Evaluation of Commercial Data Sources for Obtaining Individual Residential Histories for Cancer Research

Dave Stinchcomb[1], Zaria Tatalovich[2], Mandi Yu[2], Allison Roeser[1]

[1] Westat, Inc. [2] National Cancer Institute

June 16, 2016

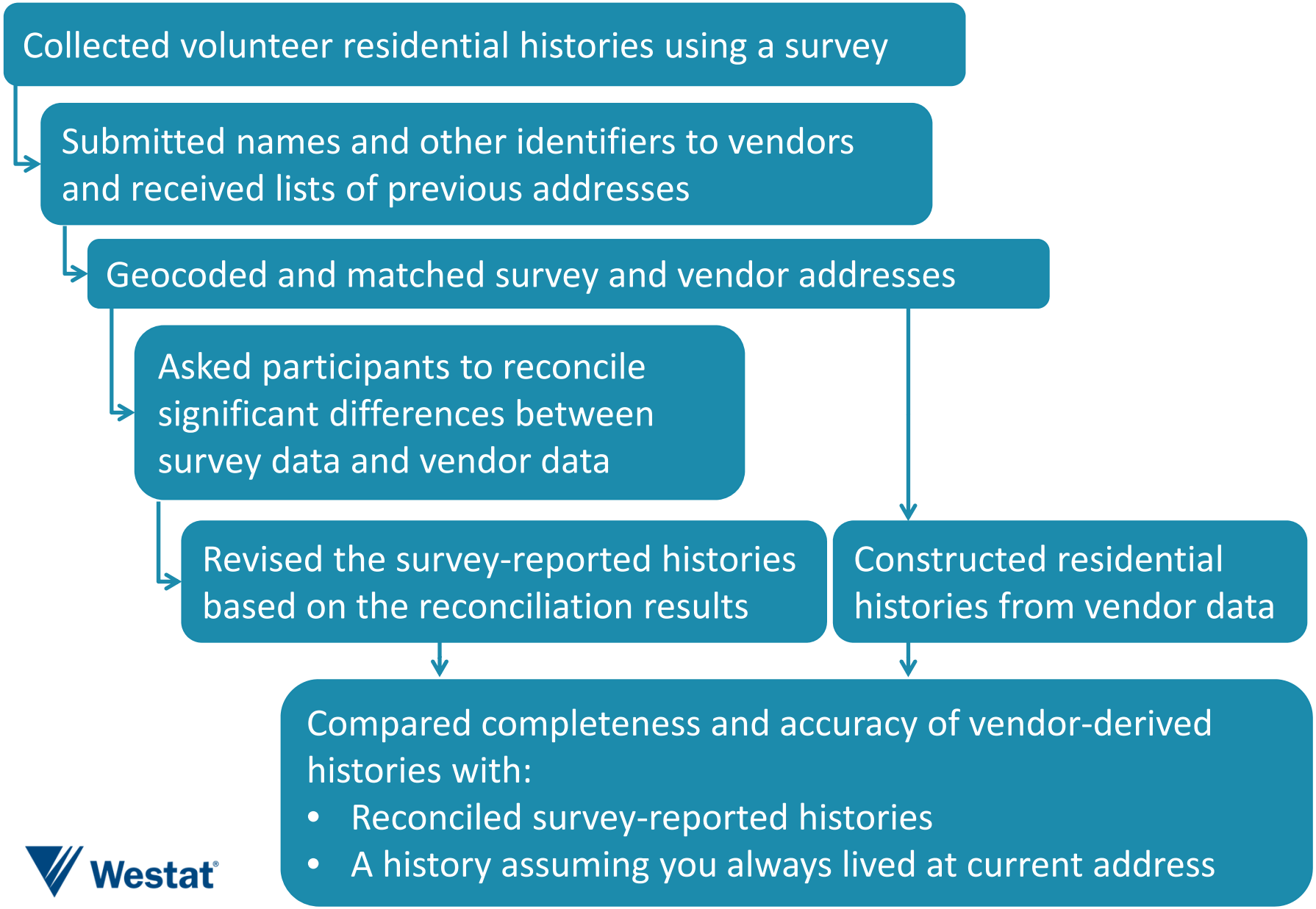
Introduction

- Interest in sources of residential history data to link with cancer registry data
- Westat was engaged to:
  - Conduct an environmental scan to identify possible sources of residential histories
  - Evaluate the quality of the identified sources
- Environmental scan results
  - Identified and evaluated over 100 possible data sources
    - Big data and social media sources
    - Open data providers
    - Commercial data providers
  - Found ten likely candidates – all commercial data providers
  - Only three could actually provide the needed data

Goals and Research Questions

- Goals:
  - To evaluate the accuracy and completeness of the three commercial data sources for linkage with cancer registry data
  - To evaluate the feasibility of generating residential histories for cases in cancer registries
- Research questions:
  - How does the quality of the different vendors compare?
  - Is there value in combining data from multiple sources?
  - Is there data available on deceased individuals for cancer research studies involving highly fatal cancers?
  - How does the quality vary by time period?

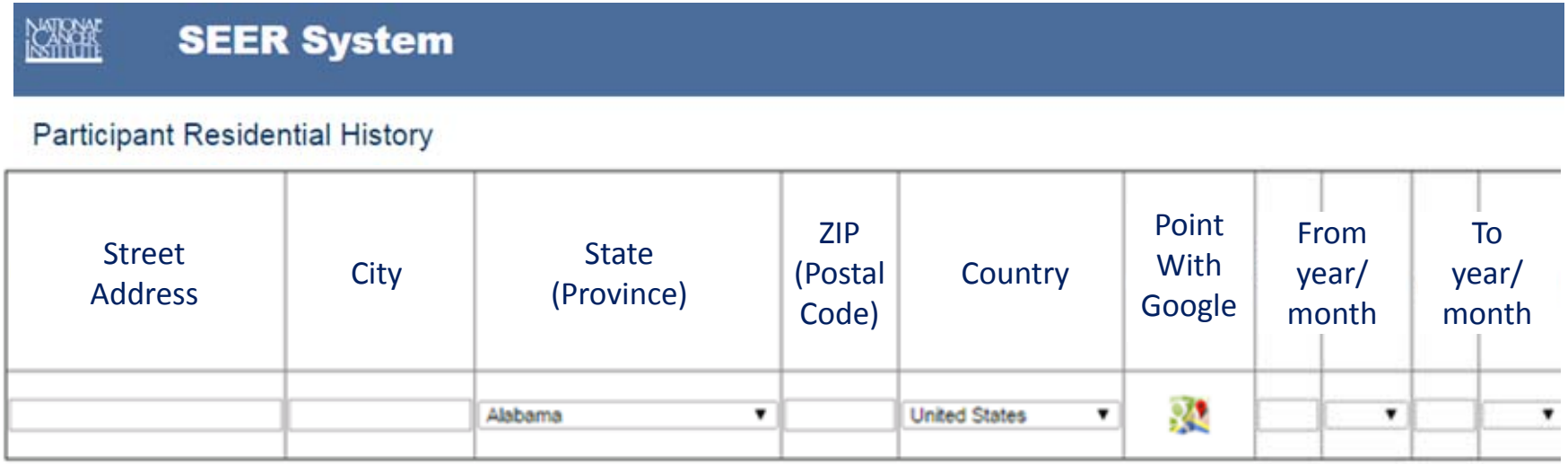
Methods Overview



Westat

Survey to Collect Residential Histories

- Volunteers from NCI and NIEHS
- Asked to provide their life-time residential history
  - Optionally, also provide the history for a deceased relative
- Web-based data collection
  - Individual authenticated login
  - A series of addresses with start and end dates
  - Optional Google Maps “point and click” interface
- Received 66 residential histories (10 deceased relatives)



Vendor Data Requests

- Sent identifiers of participants to vendors:
  - First and last names, date of birth, SSN*, current address
    * Vendor 2 would not accept SSNs
  - These data elements are typically available in registries

- Different types of information returned by each vendor:

Vendor	Address Information	Other Information
Vendor 1	Current and all known previous addresses with from and to dates for each (month and year)	Other names, phone numbers, date of death if deceased
Vendor 2	Current and up to 5 previous addresses with a single effective date (start date) for each	
Vendor 3	Current and previous addresses with from and to dates for each (day, month, and year)	Date of death if deceased

Westat

Person-Match Rates

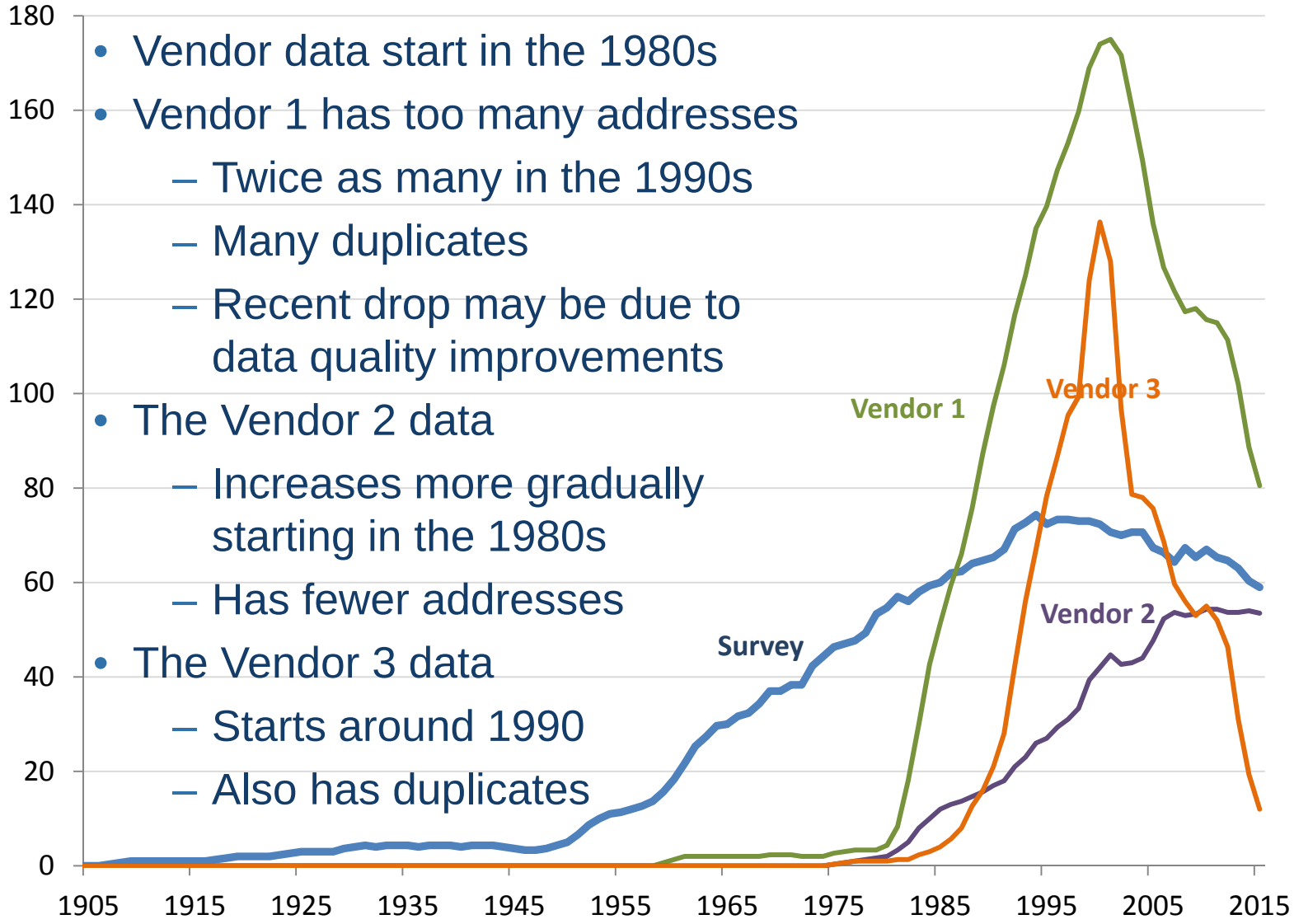
	Total number of individuals		Living individuals		Deceased individuals	
	N	Pct.	N	Pct.	N	Pct.
Survey	66	100%	56	100%	10	100%
Vendor 1	64	97%	56	100%	8	80%
Vendor 2	52	79%	48	86%	4	40%
Vendor 3	57	86%	50	89%	7	70%

- Vendor 1's match rate is very high (97%)
- Vendors 2 and 3 have good match rates (79-86%)
- All vendors have data for deceased individuals
- No evidence of any false positive matches (there were always some common addresses)

Westat

Total Addresses per Year

- Vendor data start in the 1980s
- Vendor 1 has too many addresses
  - Twice as many in the 1990s
  - Many duplicates
  - Recent drop may be due to data quality improvements
- The Vendor 2 data
  - Increases more gradually starting in the 1980s
  - Has fewer addresses
- The Vendor 3 data
  - Starts around 1990
  - Also has duplicates



Westat

Reconciliation

- Asked participants to review discrepancies between survey-reported data and vendor data
- 52 responses (79%)
- Answers to 335 specific questions
- Summary of results (all vendors)
  - Vendor address details were correct 77% of the time
  - 23% of unreported addresses should have been included
  - 27% of divergent vendor start dates were correct

Westat

Constructing Residential Histories

- Algorithm to derive a residential history from vendor address information
  - Resolve duplicates and conflicts
  - Fill time gaps
- Basic steps:
  1. Match addresses within and across vendors
  2. Combine matched addresses from all of the vendors
  3. Decide on a time-frame for each address
     - Optionally trim time-frame based on vendor frequencies
  4. Weed out short duration addresses
  5. Build residential history working backwards from the most recent address
     - Use current start date as end date for the previous address

Westat

Comparison Measures

- Completeness:
  - Proportion of time with survey-reported locations that we also had locations from vendor data ("coverage")
- Accuracy based on distance:
  - Time-weighted distance error: mean, median, percentiles
  - Within distance thresholds: 0 km, 1 km, 5 km, 10 km
- Accuracy based on changing geographic areas:
  - In different census tracts
  - In different ZIP codes
  - In different counties
- Comparison with a history based on the assumption that people have lived at their current address all of their lives

Westat

Full Life Span Comparisons by Vendor Combo

Full life span:

Vendors used	Percent time period coverage	Percent of covered time distance=0	Percent of covered time within 1 km	Percent of covered time within 5 km	Percent of covered time within 10 km
V1	58.7	56.0	68.4	74.0	81.3
V2	35.4	49.1	56.7	70.0	76.3
V3	35.5	31.6	77.0	84.0	90.3
V1,V2	58.9	50.0	64.3	71.0	78.5
V1,V3	58.8	25.2	65.9	73.1	80.3
V2,V3	46.9	26.2	68.1	78.1	84.0
V1,V2,V3	59.1	28.4	64.2	72.4	80.1
Cur Res	NA	23.9	27.6	32.5	37.4

Westat

Vendor 1 Comparisons by Time Period

V1 alone:

Time span	Percent time period coverage	Percent of covered time distance=0	Percent of covered time within 1 km	Percent of covered time within 5 km	Percent of covered time within 10 km
Full life span	58.7	56.0	68.4	74.0	81.3
1986 to 2015	89.7	57.5	69.9	75.8	82.0
1996 to 2015	96.8	63.2	75.4	80.6	86.2
2006 to 2015	98.9	72.6	83.2	87.8	90.1

Current residence:

Time span	Percent time period coverage	Percent of covered time distance=0	Percent of covered time within 1 km	Percent of covered time within 5 km	Percent of covered time within 10 km
Full life span	NA	23.9	27.6	32.5	37.4
1986 to 2015	NA	36.9	40.3	47.6	53.2
1996 to 2015	NA	49.0	52.0	59.5	64.8
2006 to 2015	NA	68.1	71.2	76.2	78.2

Westat

Conclusions – Data Availability

- Data seem to start in the 1980s – very little before then
- All vendors have data on deceased relatives, Vendors 1 and 3 have more complete data
- Mostly U.S. addresses only – no foreign addresses
  - There are some military APO addresses (not used)
- All vendors had good person-match rates – no evidence of false positive matches
- Vendor address data includes duplicates, conflicting information, and gaps in time
  - An algorithm is needed to construct a plausible residential history from the data
- Reconciliation of survey-reported histories with vendor data can improve survey-reported histories

Conclusions – Accuracy and Completeness

- Vendor 1 has the best coverage
- Vendor 1's accuracy is reasonably good
  - Significant improvement over the current-residence assumption
  - More accurate for more recent time periods
    - But there is less of an improvement over the current-residence assumption
- Including data from Vendors 2 and 3 did not improve Vendor 1 results

Limitations

- Small sample size
- Sample of convenience – not representative
  - College educated, most have advanced degrees
  - Middle class / upper middle class
  - Limited range of ages – no one who is very young
  - People currently living in the DC area or North Carolina: majority of the addresses from the Eastern states
  - More foreign born individuals than U.S. average
  - Key issues: we don't know about vendor data quality for:
    - The very poor – folks without credit cards and mortgages
    - Children, teens, young adults
    - Older people where someone else has power-of-attorney
- The history generation algorithm may not be optimal
  - Could seed the algorithm with known previous addresses (e.g., address at diagnosis)

Use by Cancer Researchers

- Cancer registries have patient identifiers (names, SSNs)
  - IRB process to get access for a research study
  - The registry, the researcher, or a third-party broker could get vendor data and construct residential histories
- Plan to make the algorithm for constructing residential histories available for registries/researchers
  - Make available as a SAS program
  - Preprocessing of vendor data: need to geocode
  - Address matching: option to manually review possible matches
  - See https://gis.cancer.gov/tools/residential-histories.html
- Residential history research objectives have been added to NIH/NCI **PA-16-175**: Exploratory Grants in Cancer Epidemiology and Genomics Research (R21)

Westat

Summary

- Commercial data can be used to generate residential histories for cases in cancer registries
  - An algorithm in needed to process vendor address data
  - Data is available for deceased individuals
- Residential histories derived from commercial data are reasonably complete and accurate
  - Substantially better than assuming that people never move
  - Data start in the mid 1980s
  - Primarily just U.S. addresses
  - Currently, Vendor 1 has the most complete and accurate data
  - Including data from Vendors 2 and 3 did not improve the results

Acknowledgements

- NCI:
  - Zaria Tatalovich
  - Mandi Yu
  - Rocky Feuer
  - Li Zhu
  - Benmei Liu
  - 33 volunteers who provided residential histories
- Westat:
  - Allison Roeser
  - Katie Genoversa-Wong
  - Kiran Amireneni
  - Stephen Leard
  - Michael Giangrande
- NIEHS:
  - Aubrey Miller
  - April Bennett
  - 23 volunteers who provided residential histories

Westat

Thank You.

DavidStinchcomb@westat.com

Westat