Predicting US state teenage birth rates using search engine query data on pregnancy termination and prevention

| Daniel D Reidpath[1], Pascale Allotey[2]

Background To determine whether aggregated searches for pregnancy prevention or pregnancy termination predicts US State teenage birth rates.

Methods US birth rate data for the 50 states, and search engine query data (Google Trends) for "condom" and "abortion" were used in an ecological analysis. Multivariable ordinary least squares regression was used to predict state-level birth rates from state-level searches for condom and abortion.

Results The final model accounted for 35% of the variance (R^2=0.347). Abortion and condom had similar, absolute, standardized parameters (ß≈0.5). High state-levels of searches for abortion were associated with higher teenage birth rates, whereas high state-levels of searches for condom were associated with lower teenage birth rates.

Conclusions Google Trends data for abortion and condom can be used to model US state-levels teenage birth rates. This raises the possibility of well targeted, accessible and relevant information for populations wanting to avoid unwanted pregnancies.

[1] Global Public Health, School of Medicine and Health Sciences, Monash University Malaysia, Bandar Sunway, Malaysia
[2] International Institute for Global Health, United Nations University, Cheras, Malaysia

Search engine query data (SEQD) include the information collected by search engine companies about search terms used, geographical location, and the date of the search. With an increasing need for data to monitor emerging global health risks, as well as evaluate the effectiveness of various interventions, SEQD provide important proxy, descriptive data. SEQD in health research are just beginning to be explored, most famously to predict influenza outbreaks (*1*, *2*). Most subsequent research has similarly focused on communicable diseases, although there is some limited use in other areas of public health (*3*).

Teenage births (15-19 years old) are generally considered to carry significant health and social risks for mother and child; although the risks are often outcomes of other social causes (*4*). Teenage births are associated with higher obstetric risk (*5*) related to delayed antenatal care (*6*), poorer long term health outcomes for the mother (*7*), poorer long term health outcomes for the child (*8*), as

CORRESPONDENCE TO:

Prof Daniel D Reidpath | Global Public Health | School of Medicine and Health Sciences
Monash University Malaysia | Jalan Lagoon Selatan | Bandar Sunway | Selangor DE |
Malaysia 46150 | daniel.reidpath@monash.edu

well as a range of negative social sequelae (*9*). The US teenage birth rate (TBR) has declined steadily since 1970 (68.3) to 2010 (34.2); but it remains high by Western European standards. Declining rates can nonetheless be misleading if one imagines that a single TBR applies to the whole of the US, because the rates are unevenly distributed across the states from a low of 15.7 in New Hampshire to a high in Mississippi of 55.0 (*10*).

Aggregated, internet searches reflect the information that is being sought by a population. In the same way that populations in which flu-like symptoms are occurring may search for information about those symptoms, one can equally well imagine that populations wishing to avoid unwanted pregnancies are more likely to seek information about prevention of pregnancies, and those experiencing unwanted pregnancies would seek information about their options, including the termination of pregnancies. This raises the intriguing possibility that the state-level variation in searches for pregnancy termination and pregnancy prevention may predict the state-level variation in teenage birth rates.

METHODS

Given the broad nature of sexual and reproductive health, and the volume and range of results using more general terms like contraception, sex and sexually transmissible infections (STIs), we restricted the search to as narrow a range as possible that would still reflect the question of interest. In imposing this narrow focus, we acknowledge that instances of other searches that relate to termination of pregnancy and contraception would be lost. Nonetheless, the general approach remains illustrative of the potential of SEQD. For this study, we used the single search words: 'abortion' and 'condom', as markers for pregnancy termination and pregnancy prevention respectively.

State-level TBRs for 2010 were obtained from The US Centers for Disease Control, *National Vital Statistics Report* (*10*). The birth rates were reported per 1000 estimated women in the age range 15–19 in each state (p. 42).

The SEQD were obtained from "Google Trends" (http://www.google.com/trends/). Search data from each state were aggregated over the five year period 2006–2010 to smooth out any year to year variation. The data were returned in a normalized and rescaled form. Google normalizes SEQD so that regional variation in search counts are not simply attributable to population size. The rescaling gives the state with the highest number of normalized searches a value of 100. Every other state's rescaled value is then represented as a percentage of the maximum number of searches. As an example, the Google Trends strategy used for "condom" was: http://www.google.com/trends/explore?q=condom&geo=US&date=1%2F2006%2060m&cmpt=date

Multivariable, ordinary least squares (OLS) regression was used to examine the independent predictive value of the state numbers of searches for "abortion" and "condom" in the estimation of state teenage birth rates.

The research used aggregated, anonymous, publicly available data and was not a study of human subjects. The research neither required nor received Institutional Review Board approval.

RESULTS

A visual inspection of histograms of the TBR, condom and abortion data showed broadly unimodal, Gaussian distributions; summary statistics are shown in Table 1. There were no missing data.

The correlation between the two predictors, abortion and condom, was modest (r=0.33; 95% CI=0.059 to 0.559). In the OLS model the abortion and condom values were regressed against the TBR. The model accounted for around 35% of the variance (R^2=0.347, $F_{(2, 47)}$=2.47, $P<.0001$). The absolute value of the standardized coefficients (ß) for abor-

Table 1. Summary statistics for the outcome variable (TBR) and the predictor variables (abortion and condom)

	MINIMUM	MEDIAN	MEAN	SD	MAXIMUM
TBR	15.7	33.0	34.0	10.1	55
Abortion	55.0	78	78.2	9.5	100
Condom	50.0	74	73.5	11.3	100

SD – standard deviation, TBR – teenage birth rate

tion (0.48; 95% CI=0.230 to 0.720) and condom (-0.54; 95% CI=-0.785 to -0.294) were close – both around 0.5.

DISCUSSION

In the US the internet search behavior of state populations for pregnancy prevention (condom) and pregnancy termination (abortion) information is significantly associated with state-level teenage birth rates. Furthermore, the association with the TBR is balanced, such that seeking pregnancy prevention information is independently associated with a *reduction* in the TBR, and seeking pregnancy termination information is independently associated with an *increase* in the TBR.

Google trends data has not previously been used in quite this way (looking at state-level associations), but the broad finding is consistent with the infectious disease studies of SEQD. The ecological nature of the study precludes links being made between individuals searching the internet for pregnancy prevention or pregnancy termination information, and subsequent individual teenage births. It does, however, suggest the importance of information in the population, and the kinds of information being sought (*11*, *12*). Populations seeking knowledge about pregnancy prevention were less likely to experience teenage pregnancy and therefore teenage births, and populations seeking information about pregnancy termination were more likely to experience teenage pregnancy and therefore teenage births. While the application of knowledge about pregnancy termination is likely to lead to the termination of individual, unwanted pregnancies, searches for information about termination are likely to simply indicate the presence of an unwanted pregnancy.

Most SEQD studies have been used to explore temporal dimensions of the data, and if weekly birth rate data were available, it would certainly be possible to extend the research in that direction. The possibility of time series is particularly interesting, because if the birth rate time series were associated with searches for pregnancy prevention and pregnancy termination, one would expect to see appropriate temporal ordering of prevention searches occurring prior to termination searches.

There are limitations to this study, and future lines of inquiry to which the results point. The ecological nature of the data, limits the kinds of conclusions that may be drawn; and it would be dangerous to infer anything about individual behavior from the aggregate state search data. There is also the possibility of confounding. Notwithstanding these limitations, the results raise the potential of SEQD data as an extra tool to inform public health predictive and preventive approaches beyond teenage pregnancy.

CONCLUSIONS

Search engine query data may provide novel methods for exploring adolescent health issues. In this case, high search rates for prevention and low searches for termination were associated with the lowest state level birth rates. Types of information, access to information, and information utilization may remain key areas for targeted prevention.

REFERENCES

1 Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4. PubMed https://doi.org/10.1038/nature07634

2 Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One. 2011;6:e23610. PubMed https://doi.org/10.1371/journal.pone.0023610

3 Huang J, Zheng R, Emery S. Assessing the impact of the national smoking ban in indoor public places in China: evidence from quit smoking related online searches. PLoS One. 2013;8:e65577. PubMed https://doi.org/10.1371/journal.pone.0065577

4 Lawlor DA, Shaw M. Too much too young? Teenage pregnancy is not a public health problem. Int J Epidemiol. 2002;31:552–4. PubMed https://doi.org/10.1093/ije/31.3.552

5 Leppälahti S, Gissler M, Mentula M, Heikinheimo O. Is teenage pregnancy an obstetric risk in a welfare society? A population-based study in Finland, from 2006 to 2011. BMJ Open. 2013;3:e003225. PubMed https://doi.org/10.1136/bmjopen-2013-003225

6 Fleming N, Ng N, Osborne C, Biederman S, Yasseen AS 3rd, Dy J, et al. Adolescent pregnancy outcomes in the province of Ontario: a cohort study. J Obstet Gynaecol Can. 2013;35:234–45. PubMed https://doi.org/10.1016/S1701-2163(15)30995-6

7 Chang T, Choi H, Richardson CR, Davis MM. Implications of teen birth for overweight and obesity in adulthood. Am J Obstet Gynecol. 2013;209:110.e1–7. PubMed https://doi.org/10.1016/j.ajog.2013.04.023

8 Coyne CA, Långström N, Lichtenstein P, D'Onofrio BM. The association between teenage motherhood and poor offspring outcomes: a national cohort study across 30 years. Twin Res Hum Genet. 2013;16:679–89. PubMed https://doi.org/10.1017/thg.2013.23

9 Francesconi M. Adult Outcomes for Children of Teenage Mothers. Bonn: Institute for the Study of Labor (IZA); 2007.

10 Martin JA, Hamilton BE, Ventura SJ, Osterman MJK, Wilson EC, Mathews TJ. Births: Final data for 2010. Natl Vital Stat Rep. 2012;61:1–72. PubMed

11 Aguirre PE, Coelho M, Oliveira T, Rios D, Cruvinel AF, Cruvinel T. What can Google inform us about people's interests regarding dental caries in different populations? Caries Res. 2018;52:177–88. PubMed https://doi.org/10.1159/000485107

12 Phillips CA, Barz Leahy A, Li Y, Schapira MM, Bailey LC, Merchant RM. Relationship between state-level Google online search volume and cancer incidence in the United States: Retrospective study. J Med Internet Res. 2018;20:e6. PubMed https://doi.org/10.2196/jmir.8870