Supplemental content

Linkage procedures

Data were submitted to six cancer registries for linkage: New York, Pennsylvania, California, Ohio, Texas, and Florida. Linkage is a process used to determine if a record in our cohort file matched to one or several records in any of the cancer registries. All six cancer registries are members of the National Program of Central Cancer Registries, and meet the Centers for Disease Control and Prevention high quality data standards.

All registries used Link Plus software. This program enables them to perform both deterministic and probabilistic linkages. In addition to linking on exact matches between records, probabilistic linkage can also match records that are not perfect matches on all variables, such as when two digits in a social security number are transposed. Link Plus computes probabilistic record linkage scores based on the theoretical framework developed by Fellegi and Sunter.[1]

Link Plus handles missing values of matching variables by treating null or empty values as missing data automatically, and allows the user to indicate additional values to treat as missing data. Link Plus also facilitates a simple and efficient blocking mechanism ("OR blocking") by indexing the variables for blocking and comparing the pairs with the identical values on at least one variable.

The matching variables used for the linkage varied. Most registries used social security number, date of birth, last, first, and middle names. The use of nicknames in the first name matching method is allowed. Link Plus provides a function for manually reviewing uncertain matches that increases the accuracy of the match, since additional data items can be reviewed.

A linkage score was assigned when the linkage configuration was created. For a comparison pair, a higher score means a higher likelihood of being a match. Records with a score below the assigned value were considered definite non-matches.

We submitted 2,017 records to each cancer registry. New York State provided 41 matches, Florida 6, and the rest occurred in the other four states.

Sensitivity analyses

The primary analysis used data from six cancer registries to identify cases, assumed that workers lived in New York State while employed at the plant, assigned gaps in the residential history by splitting the gap at the midpoint, and only considered person-time in the first (initial) risk period (i.e., workers were censored when they first left the catchment area). In the absence of complete residential histories, Bender *et al*.[2] recommended conducting uncertainty analyses to understand the limitations of the available residential history information. We conducted several sensitivity analyses to consider alternative methods or different assumptions:

S1: To evaluate the decision to expand the cancer registries beyond New York State, life-table analyses were repeated using just the New York State Cancer Registry (and defining the catchment area to be New York 1976-2007).

S2: Because state cancer registries generally will not release information about tumors only known to them through other state registries, we evaluated the potential under-ascertainment of incident cases by repeating life-table analyses additionally including deaths from malignant bladder cancer identified from our earlier mortality study that occurred in any of the six cancer registry states (Carreón et al, unpublished data, 2013) that may not have been included as cases in the primary

analysis. For these, we requested death certificates from the state vital statistics offices to identify an approximate diagnosis date, if noted on the death certificate; otherwise, date of death was used.

S3: The primary analysis was limited to person-time in the first (initial) risk period (i.e., workers were censored when they first left the catchment area). As a consequence, nine cases with diagnosis dates after the dates they first left the catchment area were excluded from the primary analysis because they were treated as censored on the date they first left the catchment area. Since others have considered disjoint risk periods when estimating SIRs,[3] we performed additional life-table analyses that considered all person-time while ever residing in the catchment area. In this analysis (that allowed for disjoint risk periods), six of these cases were considered because they were known to have re-entered the catchment area [e.g., a hypothetical worker first known to be living in Florida in 1985 and diagnosed with bladder cancer in Florida in 2000 would not be considered a case in the primary analysis because of leaving the catchment area in 1985, but would be considered a case in the supplemental analysis because of re-entering the catchment in 1997 when Florida joined the catchment area].

S4: The primary analysis assigned state of residence to gaps in the residential history by splitting the gap at the midpoint. To evaluate this decision, we repeated the life-table analyses assuming workers resided in a state until the first date known to be in another state (i.e., residence throughout the gap was assigned to the earlier state).

S5: Next, we repeated the life-table analyses assuming workers left a state just after the last date known to be in that state (i.e., residence throughout the gap was assigned to the next state).

S6: Finally, the primary analysis assumed that workers resided in New York while employed at the plant; however, given the proximity of the study plant to Canada, we repeated the life-table analyses after excluding a small number of workers (n=7) known to have lived outside of the United States.

Compared to the referent population, the incidence of bladder cancer was elevated in the cohort (SIR 2.87, 95% CI 2.02-3.96); similar elevations were observed in the six sensitivity analyses that considered alternative definitions of the catchment area, case ascertainment using death certificates and registries, and different assumptions for assigning state of residence (see supplemental table).

Supplemental Table. Bladder cancer standardized incidence ratios for primary and supplemental analyses

Analysis	No. workers	PYAR	OBS	EXP	SIR	95% CI
Primary*	1786	35,155	37	12.89	2.87	2.02-3.96
S1: Included only cases identified from the New York State Cancer Registry†	1777	34,176	34	12.27	2.77	1.92-3.87
S2: Included primary cases and cases identified from death certificate data, who resided in any of the six registry states‡	1786	35,155	38	12.89	2.95	2.09-4.05
S3: Included all risk periods§	1786	38,985	43	15.25	2.82	2.04-3.80
S4: Assigned the entire gap to the earlier state¶	1800	40,015	37	14.29	2.59	1.82-3.57
S5: Assigned the entire gap to the later state**	1631	34,371	37	13.81	2.68	1.89-3.69
S6: Excluded workers ever known to have lived outside the United States††	1779	35,095	37	12.84	2.88	2.03-3.97

*The primary analysis included cases identified using the six state cancer registries (NY, PA, CA, OH, TX and FL); split any gaps in the residence history at the midpoint and assigned the first half of the gap to the earlier state and the second half of the gap to the later state; and limited person-time at risk to the initial risk period (i.e., person-time at risk was censored at the date the worker was first known to be living outside the catchment area).

†S1 limited cases to those identified using the New York State Cancer Registry and censored person-time at risk at the date first known to be living outside of New York State.

‡S2 was like the primary analysis except that it additionally included cases from the six registry states who were identified using death certificates.

§S3 was like the primary analysis except all risk periods were included (i.e., all person-time at risk in the catchment area contributed to the denominator).

¶S4 was like the primary analysis except gaps in the residence history were assigned to the earlier state.

**S5 was like the primary analysis except gaps in the residence history were assigned to the later state.

††S6 was like the primary analysis except workers known to have lived outside the US were excluded.

PYAR, person-years at risk; OBS, observed cases; EXP, expected number of cases based on rates for New York State excluding New York City; SIR, standardized incidence ratio; CI, confidence interval.

REFERENCES

1. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969;64:1183-210.
2. Bender TJ, Beall C, Cheng H, Herrick RF, Kahn AR, Matthews R, et al. Methodologic issues in follow-up studies of cancer incidence among occupational groups in the United States. *Ann Epidemiol* 2006;16:170-9.
3. Bender TJ, Beall C, Cheng H, Herrick RF, Kahn AR, Matthews R, et al. Cancer incidence among semiconductor and electronic storage device workers. *Occup Environ Med* 2007;64:30-6.