Universal Mechanism to Link Multiple Datasets with Synthetic Populations for Clinical and Epidemiological Research

Georgiy Bobashev (bobashev@rti.org)
RTI International

Emily Hadley
RTI International

Alan Karr
AFK Analytics, LLC

Caroline Kery
RTI International

James Rineer
RTI International

Research Article

Keywords: Synthetic Populations, Clinical, FYSAS

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-456335/v1

ABSTRACT:

Many modern predictive models require knowledge acquired from multiple datasets, yet data linkage can be a substantial challenge. Datasets exist in a wide variety of formats and linking them directly is often infeasible or restricted by privacy requirements. One solution is to map variables from different datasets onto a synthetic population, producing a dataset that contains information from multiple sources sufficient for reliable statistical inference with quantifiable uncertainty. This approach is universal because it is applicable to a broad range of datasets and has good potential for privacy protection.

We consider datasets with information about individuals and describe three methods for building linked synthetic data: resampling, modeling predictors independently, and modeling predictors sequentially. We apply these methods to the prediction of the prevalence of Florida youth vaping by state, county, and census tract using the 2018 Florida Youth Substance Abuse Survey (FYSAS) and the 5-Year American Community Survey (ACS). We find that resampling and sequential modeling most closely approximate the 2018 survey results, and that the sequential model captures more variability. We find that the order of predicting the variables in sequential modeling does not substantially impact the outcome.

Universal Mechanism to Link Multiple Datasets with Synthetic Populations for Clinical and Epidemiological Research

Bobashev, Georgiy*[1]

Hadley, Emily[1]

Karr, Alan[2]

Kery, Caroline[1]

Rineer, James[1]

[1] RTI International

[2] AFK Analytics, LLC

*Send all correspondence to:
Georgiy Bobashev
RTI International
3040 E. Cornwallis Road
P.O. Box 12194
Research Triangle Park, NC 27709
Voice: (919) 541-6167
Fax: (919) 541-6722
E-mail: bobashev@rti.org

1. INTRODUCTION

Many modern research questions require knowledge acquired from multiple datasets. This need arises when obtaining a single dataset with all the required information is either difficult or impossible, or when the data already exists in multiple data sources. For modeling disease progression in a population, there is often a need to link population (survey) data with longitudinal clinical data because clinical data contains information about disease progression and the survey dataset contains data on the population. Examples include predicting the effect of new cancer screening on the US population, identifying obesity hotspots (i.e., geographic areas with unusually high prevalence of high BMI individuals), and predicting the effect of interventions aimed to reduce opioid-related deaths.[1-6] In each case, predictive models need to use data on population demographics and geographic locations, along with the natural history of the disease, administrative data, etc. We can thus formulate the main question as follows:

Imagine that we want to jointly analyze variables $X_1,\dots,X_m$. However, in one dataset (Dataset A) we have part of the variables $X_1,\dots,X_{m1}$ and in another dataset (Dataset B) we have other variables $X_{m1+1},\dots,X_m$. How can we develop a dataset C which would have the full set of variables $X_1,\dots,X_m$?

Producing knowledge from multiple datasets poses a two-fold challenge. (1) How do we link two or more datasets together to enhance information contained in each of them? (2) How do we protect privacy of the subjects when cross-linking multiple datasets can potentially identify unique and identifiable records? In this paper we focus on the first challenge of linking multiple datasets.

In rare cases (e.g., clinical studies and medical records) where it is known that the same subjects were present in two or more datasets, it is possible to link records directly by matching unique IDs either directly or probabilistically.[7] In this paper, however, we are not trying to reconstruct real individual profiles, but rather we aim to create a dataset that contains information from multiple datasets sufficient to produce reliable statistical inference with quantifiable uncertainty, while maintaining the privacy of individuals from whom the data was collected.

In this paper, we show that if both datasets have a set of common variables, it is possible (with varying but quantifiable uncertainty) to create a dataset where all variables of interest from the datasets are present. We summarize the methods to answer the main question, illustrate how this methodology is used in practice, and lay ground for future directions of research in this area. While the technical details of the methodology are subjects for further improvements and adaptations, the overall approach is universal because it is not limited to specific types of data and can thus be applicable to a broad range of data.

2. METHODOLOGY

We have found it useful to link auxiliary dataset to nationally representative datasets which can be subset to represent any specific subpopulation. We therefore introduce a synthetic population dataset which statistically represents a full, spatially explicit, enumeration of all households and household persons for the entire United States based on methods advanced by RTI International.[8] This dataset provides person-level demographic data upon which subpopulations can be selected. In contrast to linking directly to PHI or PII records, with this approach there are no inherent privacy limitations because all records are synthesized from already anonymized census information. As a baseline dataset, then, the synthetic population can safely serve as the basis for linking multiple datasets.

Synthetic Populations

Synthetic populations are representations of every household and person in a population. For example, a US synthetic person population dataset contains over 300 million rows representing each person in the US population. Column variables contain variables such as geographic coordinates, household size, age, race, gender, and education for each individual. This US dataset can be produced from the aggregate public use American Community survey data. In the aggregate, block group level, demographics of the synthetic population point data will match the same demographics in the ACS for the household variables used to build the dataset. The dataset is created leveraging iterative proportional fitting algorithms and resampling from the public use microdata for the current year being synthesized as described in.[8] Because the data are entirely computer-generated, matching of synthetic characteristics to real individual characteristics are purely coincidental.

RTI has created nationwide synthetic populations for the US as well as for select international settings. Synthetic populations have been used by governments and academia as the basis for population simulation modeling (microsimulations and ABMS), including in support of recent COVID-19 modeling and analysis work.[9-11] Of relevance here, synthetic households and persons have education and health status or other characteristics and behaviors assigned to them based on covariates from surveys or other sources. In addition, a geographic location at multiple times during the day is available.

Linking Multiple Datasets Approaches

Notation:

Dataset A. A recipient dataset

Dataset B. A donor dataset

Dataset C. The resulting dataset that is based on datasets A and B. Usually it is Dataset A with added variables based on Dataset B.

{X} a set of variables of interest.

{V} a set of variables that are common in Datasets A and B.

V^{A}_{ij}, V^{B}_{ij}, X^{A}_{ij}, X^{B}_{ij} correspond to an observation *i* for variable *j* in dataset A, B.

Because synthetic populations are representative of the US population, they can be used as the basis to which other datasets are linked. Specifically, we examine the following scenario:

Consider a dataset (e.g., a synthetic population) with demographic variables but lacking columns of interest. We call it Dataset A. Also consider another dataset (e.g., a survey dataset) that contains variables of interest and some columns in common with Dataset A. We will call this Dataset B and refer to the common variables as "linkage variables" and denote them as {V}. The number of rows could be different between datasets A and B and dataset B could also contain sampling weights. Our goal is to create a dataset C that will contain as many subjects as in Dataset A but also all variables of interest from Dataset B. If dataset A is a synthetic population dataset, it is natural to use it a "recipient" dataset, i.e., dataset C will be dataset A with augmented variables of interest being obtained from a "donor" dataset B.

Inherent to multiple dataset analysis are the issues related to data preprocessing to harmonize coding, quality, and compatibility. For example, imputation of deletion of observations with missing values, different coding of similar variables (ages vs. age groups, race ethnicity as two different variables vs. race-ethnicity combined, etc.). Thus, before the datasets are linked, they need to be harmonized, i.e., contain measures with the same coding and units for common variables. These procedures are broadly discussed elsewhere[12] and are not the subject of the current study.

After assuring that the datasets contain the same variables, we can now start linking the datasets. Dataset C starts as Dataset A with available values of $\{X^{C}_{ij=1,..,m1}\}$. Each subject *i* in this dataset will then be assigned more values $\{X^{C}_{ij=m1+1,..,m}\}$. We will consider two approaches: resampling and modeling. The resampling approach considers assigning real data from the donor dataset B to observations in the recipient dataset A; the modeling approach first develops generative models of the observations based on the donor dataset and then the generative model assigns augmented values to the rows in the recipient dataset. Both approaches are variants of commonly known approaches to dealing with missing data.[13]

Approach 1. Profile Matching (Resampling)

This approach is sometimes called "Hot-deck" assignment. Essentially, for each profile of common variables $V^{A}_{i.}$ in Dataset A, it finds a multitude of K matching profiles $V^{B}_{K.}$ in Dataset B and randomly picks up a single profile k<{K}. $X^{B}_{k.}$ is assigned to subject *i* in Dataset C, so that $X^{C}_{i.}= X^{B}_{k.}$. Matching could be exact, if possible, or based on some measure of similarity $D=\| V^{A}_{i.}– V^{B}_{.j}\|$ a subject k with the minimal value of D will donate the values of $X^{C}_{i.}= X^{B}_{k.}$. If minimal values are achieved on multiple profiles, then the candidate profile is randomly selected from the candidates. When dataset B is a probability-based sampled survey, appropriate sampling weights should be used to determine correct sampling probabilities. This approach is commonly used in missing value imputation where distance D could take complex forms, sometimes with the assistance of machine learning techniques.

The advantage of this approach is that it considers the entire profile of real observations which represents a sample from the unknown multivariate distribution and thus preserves the correlation structure. Another advantage is that there might be a certain interdependence between the variable that goes beyond simple correlation. For example, individuals with higher household income might be less likely to take a student loan. At the same time, the major limitation is that the number of unique profiles might be limited and may not cover the entire multivariate distribution. Thus, this approach might underestimate the variability of profiles.

Approach 2. Modeling

Under this approach, the values of $X^{C}{}_{im1,.,m}$ will be modeled from dataset B using a set of common variables $V^{B}{}_{.j}$ and a predictive model $X^{C}{}_{ij} = F(V^{B}{}_{.j}, \varepsilon)$, where ε is an error term that follows a known estimated distribution. In a simple case, where $X^{C}{}_{ij} = F(V^{B}{}_{.j}) + \varepsilon$, the prediction of the actual value implies an estimation of the mean, and drawing a random value from the known distribution. In a more general case, the entire function F becomes a complex distribution with parameters depending on the common variables and estimated parameters. Identification of a right model can be ambiguous as well as the identification of the right distributions. However, as described later in this section, model validation on independent datasets provides a mechanism to choose better performing models. If dataset B is a probability-based sampling survey, the use of sampling weight in fitting the model F is not as clear as in case of profile resampling. A review of diagnostic tests in weighted and unweighted regression analysis could be found in K. Bollen et al. (2016).[14]

The modeling approach has the advantage of being flexible in terms of model selection, and it provides a broader variation because of the random draws from the distributions. The disadvantage is that this approach generally ignores estimation of joint distribution between variables $X^{C}{}_{.j}$. This shortcoming can be partially alleviated by the use of independent, sequential, or joint estimation approaches as described in the next section.

Implementation Methods

Each approach can be implemented independently, sequentially, or jointly/partially sequentially.

Independence method implies that each variable is independently added using only the original set of common variables. Independence here is implied in a conditional sense, i.e., it is assumed that variables are conditionally independent given the set of common variables. Assume that there are p common variables such that j=1,..,p. In this approach, each newly created variable $X^{C}_{.k}$ will be a function of the same set of variables $V^{B}_{.j}$ where j=1,…p. In the resampling approach, this implies that the profiles to find similar observations in dataset B will be based only on the set of pre-defined common variables. For the Modeling approach, this will mean that $X^{C}_{.k} = F_k(V^{B}_{.j}, \varepsilon)$, where functions F_k could be different, but the set of predictor variables $V^{B}_{.j}$ remains the same. The method is called "independence" because the order of which variables are linked to a dataset is irrelevant. If the added variables are truly conditionally independent given the shared variables, this approach is exact.

Sequential method implies that the variables are linked in sequence and the next imputed variable considers the growing set of common variables. After the first variable X^{C}_{im1} has been created in dataset C, this dataset could be considered as a new dataset A, and the set of common variables between new dataset A and dataset B is now $\{V_{.j}, X_{.m1}\}$. This new dataset could be now used to determine either new similar profiles as in resampling approach or a new predictive model as in the modeling approach. As more variables are added to dataset C, the set of common variables grows. As in EM methodology for imputation, this process can continue for several rounds until convergence in some sense will be achieved, i.e., after the last variable $X_{.m}$ is added to dataset C, the first added variable $X_{.m1}$ could be recalculated using the set of common variables $\{V_{.j}, X_{.m1+1,\dots,m}\}$, etc.

Sequential methods combined with the modeling approach can lead to hierarchical and potentially convoluted distributions with accumulating error.

$$X^{C}_{.k+1} = F_{k+1}(V^{B}_{.j}, X^{C}_{.k}, \varepsilon)= F_{k}(V^{B}_{.j}, F_{k}(V^{B}_{.j}, \varepsilon_{1}), \varepsilon_{2})$$

In sequential linkage of multiple variables, there are some additional questions. Which variable to start with? Which one will be the second, and so on? A simple solution is to identify which variable of the addition list can be best predicted with the existing list of variables. As the list of included variables grows, the accuracy of the models increases. However, the predictive accuracy is usually estimated through some validation mechanism, e.g., 10-fold cross-validation. Conducting this procedure for large datasets such as national Household Survey of Drug use and Health (NSDUH) with several hundred of variables and 55,000 subjects can become computationally intensive.

Joint/partially sequential approach

In this approach, it is possible to model simultaneously two or more variables and add them to dataset C. This method can combine resampling and modeling approaches where some variables are generated using the modeling approach, and the others are taken as profiles with a resampling approach. Additionally, it is possible to model multivariate profiles with a joint model (e.g., using copulas[15]) but this method could quickly become too cumbersome. An additional issue with this approach is that uncertainty in the resulting variables could be difficult to assess.

Accuracy of Predictive Models and Uncertainty

Predictive models need to be validated for accuracy and uncertainty. Measures of accuracy for predictive models have been well-established and are based on the comparison between predicted and true outcomes. These measures could be in the form of a confusion matrix, F1, etc. measures for categorical variables, and pseudo-R^2 for continuous ones. The dataset is split into training and test sets, and the accuracy is measured on the test set. Because there are no “true” values for the combined data set, one needs to find other data sources that contain the same variables and apply the same procedures on that dataset, or simulate a dataset with known structure and test whether the linking approach successfully reproduces the simulated data.

Validation is an important step, obviously for checking the accuracy of the predicted values, but also for forming a thought process about the levels of uncertainty and what is “good enough” for practical purposes. The most traditional way to validate the model is to predict a known outcome and assess the quality of prediction. This assessment could be done at the individual level (e.g., for each person assess the accuracy of prediction), or at the population level, (i.e., comparing the difference in a specific statistic). The former is, of course, a stronger validation approach than the latter. In our study, we considered both: individual-level validation by generating and evaluating a confusion matrix and group-level validation, to assure that main marginal statistics to be satisfied.

Uncertainty in predictive models comes from many sources. Some uncertainty is inherent in the datasets that are being linked and are routinely considered in data analysis. For example, survey data contains uncertainty produced by the survey process *per se*, including sample selection procedures, sample weights, cognitive biases, non-response, and so forth. Another source is the predictive modeling process. Models based on the data are not perfect and result in variability of their own associated with the residual error and the standard errors in the parameters, even if the underlying data is perfect. In sequential modeling process these errors are accumulating because the next variable is conditioned on the predicted value of the previous variables.

Assessing the partial and total uncertainty could be done with multiple replications of the process, where at each step the values are drawn from the corresponding distribution. This assessment also goes back to the use of multiple synthetic populations because each synthetic population is just a realization of what the population could have looked like given the marginal distributions. Finally, there is "deep" uncertainty that contains factors that are unknown to researchers and that are unfeasible to assess. These factors could include natural and man-made disasters, abrupt changes in politics, economics, and such. In many applications, such deep assessment is not practical and uncertainty estimation is limited to multiple replications of conditional models. While it could be sufficient for practical purposes, it is important to be explicit about the level of uncertainty considered.

3. APPLICATIONS AND AN EXAMPLE TO 2018 FYSAS DATASET ON YOUTH VAPING IN FLORIDA

We applied synthetic dataset linking method to several projects in the past, including linking NHIS data that contains data on cancer screening and demographic with SEER containing incidence and prevalence,[5,6] BMI mapping and cluster detection (https://synthpopviewer.rti.org/obesity). In this section, we illustrate the details of its work on a recent example predicting vaping of youth in Florida counties. We specifically explore the resampling approach and compare it to the independent and sequential methods of the modeling approach.

3.1 Estimating Electronic Vaping Usage and Relevant Predictors Among Youth in Florida

The increased use of electronic vaping products is of considerable interest to public health researchers. In this example, we sought to link a synthetic population of Florida youth aged 10 to 19 with the weighted 2018 Florida Youth Substance Abuse Survey (FYSAS) estimates of electronic vaping and other associated variables.

The synthetic population used in this scenario was built using the 2017 American Community Survey (ACS) three-way tables with age, race, and gender by census tract. The population was limited to youth ages 9 to 18. To create a 2018 synthetic population (the 2018 ACS data was not available at the time of development), the 2017 synthetic population was aged up by one year. An adjustment for growth was made using county-level population growth estimates from the Florida Division of Public Health Statistics and Performance Management (see http://www.flhealthcharts.com/). The 2018 synthetic population was validated through comparison with the 2018 statewide demographic estimates from the Florida Department of Health.

The outcome variable of interest is the proportion of youth who vaped in the last 30 days. Half of the FYSAS data was used to train the models, and **Table 1** summarizes the mean and standard deviation of this outcome at the state level for each method averaged over 100 runs. Each method is described in detail below. **Table 2** highlights the confusion matrix results for the three methods that were evaluated as a part of this analysis.

Table 1. Average proportion of respondents that vaped in last 30 days by method

	2018 FYSAS Survey (actual)	**Resampling**	**Modeling: Independent**	**Modeling: Dependent**
Vaped in Last 30 Days	0.139[1]	0.131 (0.00017)	0.099 (0.00019)	0.136 (0.00022)

[1] This result is different from the 2018 FYSAS statewide value (0.137) because the dataset was split in half for the purposes of model building. The models are compared with the result from the half of the dataset that they were tested on.

Table 2. Confusion Matrix Results

	True Positive	False Positive	True Negative	False Negative
Resampling	23,630 (3.78%)	64,233 (10.27%)	474,427 (75.83%)	63,384 (10.13%)
Modeling: Independent	12,530 (2.00%)	50,638 (8.09%)	488,022 (77.90%)	74,484 (11.90%)
Modeling: Dependent	16,891 (2.70%)	70,997 (11.35%)	467,663 (74.75%)	70,123 (11.2%)

3.1.1 Results from Resampling Method

In the resampling method, we considered the demographic characteristics (age, gender, race) of each member of the synthetic population and sample a record with replacement from the weighted FYSAS data with the same demographic characteristics from the same county. We appended the data for the FYSAS sampled record (including the variable for vaped in last 30 days) to the record for the synthetic population member. An exact match was sampled from the FYSAS records with the same demographic characteristics and the same county for 85% of the synthetic population.

For synthetic population members without a match, we first expanded the age variable to include plus/minus one year and sampled a matching record within the same county. This was used for 9.6% of the synthetic population. For members who still lacked a match, we expanded the sample to the entire state and then sampled a matching record. This method was used for 5.4% of the synthetic population.

Since we sampled with replacements from the weighted FYSAS records, some of the 1,251,348 weighted FYSAS records were sampled multiple times while other records were never sampled. **Table 3** highlights that 29.8% of records were never sampled. The most sampled record was a 10-year old Hispanic female from Desoto County whose profile was sampled 167 times.

Table 3. Number of Times that the Record was Sampled from Weighted Sample

Number of Times Record was Sampled	Count of Records	Percent of Records
0 times	373,218	29.8%
1 time	354,368	28.3%
2 to 10 times	503,405	40.2%
11 to 100 times	20,287	1.6%
More than 100 times	70	< 1%

After completing the linkage between the synthetic population and FYSAS data using the resampling method, we then calculated the proportion of the synthetic population who vaped in the last 30 days as seen in **Table 1**. This proportion of 0.131 is smaller than, but close to, the actual proportion of 0.139. This approach also has the smallest variance of the three considered methods.

3.1.2 Results from Modeling Method

In the modeling method, we modeled the outcome of vaping in the last 30 days using predictors common to both the synthetic population and to FYSAS. Initially, only the demographic characteristics (age, gender, race) and county were common to both datasets. The model built with these characteristics alone was limited in its accuracy. We therefore sought to add variables to the synthetic population to allow for more shared variables between the synthetic

To reduce the number of variables considered for adding to the synthetic population, we completed the following steps:

1. Trained a decision tree variable on the FYSAS demographic data with vaping status as the outcome and create a “leaf” variable to account for interactions in demographic predictors
2. Identified the FYSAS variables most highly correlated with electronic vaping in the last 30 days
3. Used Lasso regression as a feature selector to narrow the list of variables
4. Used K-Modes cluster analysis to assign each respondent a cluster based on the variables selected in Step 2
5. Ran Principal Component Analysis as an additional data reduction technique and select the first two components
6. Used a second Lasso regression with the set of correlation variables from Step 1, the cluster assignments from Step 3, and the principal components from Step 4 to select the final variables

Seven variables were selected to be added to the synthetic population for the modeling process. These seven variables are the row names in **Table 4**. These variables were used as predictors for the outcome variable. The columns compare the results of assigning predictors to synthetic population members with the different modeling methods (described below).

The results in **Table 4** suggest similar proportions of the synthetic population members are assigned to the predictors important to predicting electronic vaping usage in the past 30 days.

Table 4. Proportion of all respondents with predictor by method

	2018 FYSAS Survey (actual)	**Modeling: Independent**	**Modeling: Dependent**
Assignment to Cluster 2	0.180	0.180	0.180
Assignment to Cluster 1	0.569	0.567	0.566
Never Smoked Marijuana	0.791	0.791	0.791
Used Alcohol or Illicit Drugs in Last 30 Days	0.223	0.224	0.223
No Best Friends Who Smoke Marijuana	0.647	0.645	0.646
No Best Friends Tried Alcohol	0.605	0.604	0.604
Never Drank Alcohol	0.590	0.588	0.591

3.1.2.1 Results from an Independent Model

In the independent model, we modeled each of the seven predictor variables as a function of the demographic characteristics alone. This method implicitly assumes that the predictors are independent of one another. We used a logistic regression for each of the seven predictor variables V_i, where *i* corresponds to each of the variables and ranges from 1 to 7.

$$\underline{\ln(\frac{p(V_i)}{1-p(V_i)}) \sim b_0 + b_1 * race + b_2 * gender} \tag{1}$$

We use each of the estimates together in the model for probability of vaping as illustrated in **Equation (2)**.

$$\ln\left(\frac{p(Y)}{1-p(Y)}\right) \sim B_0 + B_1 * V_1 \ldots + B_7 * V_7 + \ B_8 * leaf \tag{2}$$

We assumed that each county follows the same overall model but differs only in the baseline prevalence, thus we included a fixed effect for each county as presented in **Equation (3)**.

$$\ln(\frac{p(Y)}{1-p(Y)}) \sim B_0 + B_1 * V_1 \ldots + B_7 * V_7 + \ B_8 * leaf + \ C_i \tag{3}$$

Equation 3 produced probabilities for each member of the synthetic population, and from these probabilities we randomly assigned a binary indicator for vaping. Predicted estimates, however, showed some bias when validated on the overall prevalence. Prevalence averaged over 100 simulation runs produced the value of 0.0985 (95% range 0.098 to 0.099) while the observed proportion in 2018 FYSAS dataset was 0.139 (**Table 1**).

3.1.2.2 Results from a Sequential Model

In the sequential model, the first predictor is modeled using only demographic characteristics. The second predictor is modeled using demographic characteristics along with the first predictor. The third predictor is modeled using demographic characteristics along with the first and second predictors. The process continues through the seventh predictor. The order of the variables to enter the predictive models is determined by the goodness of fit characteristic, i.e., we select the variable that is best explained by the existing variables. This process is summarized with **Equation (4).**

$$\underline{\ln(\frac{p(V_1)}{1-p(V_1)}) \ \sim b_0 + b_1 * race + b_2 * gender} \tag{4}$$

$$\underline{\ln(\frac{p(V_2)}{1-p(V_2)}) \sim b_0 + b_1 * race + b_2 * gender + \ b_3 * \ V_1}$$

$$\ldots$$

$$\underline{\ln(\frac{p(V_7)}{1-p(V_7)}) \sim b_0 + b_1 * race + b_2 * gender + \ b_3 * \ V_1 + b_4 * \ V_2 + b_5 * \ V_3 + b_6 * \ V_4 + b_7} \\ \underline{* \ V_5 + \ b_8 * \ V_6}$$

After the predictors were selected, we used **Equation C** to calculate the probability of vaping in the last 30 days and then randomly assign a binary vaping indicator based on this probability. The use of sequential assignment produced virtually no bias when compared to the overall prevalence. The average across 100 runs produced value of 0.136 (95% 0.135 to 0.137) compared to 0.139 in 2018 FYSAS dataset (**Table 1)**

To investigate the impact of changing the order of the sequential prediction, we repeated this method for 50 randomly selected orderings of the seven variables. For each ordering, we calculated the proportion of individuals who are assigned that variable. The variability in the proportions of predictor variables as well as in the outcome variable (vape) are highlighted in **Table 5**. Overall, the variability is low, even with different orderings of variables for sequential prediction, suggesting that order did not greatly impact the predictions.

Table 5. Variability in Predictors with Different Orderings of Sequential Prediction (50 replications)

Variable	Min Proportion	Max Proportion	Standard Deviation
Assignment to Cluster 2	0.173	0.183	0.0029
Never Smoked Marijuana	0.786	0.791	0.0016
Used Alcohol or Illicit Drugs in Last 30 Days	0.209	0.215	0.0019
No Best Friends Who Smoke Marijuana	0.663	0.670	0.0022
No Best Friends Tried Alcohol	0.621	0.625	0.0008
Never Drank Alcohol	0.610	0.615	0.0018
Assignment to Cluster 1	0.589	0.597	0.0019
Vaped in past 30 days	0.134	0.138	0.0009

3.1.3 Results for Small Communities

In some cases, there is a very small number of records for a subset of interest. For example, there are cases where a county has only one individual of a specific race/gender/age combination in the synthetic population. These small samples present disclosure concerns if there is one matching record in the survey data that could be linked to this one synthetic person. This would effectively recreate the actual record, potentially exposing private information, particularly in the case of sampling. For example, for few specific combinations of county, age, race and gender in our sample we get either zero or a single subject who reported vaping in the past 30 days. With the sampling method, the value is always the same as the actual (single) student's survey value. With the independent and dependent methods, there is increased variability with the mean of 0.07 and standard deviation of 0.066, and with the mean of 0.08 and standard deviation of 0.074 respectively.

4. DISCUSSION

We presented several methods to link multiple datasets using synthetic populations. Based on our analysis and as highlighted in **Table 1**, the dependent modeling method most closely approximated the overall state proportion of youth using electronic vaping products in the last 30 days. We suspect that this is a result of the model better accounting for the correlation between variables than the independent model and allowing for more coverage than the resampling method.

Figure 1 highlights the relationship between the difference in prediction of proportion vaping for the dependent and independent methods, and the average correlation strength between the variables used in the prediction for each age/race/gender subset. Age, race, and gender subsets with fewer than 30 individuals were excluded from analysis. The positive trend suggests that the difference in predictions is greater between the dependent and independent methods when the average correlation strength between the variables is higher for a given subgroup. Capturing this correlation improves the accuracy of vaping proportion estimates.

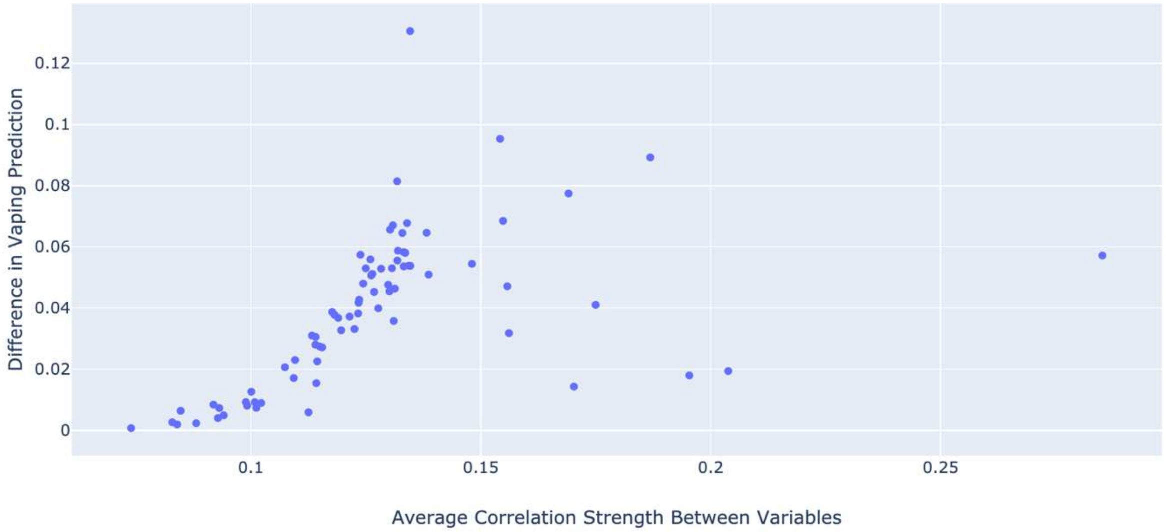

Figure 1. Difference in Prediction of Vaping Proportion Between Dependent and Sequential Methods by Strength of Correlation Between the Variables Used in the Analysis for Each Race/Age/Gender Subset

The dependent method also accounts for more variability than the resampling method. Nearly 30% of FYSAS weighted respondents were never sampled, raising concerns about the representativeness of the synthetic cohort created through this process. We also determined that, as highlighted in **Table 5**, the ordering of sequential prediction in the dependent method does not substantially impact the overall prediction. We demonstrated that it is possible, though unlikely, to recreate a record for an actual person when the subsample is very small. Common recommendations to prevent potential identification, is to either aggregate data to a larger subpopulation, or to exclude subsets with small counts. With model-based assignment it is critical to avoid models with small (e.g. less than 10) degrees of freedom to assure that all assignments are probabilistic with enough variability.

Although here we illustrated our approach with a simple example, the approach is not limited to the number of datasets that could be linked. One can link two, three, or more datasets. Whether the same methodology (e.g., resampling or modeling) should be used in linking all datasets or there are separate preferred methods for each dataset is to yet to be seen from systematic studies because accounting for error propagation under the applications of different methodologies is the subject of future research.

Our methodology can be used not only to create analysis datasets, but also to provide a basis for microsimulation modeling. For example, in Subramanian et al. 2009[6] we have considered a synthetic population representative of the US population and probabilistically linked longitudinal clinical profiles to address cancer progression in the US population under different intervention scenarios.

At the same time, these methodologies are at their early stages and need more development. So far, we have considered two major categories: profile matching (resampling) and modeling. Each has their advantages and limitations. Profile matching reproduces the joint distributions but has limitations with variability and requires control for small cells. Modeling can produce smoother estimates with broader coverages of the distribution ranges, and can lead to less bias, but is dependent on model accuracy and needs more understanding of variable selection algorithms and accumulating uncertainty. Potentially, a combination of both methods can lead to the improvements of both the capturing of variability and accuracy.

The use of synthetic population will result in profiles that have five demographic variables randomly selected. Thus, any matching of real profiles to the simulated ones is purely coincidental. In our approach we use probabilistic models to build synthetic profiles from additional datasets. For example, we might assign a colorectal cancer status to a specific person by drawing a random assignment from the modeled probability of cancer given other characteristics. The probability of revealing cancer diagnosis from a specific individual in the real population is not better than just flipping a coin with model-based probability. Thus, by its construction synthetic population allows one to link multiple datasets without breaching confidentiality. This is true for the model-based approach where all values are probabilistically generated. In the resampling approach, the use of the resampling method is safeguarded by ensuring that the selected profile is not unique, but rather is randomly drawn from a sufficiently large group of profiles with a smaller matching probability. This is a common practice in the creation of public use analytic files and does not allow one to obtain a unique matching profile. Usually, that requirement is to have at least 50 observations in the cell.

A formal examination of such coincidental matching is nevertheless needed to establish proven guidelines for the use of data linkage methods. Even with synthetic populations, the linkage processes we describe may potentially pose disclosure risks. These risks are not yet fully investigated and are subjects of our current research.

REFERENCES

1 Krauland, M. G. *et al.* Development of a Synthetic Population Model for Assessing Excess Risk for Cardiovascular Disease Death. *JAMA Netw Open* **3,** e2015047, doi:10.1001/jamanetworkopen.2020.15047 (2020).

2 Bates, S., Leonenko, V., Rineer, J. & Bobashev, G. Using synthetic populations to understand geospatial patterns in opioid related overdose and predicted opioid misuse. *Computational and Mathematical Organization Theory* **25,** 36-47, doi:10.1007/s10588-018-09281-2 (2019).

3 NIH obesity challenge 2015. *RTI International named an Obesity Data Challenge winner*, <https://debeaumont.org/news/2015/obesity-data-challenge-winners-announced/> (2015).

4 Subramanian, S., Bobashev, G. V., Morris, J. & Hoover, S. Precision medicine for prevention: Can personalized, risk-based screening decrease colorectal cancer mortality at an acceptable cost? *Cancer Causes & Control* **28,** 299-308 (2017).

5 Subramanian, S., Bobashev, G. V. & Morris, J. R. Cost implication of the new colorectal cancer screening guideline: how do we ensure optimal allocation of limited resources? *Health Affairs* **29,** 1734-1740 (2010).

6 Subramanian, S., Bobashev, G. V. & Morris, R. J. Modeling cost-effectiveness of colorectal cancer screening: policy guidance based on patients preferences and compliance. *Cancer Epidemiology Biomarkers and Prevention* **18,** 1971-1978 (2009).

7 Karr, A. F. *et al.* Comparing record linkage software programs and algorithms using real-world data. *PLoS One* **14,** e0221459, doi:10.1371/journal.pone.0221459 (2019).

8 Wheaton, W. D. *et al.* Synthesized Population Databases: A US Geospatial Database for Agent-Based Models. Methods report (RTI Press) (NIH Public Access), no. 10. 905 (2009).

9 Bobashev, G. V. *et al.* Geospatial forecasting of COVID-19 spread and risk of reaching hospital capacity. *SIG SIGSPATIAL Special Interest Group on Spatial Information* **12,** 25-32, doi:10.1145/3431843.3431847 (2020).

10 Bae, K.-H. *et al. Development of large-scale synthetic population to simulate COVID-19 transmission and response. Proceedings of the 2020 Winter Simulation Conference*. (2020).

11 Akbarpour, M. *et al. Socioeconomic network heterogeneity and pandemic policy response. Working Paper 20-025*, <https://siepr.stanford.edu/research/publications/socioeconomic-network-heterogeneity-and-pandemic-policy-response> (2020, June 15).

12 Genevieve, L. D., Martani, A., Mallet, M. C., Wangmo, T. & Elger, B. S. Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review. *PLoS One* **14,** e0226015, doi:10.1371/journal.pone.0226015 (2019).

13 Little, R. J. A. & Rubin, D. B. in *2002 Book Series: Wiley Series in Probability and Statistics* (2014, August).

14 Bollen, K., Biemer, P. P., Karr, A. F., Tueller, S. & Berzofsky, M. E. Are survey weights needed? a review of diagnostic tests in regression analysis. *Annual Review of Statistics and Its Application* **3,** 375-392, doi:10.1146/annurev-statistics-011516-012958 (2016).

15 Nelsen, R. B. An Introduction to Copulas. ISBN 978-0-387-98623-4. (Springer, 1999).

ACKNOWLEDGEMENTS, FUNDING SOURCES

This study was funded in part by a Grant from the National Institutes of Health 1R01DA047994-01A1 (Bobashev, PI)

AUTHOR CONTRIBUTIONS

GB and AK conceived the study, JR developed and provided synthetic populations, EH and CK has conducted the analysis. All authors participated in writing of the manuscript.

ADDITIONAL INFORMATION

Competing Interests

The authors have no competing interests.

Availability of Data and Materials

The 2018 Florida Youth Substance Abuse Survey data are maintained by the Florida Department of Children and Families but were used under license for the current study and are not publicly available. The 2017 RTI US Synthetic Population is maintained by RTI International and is not publicly available but may be available from the corresponding author on reasonable request. The 2010 RTI US Synthetic Household Population is public available at https://fred.publichealth.pitt.edu/syn_pops.

Figures

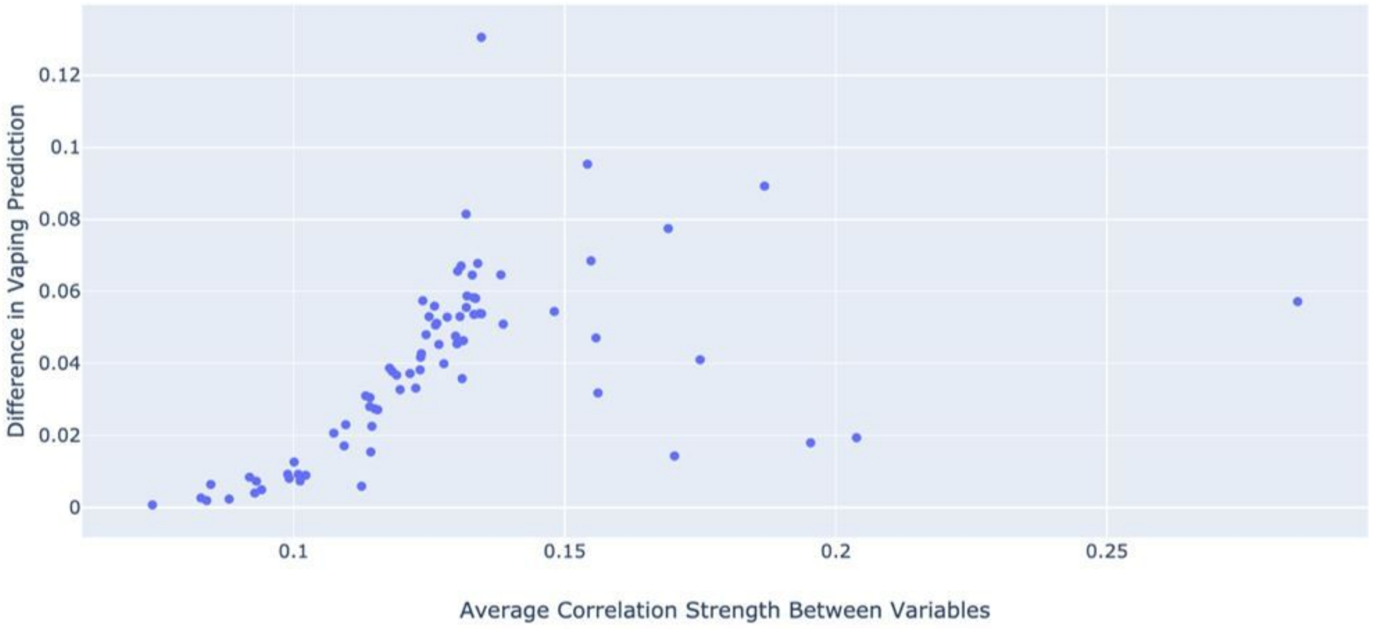

Figure 1

Difference in Prediction of Vaping Proportion Between Dependent and Sequential Methods by Strength of Correlation Between the Variables Used in the Analysis for Each Race/Age/Gender Subset