Pattern

	DLC Year in 2018															
DLC Year in 2017	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2003	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2004	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
2005	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2008	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2009	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1
2010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1
2011	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	1
2012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2013	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	2
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	19	11	0
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	7	259	1
2016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3
Total	0	0	1	1	1	2	1	0	3	0	3	0	0	28	278	11

Zone for checking | Diagnoal zone | Unusual pattern

- ***DIAGNOAL ZONE*** mostly contains deceased cases whose DLC month and/or day are missing in 2017 but updated in 2018, or vice versa.
- ***ZONE FOR CHECKING*** contains cases whose 2018 DLC year are different than those in 2017.
- ***UNUSUAL PATTERN*** displays concentrated groups of existing deceased cases changed DLC in 2018.
- For instance, one unusual case with DLC year 2008 in Submission 2017 was changed to DLC year 2011 in Submission 2018 (See table).
  - If the change is valid, the case will erroneously receive 3 years less survival than it should have in the 2017 survival analysis;
  - if the change is invalid, the 2018 survival analysis will erroneously assign 3 years more survival than the case should have.

Table 1b: Existing Deceased Cases Changed DLC in Submission 2018 of A State: Unusual Pattern

	DLC Year in 2018															
DLC Year in 2017	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
2002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2003	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
2004	0	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0
2005	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	0
2006	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
2008	0	0	0	0	0	0	0	5	1	0	0	1	0	0	5	0
2009	0	0	0	0	0	0	0	1	8	0	0	0	0	0	2	2
2010	0	0	1	0	0	0	0	0	2	4	1	0	0	1	8	2
2011	0	1	1	0	0	0	0	0	0	0	27	5	0	0	134	26
2012	0	0	0	0	0	0	0	0	0	3	2	4	58	7	0	8	4
2013	0	0	0	0	0	0	0	0	0	0	1	2	9	52	3	12	3
2014	0	0	0	0	0	0	0	0	0	0	0	0	2	8	51	33	0
2015	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	622	3
2016	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	9	13
Total	0	0	0	0	0	0	0	0	0	107	12	0	224	133	298	276	4

Zone for checking | Diagnoal zone | Unusual pattern

- In the ZONE FOR CHECKING, there is apparent surge of the existing deceased cases changed DLC year in Submission 2018
  - 134 cases reported as dead with DLC year 2011 in Submission 2017 changed DLC to 2015 in Submission 2018.
  - The boundaries of the region of the unusual pattern is well defined
- No unusual pattern is detected in the diagonal zone.
- Validation by state of these changes may be necessary.
- These changes may not significantly impact national level survival studies, but at state level analyses they might have some effects.

Table 2a: Existing Alive Cases Become Dead in Submission 2018 of A State: Normal Pattern

	DLC Year in 2018															
DX Year in 2017	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2001	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	389
2002	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	384
2003	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	452
2004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	409
2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	484
2006	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	462
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	501
2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	498
2009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	556
2010	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	556
2011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	700
2012	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	778
2013	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	989
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	1417
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	2036
2016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	623
Total	0	0	0	0	0	0	0	0	0	1	3	2	3	3	26	11234

Zone for checking | Unusual pattern

- The ZONE FOR CHECKING contains the cases with death status reporting delay.
  - The death status reporting delay refers to the death of a case is not reported on time to NPCR.
  - For example, a case diagnosed in 2010 whose death in 2011 should have been reported to NPCR by Submission 2014 (See table). However, the death status was not reported to NPCR until Submission 2018.
  - Consequently, the case was treated by presumed alive assumption as alive in the 2017 survival analysis with 2015 study cutoff date.
  - It was erroneously assigned 4 more years of survival than it should have.
- The death status reporting delay is a dominant biasing factor in the cancer survival analysis with presumed alive assumption according to NPCR internal studies.

Table 2b: Existing Alive Cases Become Dead in Submission 2018 of A State: Unusual Pattern

	DLC Year in 2018															
DX Year in 2017	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2001	9	12	7	6	6	20	23	26	9	5	3	3	4	5	2	44
2002	0	11	13	4	5	20	23	27	14	5	1	2	5	5	4	27
2003	0	0	7	6	5	23	25	35	11	2	4	2	7	7	3	41
2004	0	0	0	1	5	21	31	33	5	3	5	1	7	3	6	42
2005	0	0	0	0	1	22	29	55	11	6	1	4	8	4	7	42
2006	0	0	0	0	0	13	33	41	13	9	2	9	4	3	9	49
2007	0	0	0	0	0	0	15	87	18	5	4	5	9	5	11	53
2008	0	0	0	0	0	0	0	69	19	10	7	3	4	8	6	38
2009	0	0	0	0	0	0	0	0	16	10	7	7	4	9	6	40
2010	0	0	0	0	0	0	0	0	0	5	9	9	20	6	15	44
2011	0	0	0	0	0	0	0	0	0	0	6	13	17	13	17	68
2012	0	0	0	0	0	0	0	0	0	0	0	8	25	9	19	76
2013	0	0	0	0	0	0	0	0	0	0	0	0	21	14	29	85
2014	0	0	0	0	0	0	0	0	0	0	0	0	0	6	25	99
2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	181
2016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85
Total	9	23	27	17	22	119	179	373	116	60	49	66	135	97	192	1014

Zone for checking | Unusual pattern

- There are unusually more cases in Zone For Checking compared to the normal pattern.
- Especially there is an unusual group of alive cases become dead in Submission 2018:
  - Cases diagnosed between 2001-2009 and reported as alive in Submission 2017.
  - The same cases reported as dead with DLC year between 2006-2009 in Submission 2018.
- These cases experienced severe death status reporting delays.
- No such pattern was detected in the historical baseline which suggests the incidence is a one time occurrence.
- In this particular case, the delays are valid due to updates by NDI linkage.
- With the implementation of annual NDI linkage in NPCR states, the delays like this will get less frequent.

Results and Discussions:

1. As exhibited in Table 1 to 2, the method can facilitate quick visual detection of unusual data patterns and precisely outline the regions holding the cases of these patterns.
2. The historical baseline can be used to validate if the patterns are new or persistent to a state. The capability to provides boundary information of unusual patterns can be useful to researchers to extract these cases for further investigation. This may help resolve some data issues before data submission to NPCR.
3. In practice, multiple baselines can be obtained to assist in determining the onset and scope of some of the data issues which in turn may improve data collection practices.

Conclusions:

1. Using multiple years of NPCR submission data to monitor data quality for survival is a useful way to visually detect unusual case reporting patterns in areas that may impact survival estimates the most.
2. The method seems to be most effective if it is preformed before data submission to NPCR.
3. The method works well with a defined research goal, such as survival, where the subject knowledge helps to devise more meaningful data check algorithms. However, the method can potentially be applied beyond survival analysis to all data items for cancer data.
4. Further enhancement of this method can potentially benefit the states to use it as a tool for data quality control before submission.

Contact:

Xing Dong: xdong@icfi.com

Reda Wilson: dfo8@cdc.gov

National Center for Chronic Disease Prevention and Health Promotion

Division of Cancer Prevention and Control