**Aim:** Estimate the prevalence of sexual activity by grade, sex and county in the U.S.

**Method:**

*Data*

We used the American Community Survey (ACS), the YRBS and NSFG for all analyses. The samples were restricted to respondents aged 15-18 in all surveys. For NSFG, we used the survey cycles: 2011-2013, 2013-2015, 2015-2017 and 2017-2019. This survey provided information at the national level only. For ACS, we used the 5-year survey available for 2019 (2015-2019). Finally, we used YRBS for state level information for 36 states which had both posed question on sexual status and had an available survey for 2019.

*Variables*

Our primary outcome variable was whether a respondent had ever had sex, stratified by age, and sex. This variable was binary (‘yes’ or ‘no’). The predictor variables included those that were thought to influence sexual status, and overlapped between NSFG and ACS to enable our analysis. These included: race (other, black or white), ethnicity (Hispanic or non-hispanic), total household income category (Under $25,000, $25,000-$59,000, or $60,000 plus), number of household members (1-4 or 5 plus), metropolitan status (census defined MSA, not MSA or other MSA) and mother’s age (categorized as 24 or younger or 25 or older).

*Analysis*

Stage 1: County-level prediction

The overall aim was to produce county-level estimates of the prevalence of sexual activity by age and sex for the U.S. Currently, no survey at the county-level provides this information. Thus, in the first part of this analysis, we developed a predictive model of sexual behaviour at the national level that could ultimately be applied to counties. To do so, we used a logistic model with the covariates listed under ‘variables’ (Equation 1). The outcome was a 0/1 variable for whether a respondent had ever had sex, with separate models for survey years 2011-2015, and 2015-2019 and males versus females. This division resulted in a total of four logistic models of the form in Equation 1. The analyses were weighted using the person-sampling weights provided by NSFG, which are designed to reconstruct the full census sample.

**Equation 1**

Where Y is the logit of whether a respondent had ever had sex is the ith respondent, is the kth predictor variable, X is the predictor as listed under *Variables* (each of which had an estimated coefficient, , and is the error term.

The logistic model yielded the predicted probabilities of whether a respondent had sex based on the set of covariates *k.* Thus, we applied these probabilities to respondents with the same characteristics at the county level enabling a full set of estimates at the county level (Equation 2).

**Equation 2**

= ever had sex (1/0) in respondent *i* (imputed) in county *c.*

Stage 2: State-level adjustment

To increase robustness of the county-level estimates, we adjusted them to estimates from the YRBS at the state level. Only 36 states had an available survey with the sex outcome question in 2019. For states that had asked the question in previous years, we extrapolated the trend to determine the 2019 value. For states that had never asked the question or did not have a YRBS survey, we employed a prediction model as in stage 1 to similarly estimate the state-level prevalence of sexual activity.

**Further directions**

1. **Variable issues**
   1. Ensure alignment of household income variable between ACS and NSFG
      1. Double check definitions
      2. There was also an issue were ACS income as per the data was much higher than reported online on ACS website – still unclear
   2. Family linkages for ACS
      1. Double check definition of ‘mother’ aligns with NSFG, some mothers were younger than the children
      2. Use family linkage files to determine household composition, instead of just number of people living in the household
      3. All files are available here
   3. Combine race and ethnicity into 1 variable
2. **Analytic issues**
   1. Check difference in characteristics between those who report being **currently** in school and those who are not (ACS or NSFG)
      1. This will determine whether exclusion of non-high school students in YRBS could bias results
   2. Determine alternate predictive models using machine learning, e.g. random forest
      1. Currently using logistic model
   3. Stratification
      1. Need to stratify models by age (could combine across years to improve power, else sample sizes become very small)
   4. Time
      1. Years/secular trend are not dealt with at all in the analysis currently