Using Social Media for a Probability Sample: Is it Possible

Marcus Berzofsky

Co-Authors: Tasseli McKay, RTI
                   Patrick Hsieh, RTI
                   Amanda Smith, RTI
                   Natasha Latzman, RTI
                   Katie Grimes, RTI

www.rti.org

- Suicide is 2nd leading cause of death for 10-24 year olds.

- Sexual and gender minority (SGM) youth are 4 times more likely to attempt suicide, and half of transgender youth report thinking seriously about suicide.

- Unique risk and protective factors (e.g., bias-related victimization) likely shape suicidality among SGM youth

- *Current research methods are flawed; only biased data are available to funders and intervention developers.*

No evidence-based suicide prevention programs exist to meet the needs of SGM youth—and they can't be designed from the evidence we have.

Methodological Motivation

- Issue
  - Traditional surveys of children – such as the YRBS - use school-based probability sampling designs
  - While studies with these designs can produce reliable estimates for children, they do not usually provide enough sample in some important subdomains to allow detailed domain analyses

Social media with a Twist!

1. Develop a frame of the target population of interest on a social media platform

2. Use publicly available information on frame members to stratify them based on their likelihood of being in the subpopulation of interest

3. Apply post-survey adjustments to correct for differences in the frame population and the target population

- Issue:
  - Can a frame of users from a social media platform be created?
  - Can it be considered a random subset of the full set of platform users?

- Solution:
  - Twitter has an application programming interface (API) which allows researchers to access publicly available data from all Twitter
  - A random sample of users in the API can be drawn

- Issue:
  - What information is available to determine stratification?
  - What are the criteria which should be used for stratification?

- Solution
  - The API allows one to pull public tweets from frame members
  - An algorithm can be developed to determine likelihood person is in the subdomain of interest
 - Based on assigned likelihood strata can be formed

Step 3: Post-Survey Adjustments

- Issue:
  - No social media platform fully covers the population of interest
  - Users of a particular social media platform may be different than those who do not use it

- Solution:
  - Embed items from nationally representative probability-based studies which are correlated with the outcome of interest
  - Use items in coverage adjustment along with demographic information

- Outcome: Suicide ideation and attempt
- Target population: youth age 14 – 21 in the United States
- Subpopulation of interest: LGBTQ persons
- Social media platform: Twitter

- Randomly selected a set of Twitter IDs from the API
- Needed to select extremely large set of IDs
  - Restricted based on time zone
  - Will later need to restrict on age
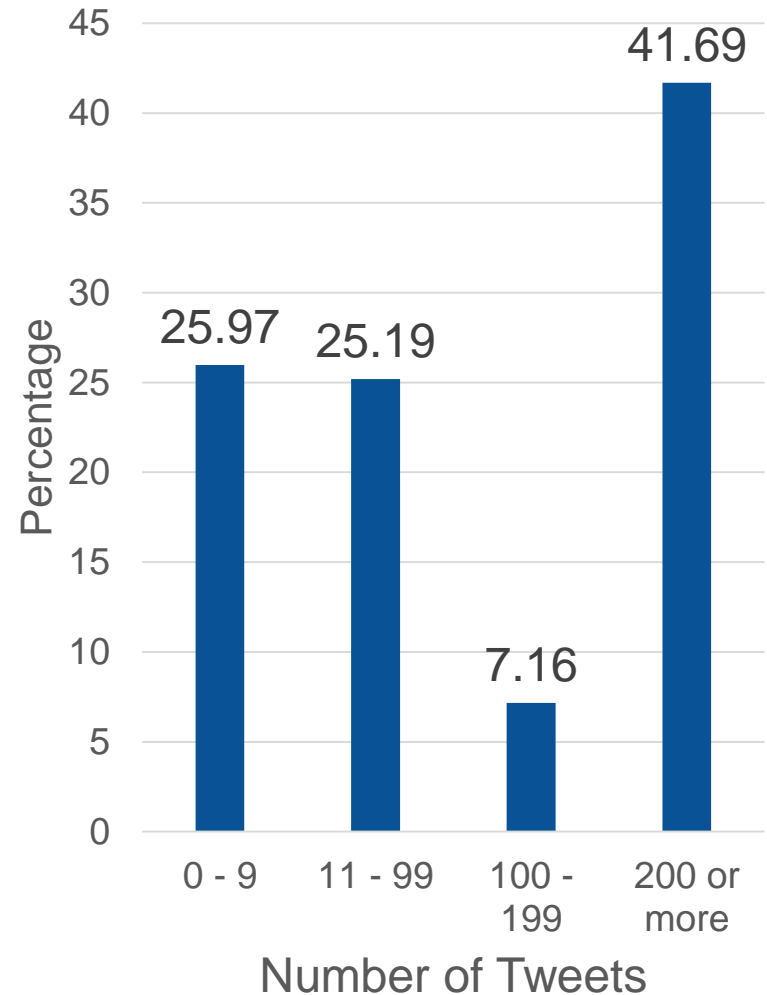
Step 2: Develop Stratification

- Held focus groups with LGBTQ youth
- Developed alpha version of stratification algorithm based on keywords focus groups identified as associated with LGBTQ persons
- Example terms
  - #NYpride
  - #queeryouth
- Based on keyword usage among frame, created 3 strata
  - Low: 0 or 1 keywords
  - Medium: 2 keywords
  - High: 3 or more keywords

Number of Keywords in Tweets
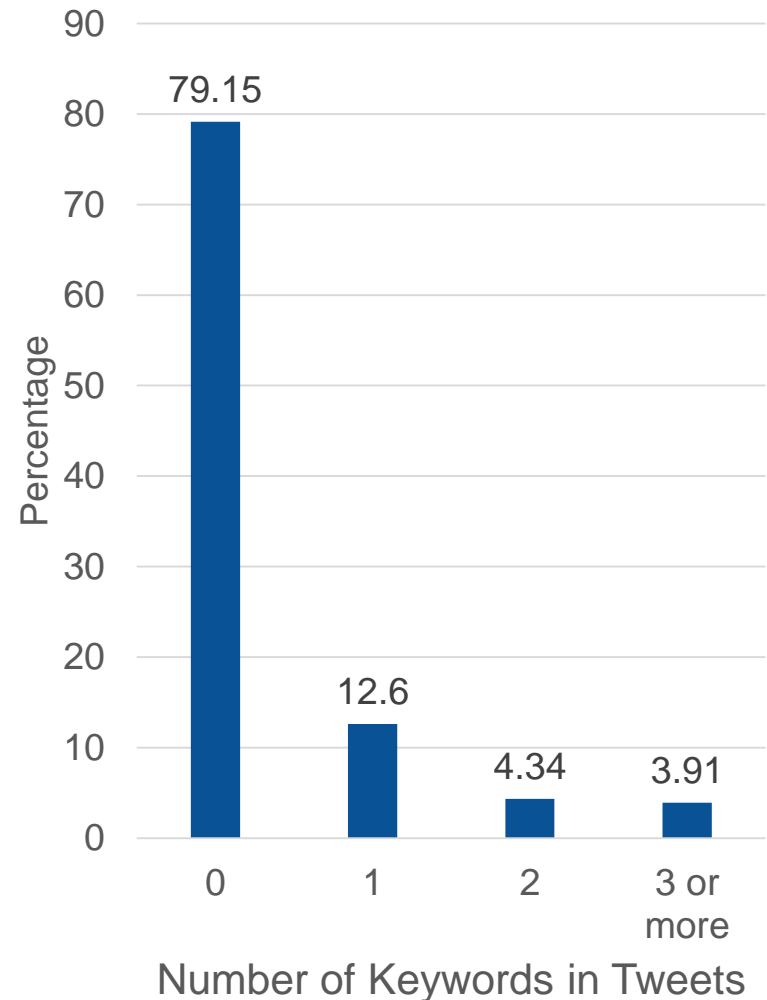
Step 3: Post-Survey Adjustments

- **Included two questions from YRBS related**
  - Youth's belief about how their parents feel about them
  - Youth's feeling about closeness to people at school

Family Connectedness (YRBS)

13. How much do you feel that your parents care about you?
    - A. Not at all
    - B. Very little
    - C. Somewhat
    - D. Quite a bit
    - E. Very much
    - F. Does not apply

School Engagement (YRBS)

14. You feel close to people at your school.
    - A. Strongly disagree
    - B. Somewhat disagree
    - C. Neither agree nor disagree
    - D. Somewhat agree
   - E. Strongly agree

- Pros
  - Easy to load sample in different campaigns to manage sample release
  - Can use Twitter to subset to age range and country of interest
 - Can use Twitter analytics to help understand sample respondents
- Cons
  - Twitter "verifies" list of users which results in large reduction of sample available to receive advertisement; reduction was as high as 90%
  - Cannot manage the number of times a sampled person sees the ad

Initial set of respondents have skewed towards the older end of age range; predominantly been White Non-Hispanic
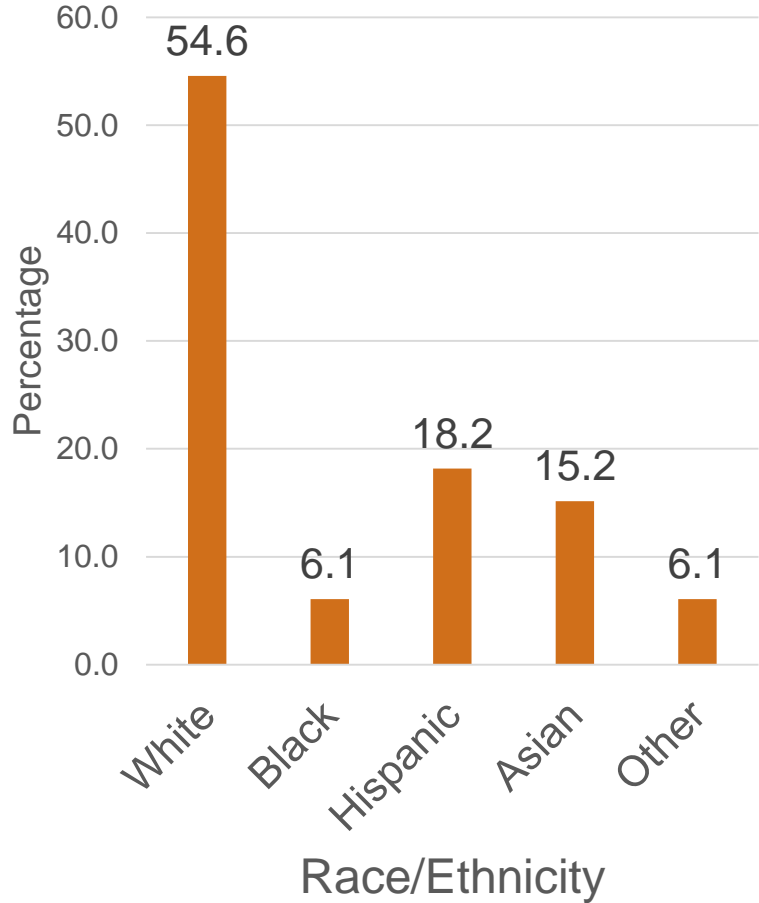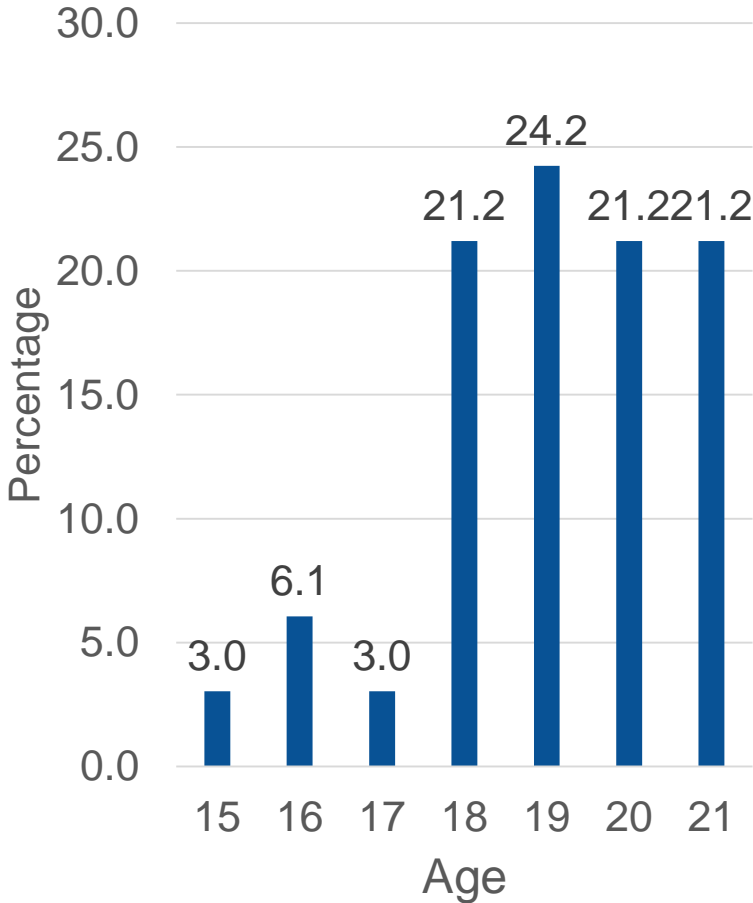
Based on preliminary results, stratification approach does seem to identify LGBTQ persons based on self-identified information
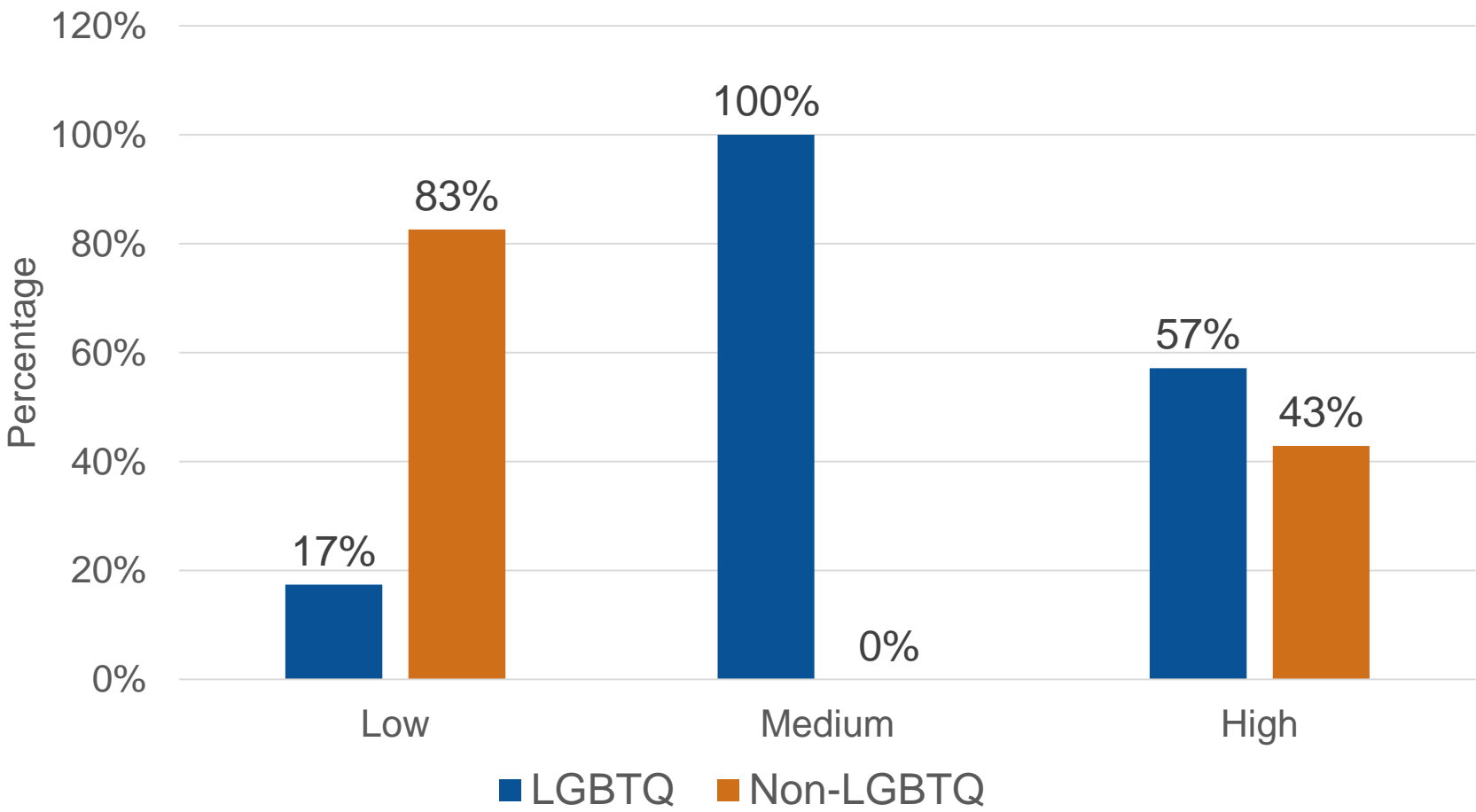
Current and Future Activities

Current

- Data collection for probability-based approach still underway
- Comparison non-probability study underway

Future

- Use API to obtain tweets from respondents to refine stratification algorithm
- Conduct post-survey adjustments and compare survey items not used in post-survey adjustments to comparable national estimates

More Information

Marcus Berzofsky

Senior Research Statistician

919.316.3752

berzofsky@rti.org