Missing Data: An Analysis of Lacks and Leaks in Health Data Collection

By: Taussia Boadi and Nino Owens

Mentor: Dr. Julia Lynch, PhD

Project Aims

- Understand the political circumstances behind government data collection
- Understand how their data collection strategies around socio-demographics allow them to see or be blind to health inequities

Significance

- Why do we care about health data collection?

“No Data, No Problem”

- Health inequities and disparities are seen through the collection and comparison of health status and mortality data

Data Lacks

- “Governments (decide to) lack certain categories of sociodemographic data when possessing (or reporting) these data would exacerbate a potentially explosive social cleavage.”

Lynch, 2022

Data Leaks

Phenomenon where governments fail to publicize certain routinely collected variables

How is Health Data Collected

US Health Data Collection

- Vital Records and Death Certificates
- Census
- Self-reported Surveys
- Medical Records
- Cohort Studies
- Disease Registries

Archival Research and Qualitative Coding

Understanding and Analyzing Health Data

Archival Research

- The process of extracting information from archival records, often data files from companies and organizations

Source Requirements

- Data collected by or at the request of a unit of the central government
- Data collected serially at least 4 times
- Data concerns the entire population of the country, not subgroups
- Data collected concerns multiple health topics, not a single disease

Qualitative Coding

- The process of searching for and identifying relationships, connections, or trends in text, media, and other data items

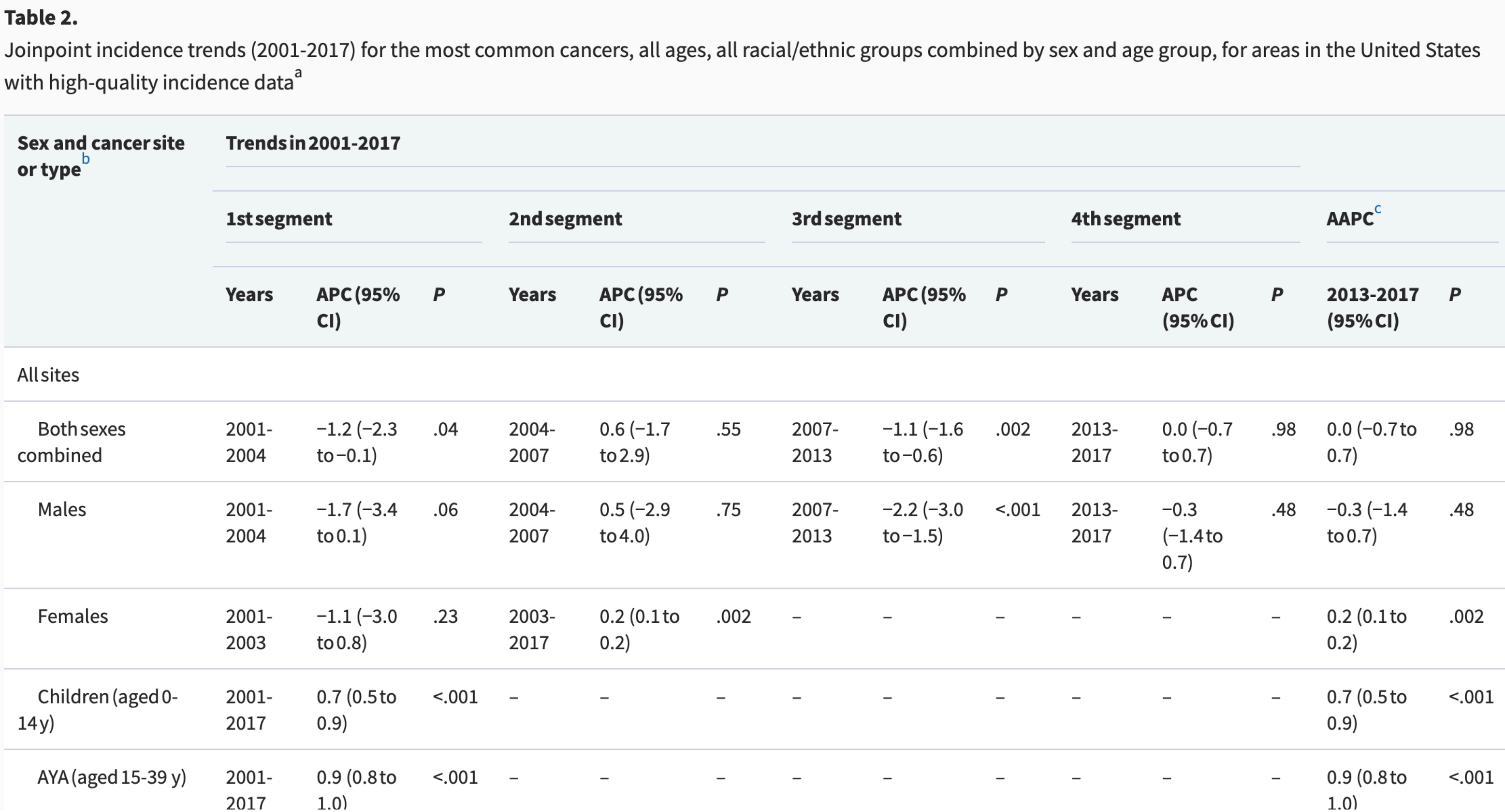
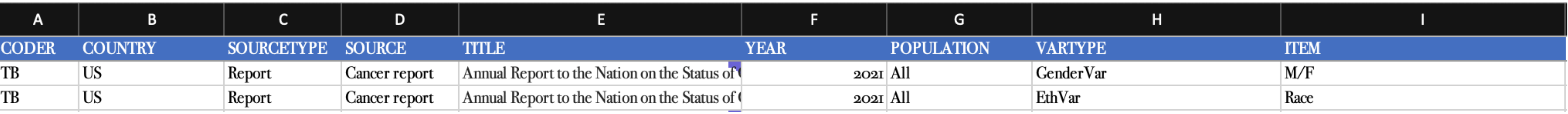


Table 2.

Joinpoint incidence trends (2001-2017) for the most common cancers, all ages, all racial/ethnic groups combined by sex and age group, for areas in the United States with high-quality incidence data[a]

Sex and cancer site or type[b]	Trends in 2001-2017														
	1st segment			2nd segment			3rd segment			4th segment			AAPC[c]	
	Years	APC (95% CI)	*P*	Years	APC (95% CI)	*P*	Years	APC (95% CI)	*P*	Years	APC (95% CI)	*P*	2013-2017 (95% CI)	*P*
All sites														
Both sexes combined	2001-2004	−1.2 (−2.3 to −0.1)	.04	2004-2007	0.6 (−1.7 to 2.9)	.55	2007-2013	−1.1 (−1.6 to −0.6)	.002	2013-2017	0.0 (−0.7 to 0.7)	.98	0.0 (−0.7 to 0.7)	.98
Males	2001-2004	−1.7 (−3.4 to 0.1)	.06	2004-2007	0.5 (−2.9 to 4.0)	.75	2007-2013	−2.2 (−3.0 to −1.5)	<.001	2013-2017	−0.3 (−1.4 to 0.7)	.48	−0.3 (−1.4 to 0.7)	.48
Females	2001-2003	−1.1 (−3.0 to 0.8)	.23	2003-2017	0.2 (0.1 to 0.2)	.002	–	–	–	–	–	–	0.2 (0.1 to 0.2)	.002
Children (aged 0-14 y)	2001-2017	0.7 (0.5 to 0.9)	<.001	–	–	–	–	–	–	–	–	–	0.7 (0.5 to 0.9)	<.001
AYA (aged 15-39 y)	2001-2017	0.9 (0.8 to 1.0)	<.001	–	–	–	–	–	–	–	–	–	0.9 (0.8 to 1.0)	<.001

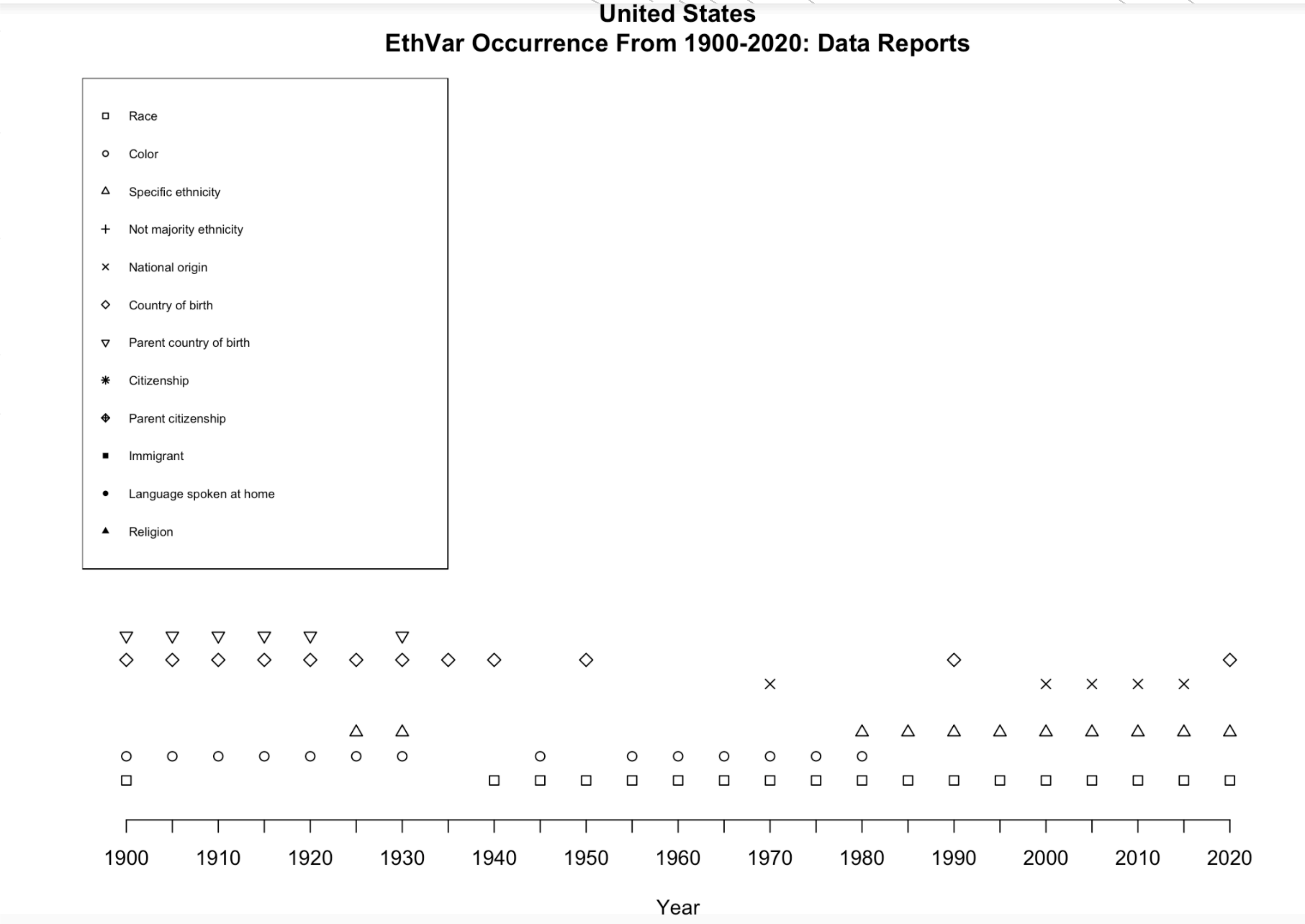
Qualitative Coding



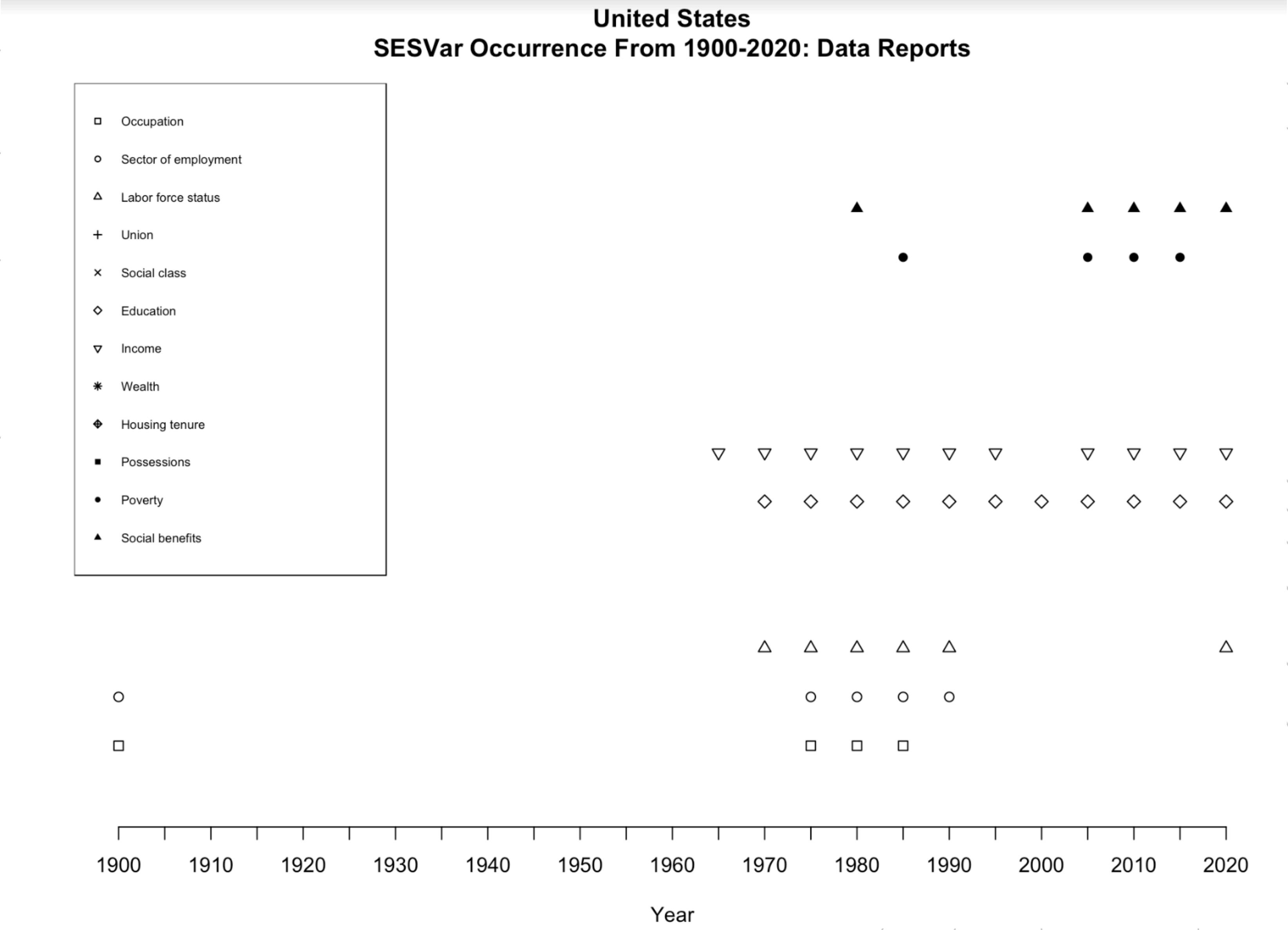
A	B	C	D	E	F	G	H	I
CODER	COUNTRY	SOURCETYPE	SOURCE	TITLE	YEAR	POPULATION	VARTYPE	ITEM
TB	US	Report	Cancer report	Annual Report to the Nation on the Status of	2021	All	GenderVar	M/F
TB	US	Report	Cancer report	Annual Report to the Nation on the Status of	2021	All	EthVar	Race

What We Look For

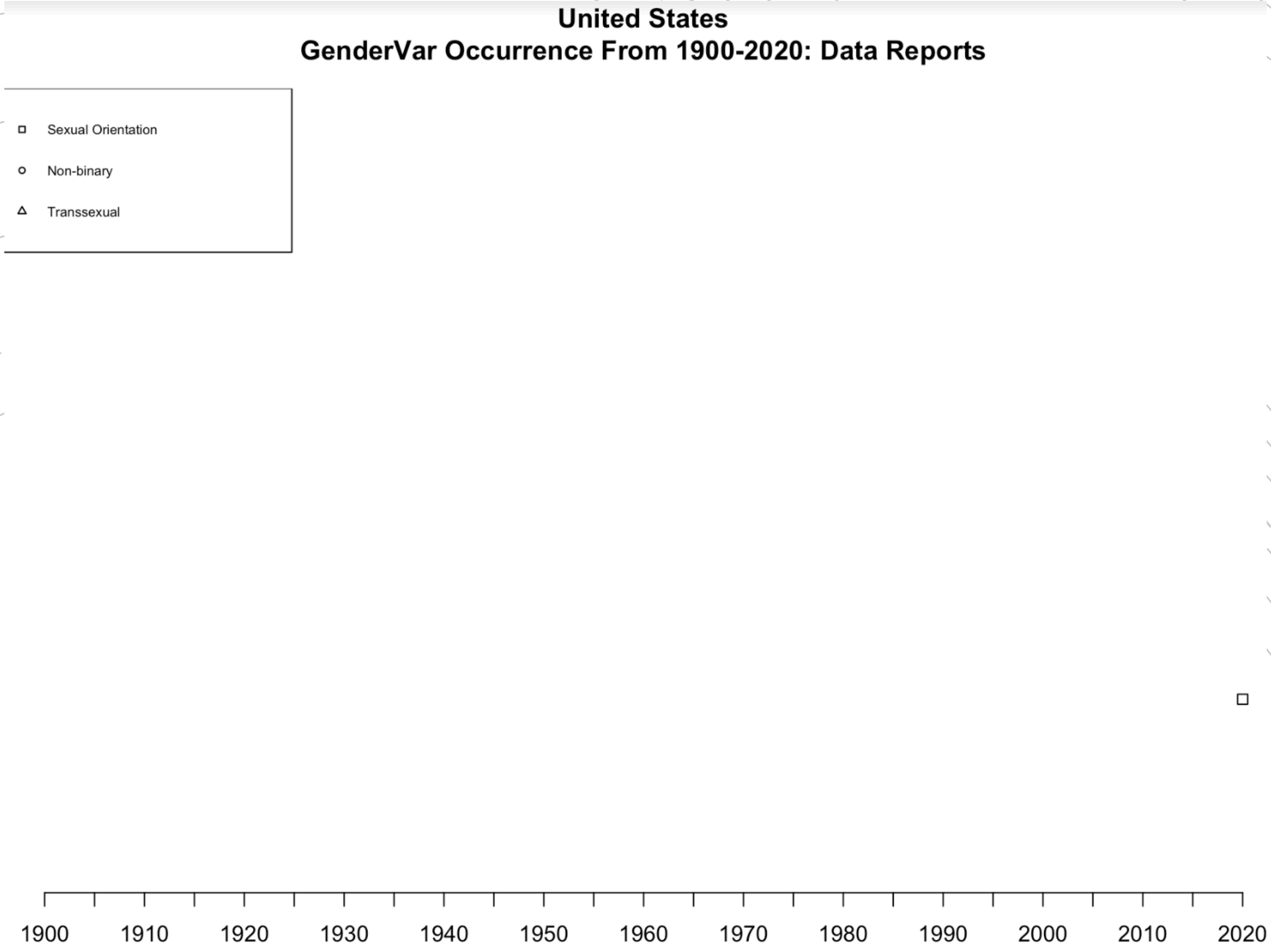
- Changes in variables coded
- Changes in language used to code variables
- Additions or removals of certain variables
- Where variables are coded



Occurrence of Ethnicity Variables in US Health Data Reports



Occurrence of Socioeconomic Variables in US Health Data Reports



Occurrence of Gender Variables in US Health Data Reports

Next Steps

- Continue coding the case studies for the United States, United Kingdom, Sweden, and France
- Research data linkage patterns
- Unearthing the motivations of various actors in the data collection process

Lessons Learned

- Expertise in qualitative coding
- Importance of communication and transparency
- Familiarity with health data collection

Acknowledgements

- Measuring Mortality Team
  - Dr. Julia Lynch, Diya Amlani, Elin Berlin, Gabriella Rabito, Ramsey Radwan, Michael Tu
- Joanne Levy
- LDI Staff
- SUMR Cohort

Questions?