The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

Document Title:	**Assessment of Sexual Assault Kit (SAK) Evidence Selection Leading to Development of SAK Evidence Machine-Learning Model (SAK-ML Model)**
Author(s):	**Julie L. Valentine Ph.D., RN, SANE-A, FAAFS, FAAN**
Document Number:	**309199**
Date Received:	**June 2024**
Award Number:	**2019-NE-BX-0001**

TECHNICAL SUMMARY
for
Research and Evaluation for the Testing and Interpretation of Physical Evidence in Publicly Funded Forensic Laboratories
Grants.gov No. NIJ-2019-15507
Award: 2019-NE-BX-0001

Utah Bureau of Forensic Services *PC: Paul Richer*

Project Title: Assessment of Sexual Assault Kit (SAK) Evidence Selection Leading to Development of SAK Evidence Machine-Learning Model (SAK-ML Model)

Award Recipient Organization: Utah Bureau of Forensic Services

Principal Investigator: Julie L. Valentine PhD, RN, SANE-A, FAAFS, FAAN

NIJ Award Number: 2019-NE-BX-0001

Project Period: January 1, 2020 – December 31, 2022 (extension granted through 12/31/2023)

Award Amount: $250,000

Acknowledgements

We acknowledge the numerous forensic nurses who provided compassionate care to the 11,715 patients included in this study. We acknowledge the work and dedication of the many forensic scientists who analyzed the evidence from 9,599 sexual assault kits referenced in this research. We wish to express our sincere gratitude for the many undergraduate and graduate students from Brigham Young University who tirelessly worked to collect and analyze this data: Connor Alder, Carolyn Allen, Nicole Asay, Emily Black, Brian Brown, Andrew Criddle, Samantha Eckery, Deborah Fry, Aubrey Gibbons, Breanna Hall, Adia Hansen, Luke Johnson, Jake Momberger, Sam Pugh, Lauren Schagel, and Whitney Wagner. Additionally, we acknowledge the invaluable support provided by Dr. Sam Payne in biostatistics and Dr. David Grimsman in computer science. Lastly, we must acknowledge the sexual assault survivors represented in this data. It is our hope that our efforts to capture and analyze their experiences will contribute to advancements in multidisciplinary practices and policies.

Table of Contents

PROJECT SUMMARY

Few studies have explored aggregated DNA analysis findings from sexual assault kits (SAKs) and predictive features of developing useful DNA information related to the foreign contributor(s). Information gleaned from evaluating DNA analysis findings have significant practice and policy implications for both forensic medical examiners/sexual assault nurse examiners and forensic scientists. Results from this innovative study were obtained by tracking SAKs from evidence collection, data from sexual assault medical forensic examinations, through DNA analysis results, data from publicly funded laboratories.

Goals and Objectives of this study were as follows:

The proposed research study addressed the gap in research on SAK evidence selection protocols to establish best practice guidelines for SAK evidence selection for analysis and also explore the development of a Sexual Assault Kit evidence Machine Learning Model (SAK-ML Model) software program. Therefore, the study had two purposes:

- To evaluate decision-making protocols on DNA evidence contained in SAKs to develop research-based guidelines regarding which swabs and how many swabs should be tested by crime lab (Part 1).
- To develop, implement and evaluate a machine learning statistical model, SAK-ML Model to guide forensic scientists within publicly funded forensic laboratories on the selection of the most probative SAK swabs to analyze (Part 2).

The overarching goal of the study was to extract and analyze information related to SAK evidence collection and analysis to inform practice and policy.

Background and Review of the Literature

Victims of sexual assault who report within five days of the assault are given the choice to have evidence collected in a (SAK). In the United States (U.S.), forensic nurses or sexual assault nurse examiners (SANEs) are specially educated registered or advanced practice nurses who conduct sexual assault medical forensic examinations (SAMFEs). While the main objective of SAMFEs is to provide trauma-informed, patient-centered care to the victim, evidence is collected, packaged, and sealed in SAKs by SANEs if victims request evidence collection. The SAKs are then given to law enforcement who decide to submit or not submit the SAK to their designated crime laboratories. Within the last decade, SAK submission rates have increased dramatically, with some states passing laws to submit all SAKs. The crime laboratories conduct testing and DNA analysis on evidence contained within the SAKs.

The primary goal of the crime laboratory in testing SAKs is to provide unbiased forensic analysis of evidence collected from the victim's body to the criminal justice community. Generally, polymerase chain reaction (PCR) short-tandem repeat (STR) DNA is the preferred analysis method as STR DNA profiles can be uploaded and searched in the Federal Bureau of Investigation (FBI) Combined DNA Index System (CODIS) database. Federal law requires crime laboratories to meet specific guidelines and accreditation standards to be eligible to upload DNA profiles into CODIS. Additionally, the evidence as well as the profiles developed from that evidence must meet specific criteria for eligibility for a CODIS upload. CODIS consists of the National DNA Index System (NDIS), State DNA Index Systems (SDIS), and in some jurisdictions Local DNA Index (LDIS) (FBI, n.d.).

To improve SAK analysis efficiency, crime labs have implemented a variety of strategies, including increasing personnel, utilizing robotics and updated processing equipment, and adopting a direct to DNA analysis approach. Additionally, many crime labs have opted for a selective swab method in which forensic analysts will select the most probative swabs within the SAKs based on their expertise, the crime scenario, and the documentation of injuries to analyze swabs more likely to provide DNA rather than analyzing all submitted swabs and associated evidence.

Few studies have been conducted on the percentage of SAKs that produce STR DNA profiles of foreign contributors entered into CODIS. In a study in Detroit, Campbell and colleagues (2020) found that 40.3% of their random sample of SAKs (n = 7,287) yielded an uploaded CODIS DNA profile. Researchers in Ohio conducted a random sample of 2,500 previously unsubmitted SAKs (representative of the entire state) and found that 57.0% yielded at least one uploaded CODIS DNA profile (Kerka et al., 2018). Researchers in Los Angeles analyzed 1,948 backlogged SAKs and reported that 35.9% produced at least one uploaded CODIS DNA profile (Peterson et al., 2012). Researchers of a similar study of backlogged SAKs in New Orleans found that 25.4% developed uploaded CODIS DNA profiles (Nelson, 2013). In Houston, researchers evaluated 491 previously unsubmitted SAKs and found that 43% were uploaded into CODIS (Davis et al., 2021). In a study testing machine-learning models for SAK forensic evidence selection, Wang and colleagues (2020) found 46.9% of SAKs developed uploaded CODIS DNA profiles. In summary, prior published studies have reported a fairly wide range, from 25.4% to 57%, of SAKs developed STR DNA profiles of foreign contributors uploaded into CODIS.

Minimal research has been published on features associated with the development of STR DNA profiles entered into CODIS. Kerka and colleagues (2018) reported statistically significant factors in predicting development of CODIS entered STR DNA profiles from previously unsubmitted SAKs, including length of time between assault and exam, length of time between evidence collection and forensic analysis, victim's age, and occurrence of consensual sex within 120 hours of evidence collection. Regarding age variable, they reported that pediatric victims and adult victims over the age of 50 years were less likely to have SAKs with STR DNA profiles entered into CODIS (Kerka et al., 2018). Wang and colleagues (2020) examined the cost-effectiveness of using a machine learning model to predict which swab samples to test from SAKs to maximize the development of CODIS eligible DNA profiles. They found that machine learning algorithms outperformed sexual assault forensic examiners at identifying the most probative samples, suggesting that the yield of CODIS eligible DNA profiles would increase by 47.2% by testing swabs selected through the algorithm rather than the selective swab approach by forensic scientists (Wang et al., 2020).

The research questions explored in this study add to the knowledge bases of the few published articles on the development of STR DNA profiles of foreign contributors entered into CODIS from SAKs and their predicting features.

Research Questions

The study contains seven research question sections assigned to either Part 1 or Part 2.

Research questions under Part 1 of the study:

- Research question #1: What differences exist between forensic scientists in the selection and prioritization of SAK swabs for analysis?
- Research question #2: What differences occur in the aggregated percentages of the development of CODIS-entered DNA profiles when testing one swab, a few selected swabs, or testing all swabs contained in SAKs?
- Research questions #3 A-C:

    - A. In cases with selected swabs for analysis, which swabs analyzed for STR DNA are more likely to yield STR DNA profiles entered into CODIS?
    - B. In cases that analyzed all swabs, which swabs analyzed for STR DNA are more likely to yield STR DNA profiles entered into CODIS?
    - C. What differences exist between the different approaches of swab selection (test 1, test selected, or test all) on which swabs are more likely to yield STR DNA profiles entered into CODIS?
- Research questions #4 A& B:
    - A. What victim and sexual assault variables are statistically significant in predicting the development of STR DNA partial or full profiles of unknown contributor(s)?
    - B. What predicting variables are associated with development of STR DNA profiles entered into CODIS based upon swab location?

Research questions under Part 2 of the study:

- Research question #5: What is the reliability and validity of the SAK-ML software program in predicting STR DNA profiles entered into CODIS using retrospective data?
- Research question #6: Which method of selecting swabs from SAKs (forensic analysts determine which swabs to analyze and number of swabs, OR use of SAK-ML Model) yields a higher percentage of STR DNA profiles entered into CODIS?
- Research question #7: What is the impact of using SAK-ML Model on the following outcomes: development of STR DNA profiles entered into CODIS, crime lab efficiency, and crime lab cost savings?

Summary of Project Design and Methods

Study Population

The study population consisted of victims age 14 years and older who received a SAMFE from one of the participating forensic nursing teams and had an unrestricted SAK collected. Years of inclusion are 2010-2022 in Utah, 2015-2020 in Orange County, and 2013-2020 in Idaho.

Study Settings

Three publicly funded crime laboratories were collaborative research partners: Utah Bureau of Forensic Services (UBFS), state crime laboratory in Utah; Orange County Crime Lab (OCCL), county crime laboratory in Orange County, California; and Idaho State Police Forensic Services (ISPFS), state crime laboratory in Idaho. As the DNA analysis interpretation methods utilized by crime labs impacts findings, it is important to note that binary interpretation approach was employed during the study period at the sites.

The primary research site was the Utah Bureau of Forensic Services (UBFS) and the SAKs collected throughout Utah from 2010 to 2022 (N=8,981, submitted SAKs of 6,865). Utah is a Mountain West state in the U.S. with a population of approximately 3.4 million (U.S. Census Bureau, 2022).

The other research sites included in this study included the state of Idaho and Orange County, California. Idaho is a Northwestern state in the U.S. with a population of approximately 1.94 million (U.S. Census Bureau, 2022). Idaho consists of urban, suburban, and many rural communities. The state crime lab is Idaho State Police Forensic Services (ISPFS) located in Meridian, Idaho. The project team for this study traveled to ISPFS from Provo, Utah, several times to extract data from the crime lab database as data collection was only available through in-person extraction. Unfortunately, the Idaho study data does not contain information from the

SAMFE charts due to the inability to obtain clearance from each forensic nursing team in Idaho. Data regarding victim and assault features were obtained from a one-page summary of the case completed by forensic examiners and/or police reports. Not all of the Idaho cases contained this additional information, so data points are missing (see Appendix A).

Orange County, California is a large county in Southern California with a population of approximately 3.15 million (U.S. Census Bureau, 2022). Substantial data was obtained from the SAMFE charts in Orange County although less data than the Utah cases. The primary data obtained from the Orange County Crime Lab (OCCL) was on the outcome findings from STR DNA analysis per analyzed swab sets. Therefore, the Orange County data has fewer data points on crime lab features than Idaho and Utah (see Appendix A).

Project Data Collection

The study was an exploratory, retrospective design with data retrieved from SAMFE charts and crime lab DNA reports. The research team extracting the data consisted of Dr. Julie L. Valentine (PI), Dr. Leslie Miles (Co-investigator), two graduate students, and six undergraduate students. The research team had already obtained several years (2010 to 2018) of Utah data before beginning this study on January 1, 2020. Memorandums of Understanding were signed by the participating agencies prior to data collection.

Utah Data Collection

The additional Utah data (2019 to 2022) was collected by manually extracting the data on collected SAKs from eight Utah counties, comprising 82% of the state's population, from forensic electronic medical records and crime lab DNA reports and coding de-identified information directly into the study's database in SPSS 28 (*N*=6885 submitted SAKs). The research team received research access to the SAMFE data in the electronic forensic electronic medical records. Data collection of the state crime lab data was initially completed by the research team at the state crime lab. When the COVID-19 pandemic occurred, data collection stopped for a few months as the crime lab was inaccessible to research personnel. In July 2020, the research team was granted remote access with protected access only granted to Dr. Valentine (PI). The research team coded the crime lab data together at Brigham Young University in Provo, Utah. A detailed codebook was developed to guide coding decisions. All data coding was conducted as a team to allow discussion of any coding questions. Approximately 10% of the cases were re-coded by Dr. Valentine or Dr. Miles to conduct Cohen's kappa test to assess interrater reliability. Cohen's kappa remained over .90 across all variables, indicating high interrater reliability.

Orange County Data Collection

Data was collected on SAKs obtained by Forensic Nurse Specialist, Inc., forensic nursing team in Orange County, and submitted to Orange County Crime Lab (OCCL) from 2015 to 2020 (*N*=1207). The initial plans to obtain the Orange County data were for the research team to travel to Orange County to extract the data from Forensic Nurse Specialists, Inc. and the Orange County Crime Lab (OCCL). These plans were not possible with the COVID-related travel flight bans imposed by Brigham Young University (academic institution of research team) from 2020 to 2021 In fall 2021, Dr. Valentine received clearance to fly to Orange County to meet with the directors of Forensic Nurse Specialists, Inc. and the OCCL to develop a data extraction plan. Dr. Valentine and the directors agreed upon selected features to collect from the SAMFE and crime lab records that would not put an undue burden on their agencies. Following completion of this data extraction, a password-protected, de-identified dataset of the Orange County data was sent via secure email to Dr. Valentine in late 2021. Following data cleaning and coding to match the study code book, the data was then exported into the SPSS 28 dataset in early 2022.

Idaho Data Collection

Data was collected from SAKs submitted to the ISPFS from 2013-2020 (N=1527). The Idaho data was obtained directly from the ISPFS database and de-identified information coded into SPSS 28. Due to the COVID-19 pandemic, travel to Meridian, Idaho, was not initially approved by the university. Travel was granted in August 2020 with mandated stipulations to protect any COVID-19 infection, including no flights, travel with single passengers in each vehicle, and single occupancy in each hotel room. The research team collected the data in person at ISPFS and supervised by ISPFS personnel. Several automobile trips to Meridian, Idaho, were made by the research team in the summers of 2021 and 2022 to fully complete data extraction and coding. Again, Cohen's kappa was calculated throughout the data coding process to assess interrater reliability and remained over .90, indicating high interrater reliability.

Methodology

Prior to analysis, the data was checked for outliers and inconsistencies with descriptive statistics (frequencies, means, modes, and standard deviations). The descriptive statistics for the three sites are reported in the *Appendix*.

The next steps in the analysis process were to develop a form of logistic regression machine-learning models to evaluate predictive features and interactions of features with the case outcome feature of foreign contributor STR DNA profiles uploaded into SDIS CODIS. Additionally, models were created to evaluate features that predicted the development of full or partial STR DNA profiles of foreign contributors by swab location (perianal, vaginal, rectal, breast(s), cervical, oral, body area not including neck or breast(s), neck, underwear, other clothing, other items not including clothing or bedding, and bedding). As this portion of the analysis required multiple steps, the description of the methodology is lengthy. The steps for developing the machine learning models are outlined below and a summary contained in the *Data Archiving Plan* on the National Archive of Criminal Justice Data (NACJD) website.

To prepare a model to predict the outcomes of swab DNA testing, we turned to logistic regression as a form of machine learning, rather than other conventional machine learning models. The purpose behind this strategy was two-fold: we wished to both predict the outcomes and explain why the predictor made the prediction it did. For most machine learning models, including K-Nearest Neighbor Classifiers, Multi-Layer Perceptrons, and Random Forest Classifiers, it is difficult to retrace the training of the algorithm to know exactly why the model made the decision it did. This methodology stands in contrast with logistic regression; with this statistical machine-learning model, we can see the impact of each answer to each question on the outcome prediction, thus helping us to understand for those swabs that were tested which questions are most important in predicting whether the DNA test would be successful.

In processing the datasets from Idaho, Orange County, and Utah, we followed similar patterns to prepare the data for analysis. Initially, because the predicting variables and the relevant swabs were distinct for individuals of different genders, we divided each dataset into two: female data and male data. The Orange County and Idaho datasets had low numbers of male victims (n=48), so we only completed modeling on data from the female victims from this site. In all three datasets, there was not a sufficient number of transgender or intersex individuals to contribute substantially to statistical analysis. Therefore, the modeling findings represent only binary gender identity: male and female. The end result were four datasets: Female Utah, Male Utah, Female Orange County, and Female Idaho.

Most of the questions in the original dataset, with some notable exceptions such as age and time between assault and exam, were categorical, primarily no (0) or yes (1). However, due to the experiences of the individuals before and during the data collection, the categorical questions also included responses of unknown" or uncertain" often due to the traumatic state and loss of consciousness or awareness, either from trauma or intoxication, experienced by the victim at the time of their assault. All of these responses (no, yes, and unknown) included important information, so to provide the best information possible to the training model, we analyzed the results of each of those columns based on whether or not the victim had a positive response in that column.

Logistic regression modeling and most other machine models cannot automatically handle unknown values in continuous variables, such as age and number of injuries. In our dataset, we found comparatively few continuous variables containing unknown values. To address the few unknown values in the continuous variables, we performed a standard mean imputation on those columns, filling those empty answers with values that had a low impact on the resulting decision.

We also dealt with many sparse columns, variables for which almost all of the responses were the same, with only a few differing values. These columns are prone to spurious correlations–for example, if only a few people answered unknown" to a question. Still, everyone received a positive result; that question would appear to be a powerful predictor even if it occurred randomly. With fewer variables, we might accept those conclusions as potentially valid. However, with the large number of features in the dataset, including multiple addressing each question, and with the relatively low number of people in the dataset for machine learning purposes, we elected to drop variables that had less than a threshold of 1/10 of their values that were different than the most common value. This process helped to reduce the number of questions that appeared to predict the outcome better than they actually did, allowing us to focus on the variables that more reliably improved our predictions.

In performing machine learning logistic regression modeling, we sought to both analyze the effectiveness of individual columns, as well as make decisions based on the combination of multiple columns. As an example, if a person had a low amount of time between the assault and the exam and they also bathed or showered between the assault and exam, that may tell us more than looking at the two variables separately. This interaction was analyzed by multiplying the values of each of the two columns and then adding that result as an additional column.

We also understand the different columns' impact by the coefficients' values that apply to that column. We sought to address two questions, each requiring different treatments of the dataset itself. The first question was, "How much does a change in the response to one variable change the prediction?" To answer this question, we ran the logistic regression on the datasets directly, once with and once without the extra multivariate columns mentioned above.

The coefficients found for each variable provided information known as the log odds, which allowed us to analyze how much a change in one variable increased or decreased our expectation of the outcome variable. The exponentiated coefficients were intercepted as change in odds ratio per unit change in the input. For example, if an exponentiated coefficient had a value of 1.5, then every 1-unit increase in the variable associated with that coefficient would result in a 1.5 times increase in probability in the outcome, whereas a 2-unit increase would result in a 3.0 increase in probability of the outcome.

The second question was, "Which predicting variables were most important in estimating the outcome variable?" In machine learning logistic regression, the coefficients generally demonstrate how much impact each variable has on the prediction, but this can be skewed if two columns have the same predictive power while one has much larger values than the other. For example, if the mean value for age is around 30 but the mean for 'Yes' on suspect action verbal is a 1, the coefficient of age will be much smaller than suspect action verbal to compensate for the difference. Thus, to evaluate which variables have the greatest predictive power, we had to first scale all the columns so that the variations of all the columns are the same size before running the logistic regression again on the scaled datasets. We scaled the datasets by subtracting the mean of each column from all of the values in the column and then divided the values in that column by the standard deviation. After we performed logistic regression, this scaling technique allowed us to rank each variable from most to least helpful in predicting by sorting the coefficients by their absolute value.

Additionally, when we normalized the data, we used min/max normalization on continuous columns only. So, for example, we normalized the "Age" variable so that the minimum age was zero and the maximum age was 1. All the other variables that were already coded as binary 1/0 values remained the same. We found improved model performance by using this method rather than a "mean & standard deviation" normalization technique.

References supporting the statistical modeling decisions are listed in the “References” section.

Summary of Results

The research results are reported under each research question. Additional findings of interest not specifically found under research questions are reported at the conclusion of the research question results.

Results From Research Questions

Research question #1: What differences exist between forensic scientists in the selection and prioritization of SAK swabs for analysis?

In exploring an answer to this question, an internal audit of an individual crime lab, UBFS, was considered. However, an internal audit could not be conducted in a way that would have implications for other laboratory systems, so instead, a comparison of swabs selected for testing within the three crime labs, UBFS (Utah), OCCL (Orange County), and ISPFS (Idaho) was done.

The table below contains swab choices within the three crime labs. Interestingly, the top three swab locations selected for analysis in UBFS and OCCL were in the same order: perianal, vaginal, and breast(s). The top three choices for ISPFS were vaginal, perianal, and rectal. The decision to test the perianal swabs varied significantly between the crime labs (52%, 45.2%, & 28.3%). As noted in Table 1, the rectal swab had substantial variability in the decision to test swabs from this location, ranging from 24.8% to 15% to 2.7%. In answering the question regarding swab selection variability between crime labs, we found some similarities and differences. All three labs were similar in the fact that perianal and vaginal swabs were the swab locations most frequently selected for analysis. Differences in the swab location percentage distributions were found among the remaining swabs. The similarities and differences found in this analysis may partially speak to the question of consistency of swab selection and prioritization among analysts and between laboratories. Consequently, a more rigorous study would need to be conducted to ascertain differences in swab selection and prioritization among forensic scientists.

Table 1. Swabs Selected for Analyses by Crime Labs

Ranking	Utah Data (UBFS) (N=6865)	Orange County Data (OCCL) (N=1207)	Idaho Data (ISPFS) (N=1572)
1	Perianal (n=3574) 52%	Perianal (n=546) 45.2%	Vaginal (n=734) 46.7%

2	Vaginal (n=3273) 47.7%	Vaginal (n=282) 23.4%	Perianal (n=445) 28.3%
3	Breast(s) (n=1503) 21.9%	Breast(s) (n=252) 20.9%	Rectal (n=390) 24.8%
4	Rectal (n=1031) 15%	Body area, not including neck and breasts (n=126) 10.4%	Body area, not including neck and breasts (n=199) 12.7%
5	Neck (n=925) 13.5%	Neck (n=112) 9.3%	Breasts (n=204) 13%
6	Body area, not including neck/breasts (n=908) 13.2%	Oral (n=63) 5.2%	Neck (n=185) 11.8%
7	Cervical (n=772) 11.8%	Cervical (n=57) 4.7%	Oral (n=313) 19.9%
8	Oral (n=442) 6.7%	Rectal (n=32) 2.7%	Cervix (n=35) 2.2%
9	Underwear (n=59) 0.9%	Underwear (n=11) 0.9%	Underwear (n=16) 1%
10	Other clothing (n=51) 0.8%	Other clothing (n=5) 0.4%	Other clothing (n=8) 0.5%
11	Other items, not clothing or bedding (n=20) 0.3%	Other items, not clothing or bedding (n=2) 0.2%	Condom (n=8) 0.5%
12	Condom (n=18) 0.2%		Bedding (n=2) 0.13%
13	Bedding (n=14) 0.2%		Tampon (n=3) 0.19%
14	Tampon (n=6) 0.09%		Other items, not clothing or bedding (n=11) 0.7%

<u>Research question #2</u>*:* What differences occur in the aggregated percentages of the development of CODIS-entered DNA profiles when testing one swab, a few selected swabs or testing all swabs contained in SAKs?

Initial exploration into this research question indicated a potential likelihood of developing profiles from foreign contributors when swabs from more than three areas of the body were analyzed. However, upon further consideration, it was determined that other important factors would need to be considered before meaningful recommendations could be made. Some of those important factors include the following: how many swabs were collected, how many perpetrators were involved with the assault, the nature of the contact involved, if there was consensual activity within five days prior to the evidence collection, etc.

We further explored the answer to this question by comparing foreign contributor profiles uploaded into CODIS (SDIS) in the three participating crime labs. Each crime lab has their own protocols for selecting how many swabs to test within SAKs. Forensic scientists at UBFS use their expertise to select the most probative swabs based upon the victims account of the assault and the SANE documentation at the time of exam as recorded in the SAMFE record. OCCL reported that their selection was based upon the assault history in the SAMFE and the expertise of the forensic analysts without a specific number of swabs as a goal. ISPFS reported that their selection was based upon information contained in a one-page summary completed by SANEs of the assault, if the document was uploaded in the crime lab database. For UBFS and ISPFS we were able to complete descriptive analysis on the number of items/swabs tested and found some differences as noted in *Descriptive Data* (Appendix). This analysis was not done with OCCL. Calculation of the mean, median, and mode found more swabs were tested per case at ISPFS compared to UBFS: UBFS mean 3.56, median 3.00, and mode 3; and ISPFS mean 4.26, median 4.00, mode 4.

Overall, **the development of uploaded CODIS (SDIS) profiles varied per crime lab site as follows: UBFS 34.2%, OCCL 46.3%, and ISPFS 33.3%**. These percentages fall within the range of uploaded CODIS profiles reported in the literature of 25.4-57.0%. The data suggests that having a higher mean of samples tested does not necessarily result in a higher percentage of uploaded CODIS (SDIS) profiles. In the comparisons between these two laboratories, selective sampling based on the case scenario yielded a higher percentage of uploaded CODIS profiles. Further research and exploration of confounding variables is needed in this area prior to drawing conclusions.

Further discussion on the varying percentages of uploaded CODIS profiles is contained in the *Applicability to Criminal Justice* section within this report.

<u>Research Questions #3 A-C:</u>

A. In cases with selected swabs for analysis, which swabs analyzed for STR DNA are more likely to yield STR DNA profiles entered into CODIS?

B. In cases that analyzed all swabs, which swabs analyzed for STR DNA are more likely to yield STR DNA profiles entered into CODIS?

C. What differences exist between the different approaches of swab selection (test 1, test selected, or test all) on which swabs are more likely to yield STR DNA profiles entered into CODIS?

After beginning data collection and analysis, we realized that these three questions were more appropriately combined into one question related to the development of full or partial STR DNA profiles of foreign contributors per swab. The DNA analysis findings of individual swabs would not be impacted by the number of swabs selected. Additionally, the outcome variable for swab analysis should be the development of full or partial STR DNA profile rather than uploaded CODIS profiles as the determination of CODIS eligibility extends beyond the DNA analysis findings to other eligibility requirements defined in CODIS requirements. Therefore, the question we answered was the following: **which swabs were more likely to produce full or partial STR DNA profiles of foreign contributors?**

To answer this question, we utilized data from UBFS and ISPFS as the research team extracted and coded the data from these crime labs in the same manner. Data received from OCCL was structured differently with less crime lab information. We evaluated each distinct swab site from selection for testing through STR DNA analysis results. We divided the swabs into categories of internal swabs (vaginal, cervical, rectal, and oral) and external swabs (perianal, breasts, neck, and other external body area). To calculate the percentage of swabs per body area that developed full or partial STR DNA profiles of foreign contributor(s), we divided the swab number of those swabs with full or partial STR DNA profiles of foreign contributor(s) with the number of swabs from that body area selected for male quant (Y-screen) testing. Findings from internal swabs are listed in Table 2 and findings from external swabs are listed in Table 3.

Table 2. Internal Swabs from Male Quant Selection to Full/Partial STR DNA Profiles of Foreign Contributor(s)

	Vaginal		Cervical		Rectal		Oral	
	UBFS	**ISPFS**	**UBFS**	**ISPFS**	**UBFS**	**ISPFS**	**UBFS**	**ISPFS**
Column A Number of Swabs Selected for Male Quant Testing	3273	734	772	35	1031	390	442	313
Column B Number of Swabs with Full/Partial STR DNA of Foreign Contributor	1206	310	336	14	253	107	54	8
B/A = % of Selected Swabs that Produced Full/Partial STR DNA of Foreign Contributor	**36.8%**	**42.2%**	**43.5%**	**40%**	**24.6%**	**27.4%**	**12.2%**	**2.6%**

The findings from the internal swabs indicate that cervical swabs (40-43.5%) had the highest yield of full or partial STR DNA profile development of foreign contributors followed closely by vaginal swabs (36.8-42.2%). Of note, recent federal recommendations advise concentrating DNA on swabs and combining cervical swabs with vaginal swabs as vaginal vault swabs (National Institute of Justice, 2017). SAMFE forms in Utah changed to vaginal vault swab collection in 2018. Rectal swabs had a lower percentage at approximately 25-27% while oral swabs had a substantially lower percentage of full or partial STR DNA profile development at 2.6-12.2%.

Table 3. External Swabs from Male Quant Selection to Full/Partial STR DNA Profiles of Foreign Contributor(s)

	Perianal		**Breast(s)**		**Neck**		**Other Body Areas**	
	UBFS	**ISPFS**	**UBFS**	**ISPFS**	**UBFS**	**ISPFS**	**UBFS**	**ISPFS**
Column A Number of Swabs Selected for Male Quant Testing	3574	455	1503	204	925	185	908	199
Column B Number of Swabs with Full/Partial STR DNA of Foreign Contributor	1317	137	607	91	351	93	278	70
B/A = % of Selected Swabs that Produced Full/Partial STR DNA of Foreign Contributor	**36.8%**	**30.1%**	**40.3%**	**44.6%**	**37.9%**	**50.3%**	**30.6%**	**35.2%**

The swab locations with the highest yield of developing full or partial STR DNA profiles of foreign contributors were the neck and breast(s) swabs. The perianal (30.1-36.8%) and other body locations swabs (30.6-35.2%) also had a high percentage of developing full or partial STR DNA profiles of foreign contributor(s).

Research questions #4 A,B

A. What victim and sexual assault (SA) variables were statistically significant in predicting the development of STR DNA partial or full profiles of unknown contributor(s)?

B. What predicting variables were associated with development of STR DNA profiles entered into CODIS based upon swab location?

During data analysis, we realized that some changes needed to be made to research questions 4A and 4B to more accurately represent useful findings. The outcome variable for SAKs was changed to development of uploaded CODIS profiles. The predicting features were assault and patient/victim variables. The outcome variable for swabs was changed to the development of full or partial STR DNA profile of foreign contributor(s) with the predicting features of assault and patient/victim variables. The revised 4A and 4B questions are as follows:

4A. What victim and sexual assault (SA) variables were statistically significant in predicting an uploaded CODIS (SDIS) profile?

4B. What victim and SA variables were statistically significant in predicting the development of full or partial STR DNA profiles of foreign contributors based upon swab location?

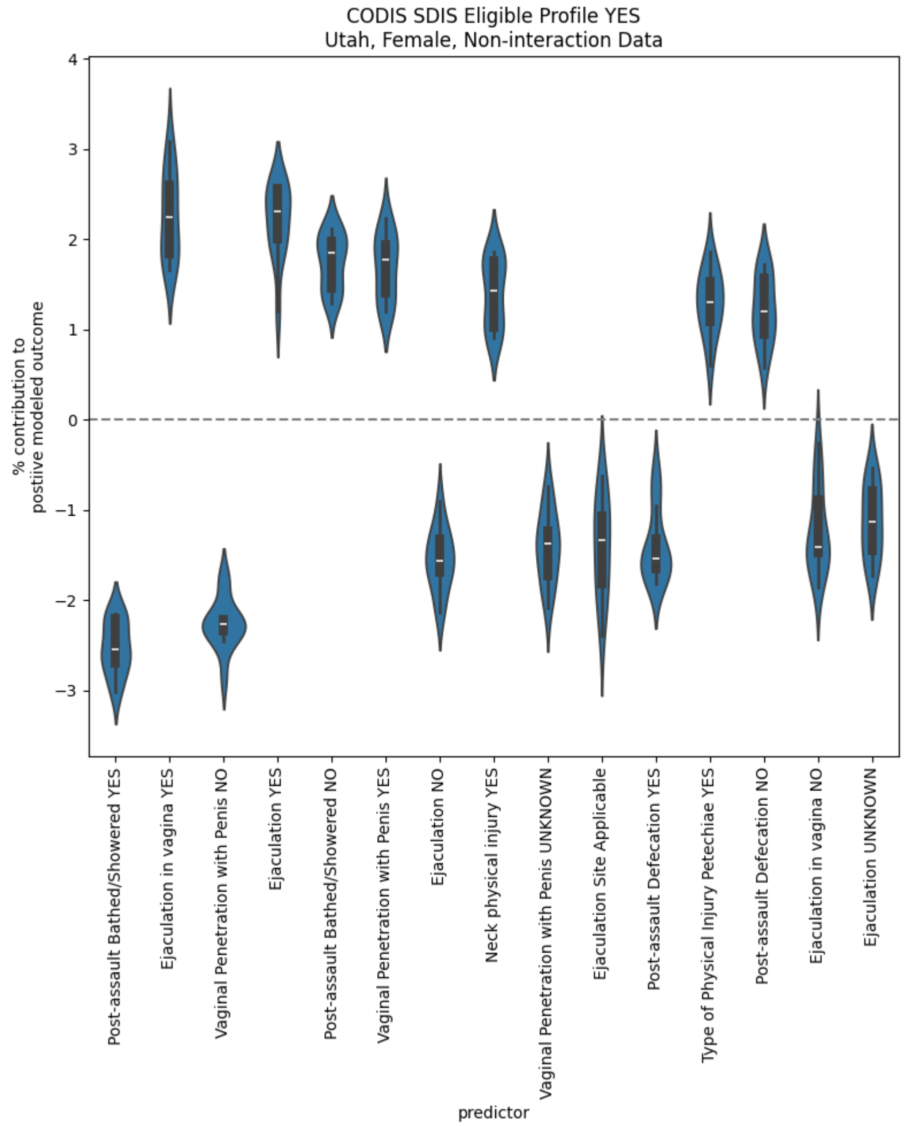
To answer these questions, we utilized logistic regression modeling as a form of machine learning and described in the previous section on methodology. As we had different data points on assault and patient/victim variables, we ran separate models for each crime lab (UBFS, OCCL, and ISP). To aid in interpretation, we trained models on both normalized and non-normalized data. Specifically, for the normalized data, we used min/max normalization on continuous variables so that the smallest value was zero and the largest value was 1. The reason for doing this was so that, for the normalized data, coefficients could be directly compared to determine the relative importance of features for determining model outcome, with a higher magnitude coefficient indicating that its associated feature contributed more than a feature associated with a lower-magnitude coefficient. Models trained on non-normalized data were important for an alternate interpretation of model coefficients, namely the “change in log odds ratio” per one-unit change in a particular feature variable, with their exponentiated value indicating the “change in odds ratio” (shown in the figures). For example, if the exponentiated coefficient on “patient age” were 0.5, then, all other factors being held equal, a 1:1 odds of developing an CODIS-eligible profile would result in a 0.5:1, or 1:2 odds of developing a CODIS-eligible if the patient were one year older, i.e. the probability of a CODIS-eligible profile would decrease. Not normalizing “patient age” in this case is key to maintaining interpretability of these coefficients but obfuscates the comparison of “patient age” to other features with different variances, hence the need for both analyses. In both the normalized and non-normalized data, categorical variables were split into separate columns with a 1 indicating “yes” for a particular value of a category, and a 0 indicating “no” for a particular value of a category. In the machine learning community this is often referred to as “one-hot encoding,” and is essential for applying machine learning techniques that rely on the topological structure of the real numbers, to categorical variables, which lack this topological structure. Because of stochasticity in the machine learning process, we trained 12 models for each outcome variable, each with different random initial conditions, to elucidate the consistency of model results. Violin plots show the density of the distributions of a particular value across these multiple models, with the interpretation of the value indicated on each y-axis. The box-and-whisker plot within the violin plots represent standard data quartiles in a traditional box-and-whisker plot.

The findings are presented by site (Utah/UBFS, Orange County/OCCL, and Idaho/ISPFS) with female findings first followed by male findings (Utah/UBFS only) for each question. The following figures represent the findings for **research question 4A**.

Utah/UBFS Data on Females:

For the first figure representation of similar models, an interpretation of the model is presented. The remaining similar models do not contain the text interpretation. A summary of the key findings across sites is presented after Figures 1-12.

Figure 1: Utah Female Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile



Below are the coefficients within Figure 1 and their percent contribution to the model decision-making process. If the coefficient is above the "0" line, then it is correlated with a positive contribution. If the coefficient is below the "0" line, then it is correlated with a negative contribution.

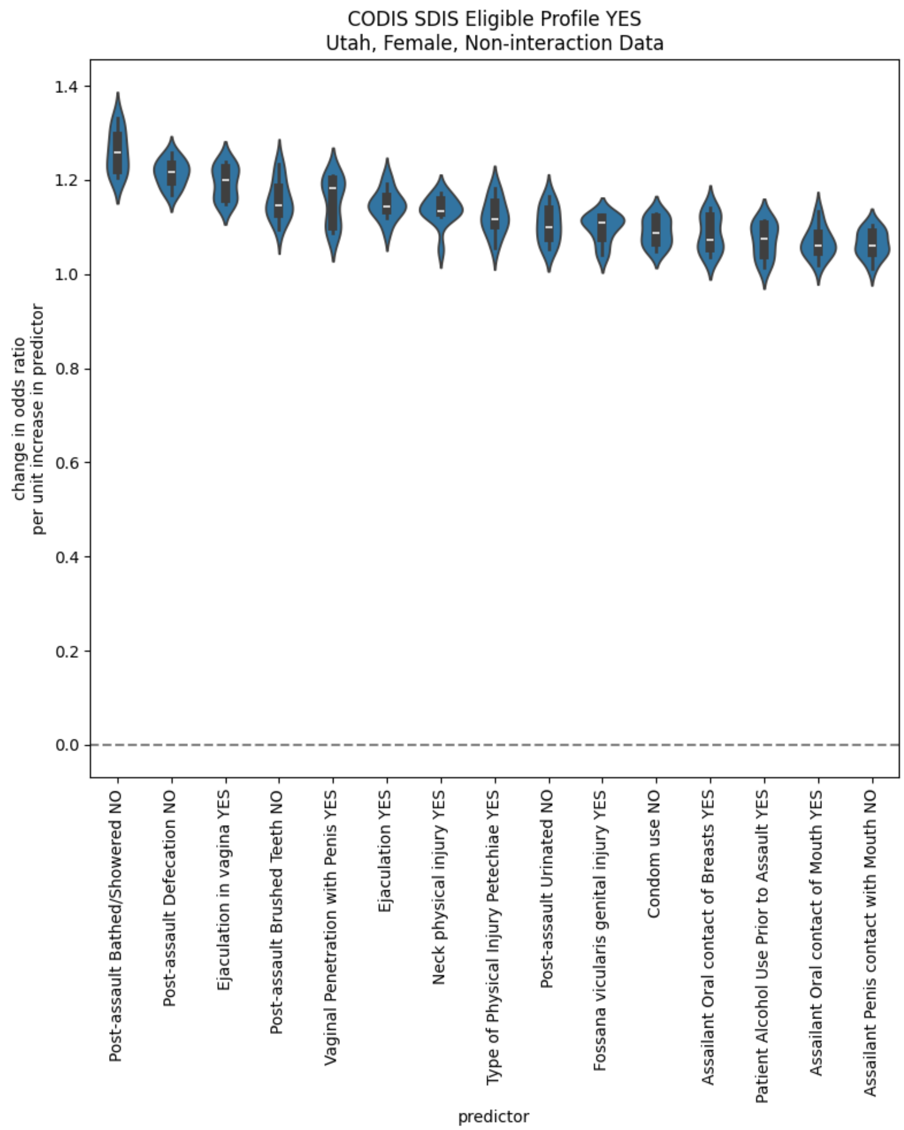
- Post-assault bathed/showered YES contributes ~2.5% of the model decision-making process, and correlated with a *negative* outcome.

- Ejaculation in vagina YES contributes ~2.2% of the model decision-making process, and correlated with a *positive* outcome.
- No vaginal penetration with penis contributes ~2.4% of the model decision-making process and correlated with a *negative* outcome.
- Ejaculation reported as YES contributes ~1.8% of the model decision-making process and correlated with a *positive* outcome.
- Patient did not bathe/shower post-assault contributes ~1.7% of the model decision-making process, and correlated with a *positive* outcome.
- Vaginal penetration with penis contributes ~1.6% of the model decision-making process and correlated with a *negative* outcome.
- Ejaculation reported as NO contributes ~1.5% of the model decision-making process and correlated with a *negative* outcome.
- Physical injury to neck contributes ~1.4% of the model decision-making process and correlated with a *positive* outcome.
- Vaginal penetration by penis unknown contributes 1.4% of the model decision-making process, and correlated with a *negative* outcome.
- Ejaculation site unknown contributes ~1.4% of the model decision-making process, and correlated with a *negative* outcome.
- Post-assault defecation contributes ~1.6% of the model decision-making process, and correlated with a *negative* outcome.
- Petechiae noted on physical exam contributes ~1.3% of the model decision-making process, and correlated with a *positive* outcome.

- Post-assault defecation did not occur contributes ~1.2% of the model decision-making process, and correlated with a *positive* outcome.
- Ejaculation did not occur in the vagina contributes ~1.5% of the model decision-making process, and correlated with a *negative* outcome.
- Ejaculation unknown contributes ~1.3% of the model decision-making process, and correlated with a *negative* outcome.

Figure 2 represents Utah data in odds ratio plots with the outcome variable of uploaded CODIS/SDIS profile. Because logistic regression solves for coefficients that represent changes in log odds ratio, the coefficients are exponentiated so as to represent changes in odds ratio. Thus, the values of the coefficients in these plots will only be positive, and whether they increase or decrease odds of an uploaded CODIS/SDIS profile depends on whether the coefficient is above or below 1, respectively.

Figure 2. Utah Female Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile



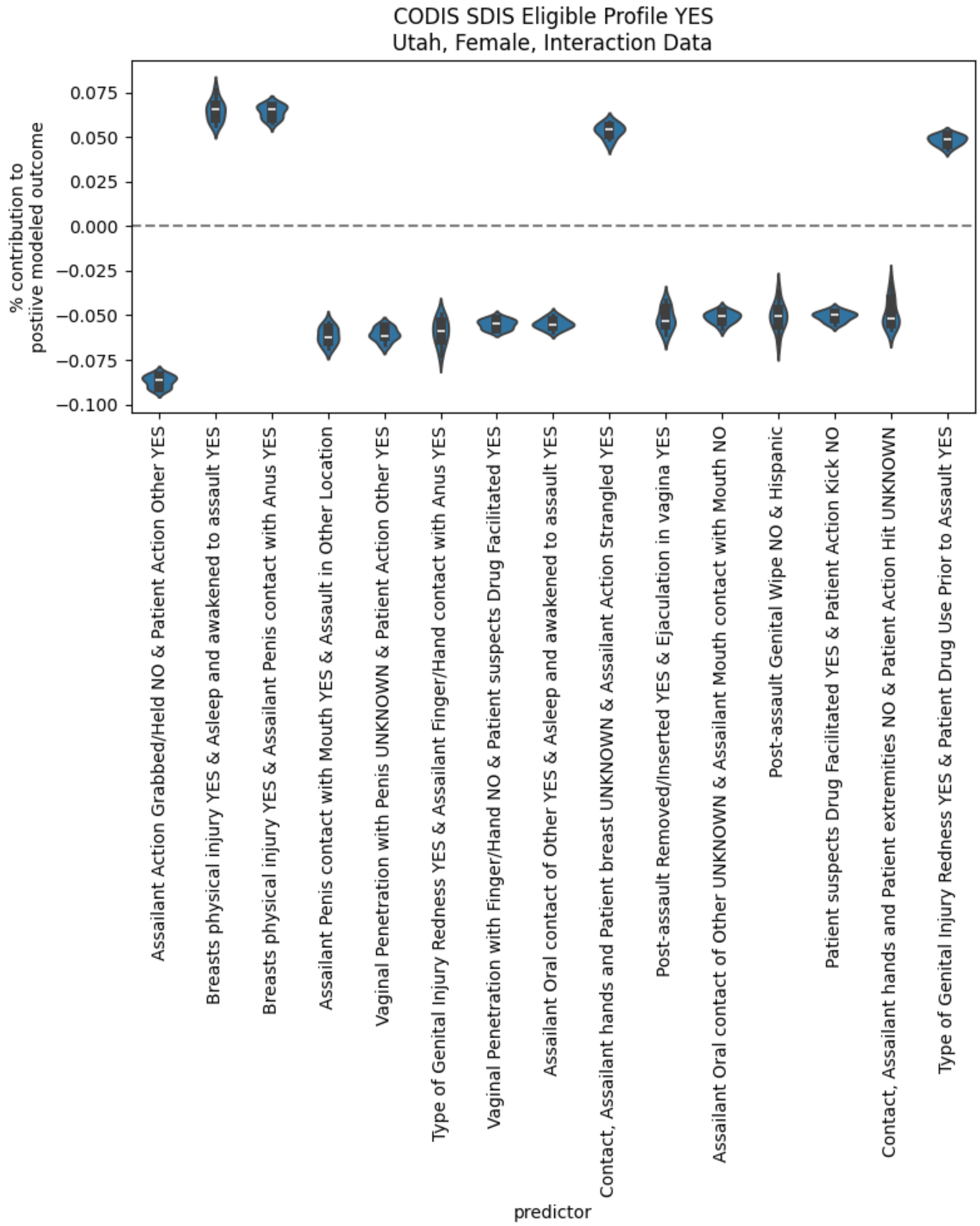
This figure represents the odds ratio of a positive outcome (development of CODIS/SDIS uploaded profile). For example, if the odds of a positive outcome to negative outcome are a:b, then, all other factors being held equal, a coefficient of c means that a one-unit change in the associated variable results in an a*c:b odds ratio. The mean value across multiple models trained using different random initializations is used to represent approximate odds ratio:

- Post-assault bathed/showered NO has an odds ratio of (a*1.25):b

- Post-assault defecation NO has an odds ratio of (a*1.2):b
- Ejaculation in vagina YES has an odds ratio of (a*1.2):b
- Post-assault brushed teeth NO has an odds ratio of (a*1.15):b
- Vaginal penetration with penis YES has an odds ratio of (a*1.2):b
- Ejaculation YES has an odds ratio of (a*1.18):b
- Neck physical injury YES has an odds ratio of (a*1.17):b
- Petechiae noted as physical injury YES has an odds ratio of (a*1.15):b
- Post-assault urination NO has an odds ratio of (a*1.14):b
- Injury on fossa navicularis has an odds ratio of (a*1.16):b
- Condom use NO has an odds ratio of (a*1.14):b
- Assailant oral contact of breasts YES has an odds ratio of (a*1.13):b
- Patient alcohol use YES prior to assault has an odds ratio of (a*1.15):b
- Assailant oral contact of mouth YES has an odds ratio of (a*1.14):b
- Assailant penis contact with mouth NO has an odds ratio of (a*1.14):b

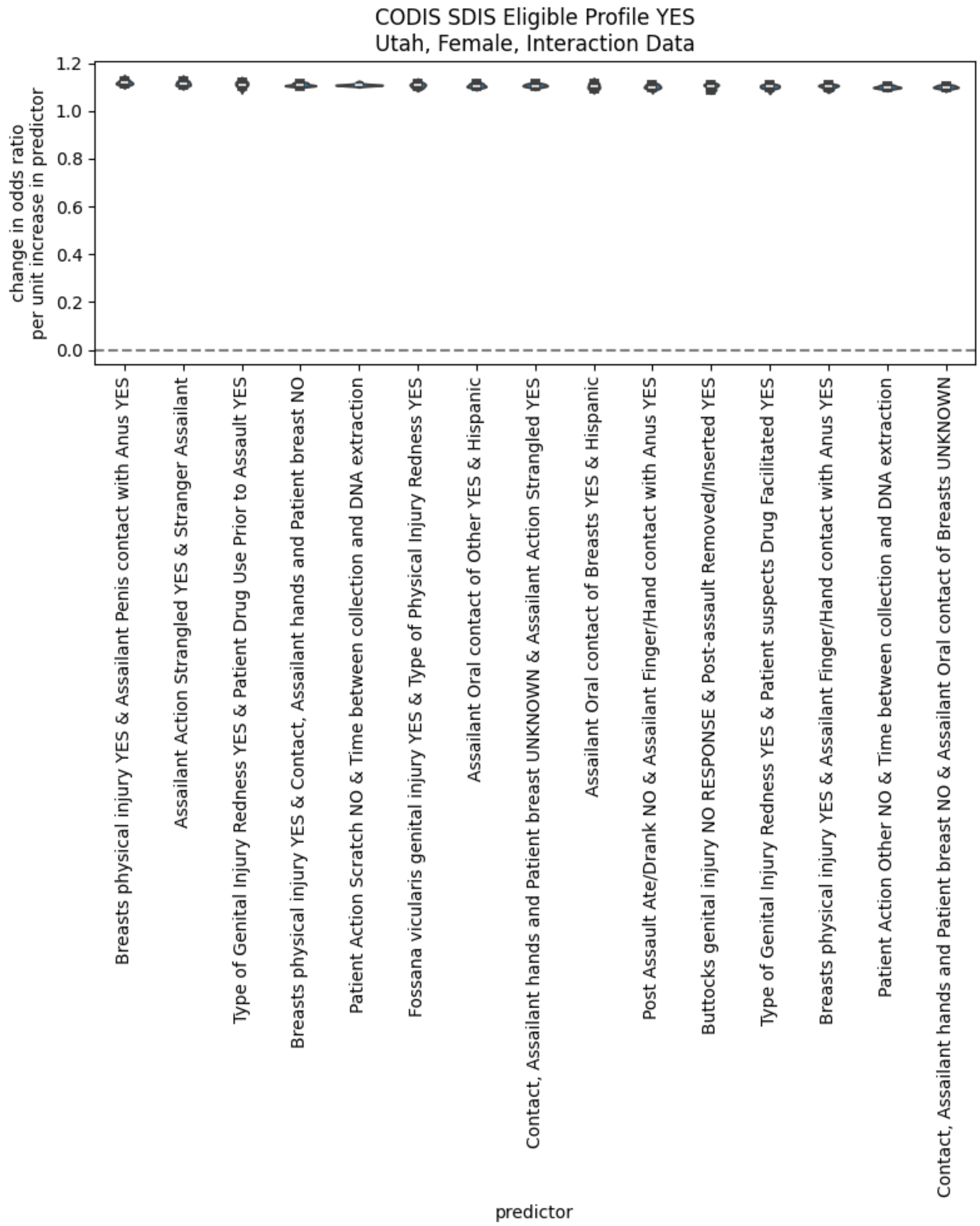
The following two Figures, 3 and 4, represent Utah female data when analyzed with interactions, meaning pair-wise multiplications of features. Several of the variables/features were found to have significant, sometimes unexpected, interactions. For these models, the data was analyzed to capture these interactions and improve model accuracy. The same interpretation approach would be implemented but looking at the variables in combination with other variables.

Figure 3: Utah Female Normalized with Interactions Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile



An example of the interpretation for Figure 3 would be: the coefficient of assailant action of grabbing or holding patient with the interaction of patient/victim other action (usually pushing or shoving assailant) contributes .085% of the model decision-making process. Note that given the large number of columns available when considering all pair-wise interactions of variables, the model decision making becomes spread across many features.

Figure 4. Utah Female Not Normalized, Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile



An example of interpretation for Figure 4 would be injury found on breasts and assailant penis contact with anus has an odds ratio of (a*1.16):b. Interestingly, in this model the interaction coefficients have approximately the same odds ratio. This indicates that the models' decision making was spread across a plurality of features, with no single feature dominating the model decision making process.

Orange County/OCCL Data on Females

Figure 5: Orange County Female Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile

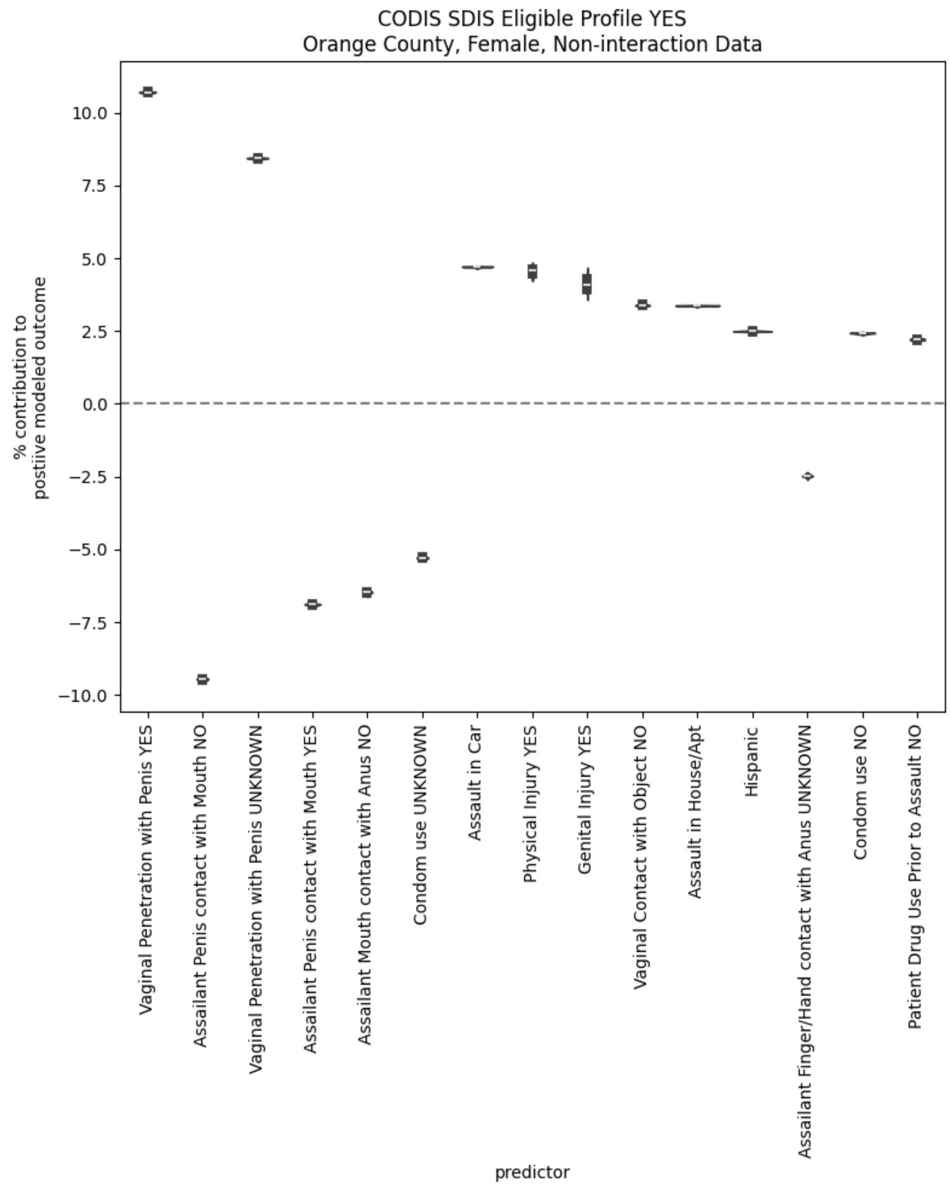


Figure 6. Orange County Female Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile

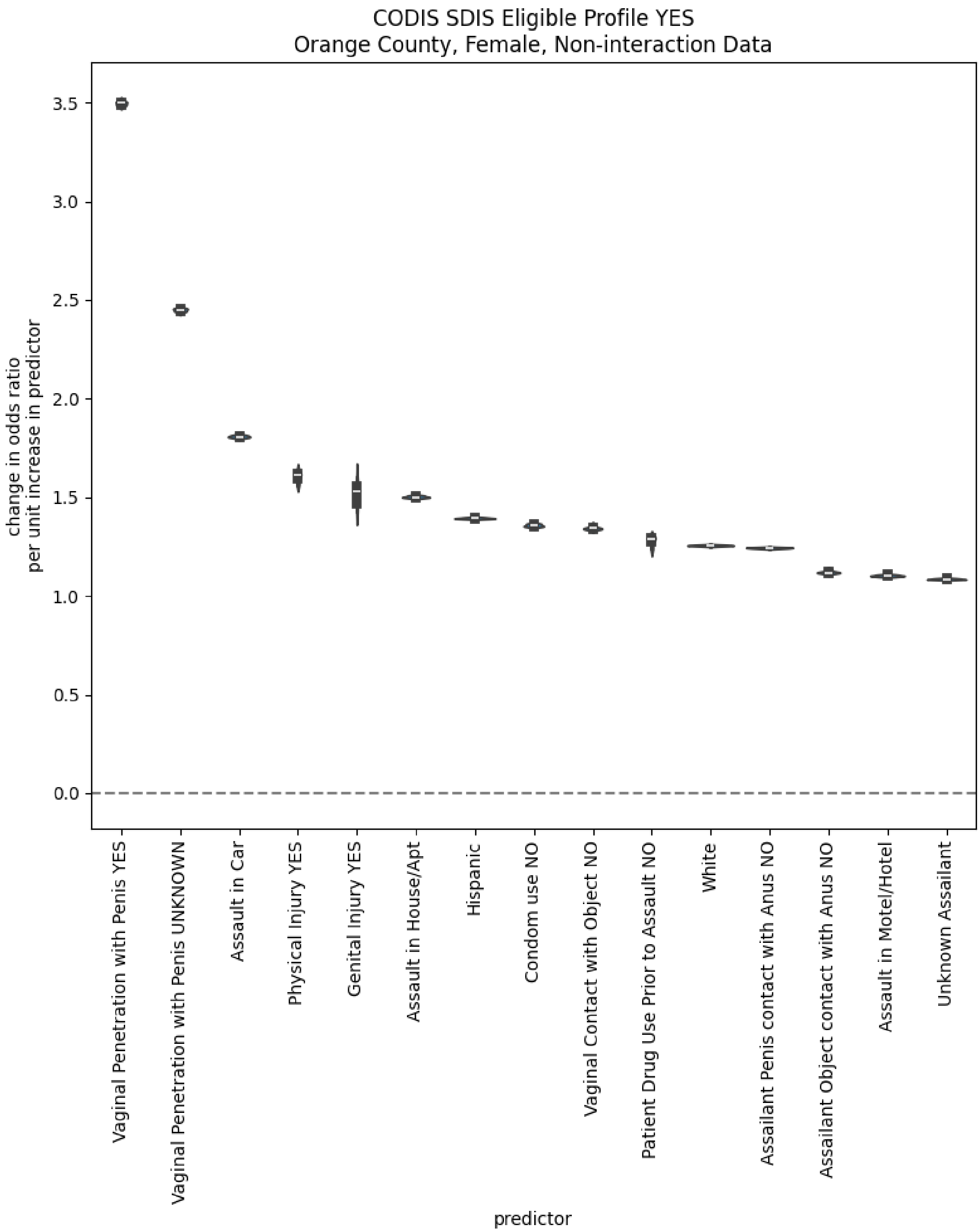


Figure 7: Orange County Female Normalized with Interactions Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile

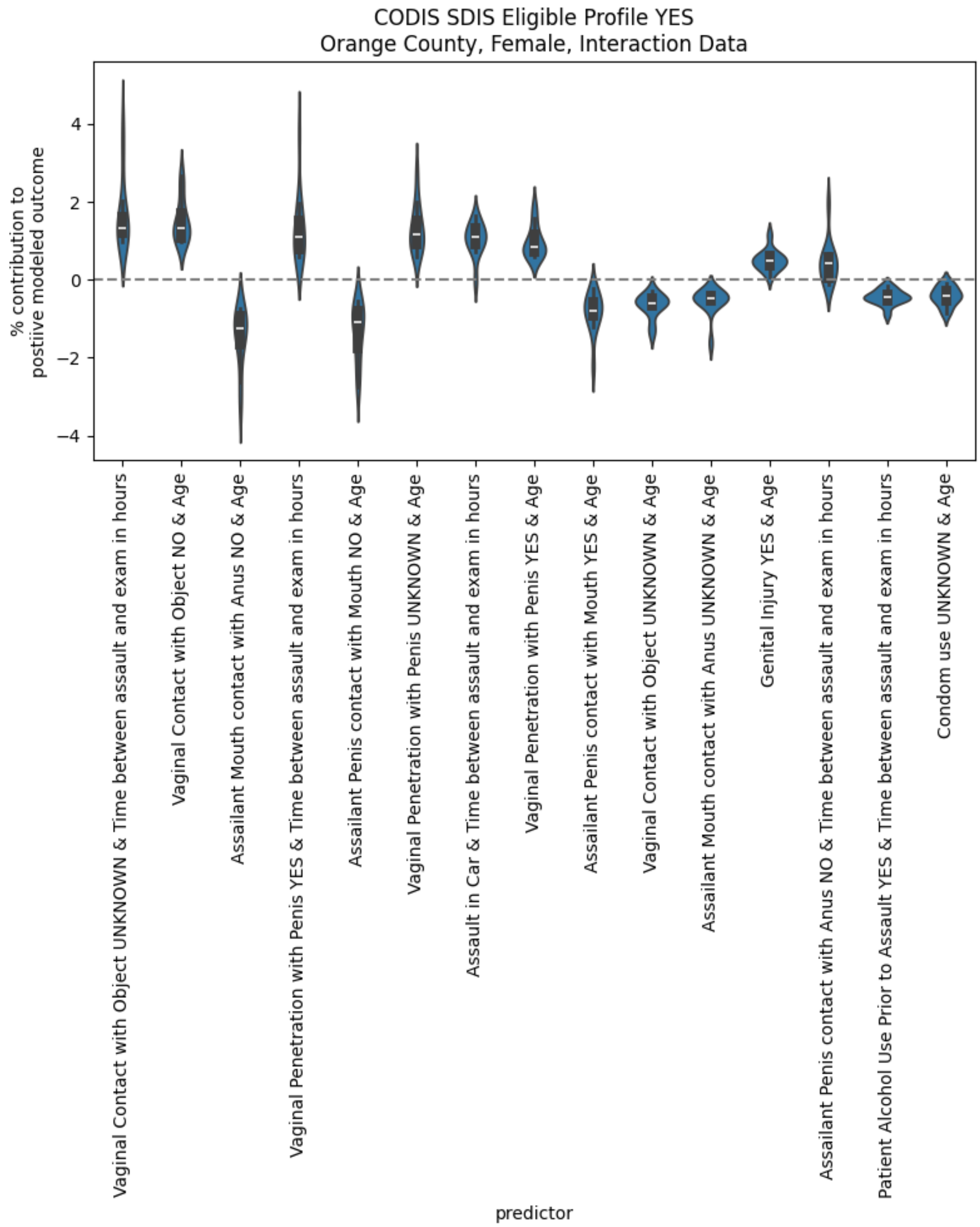
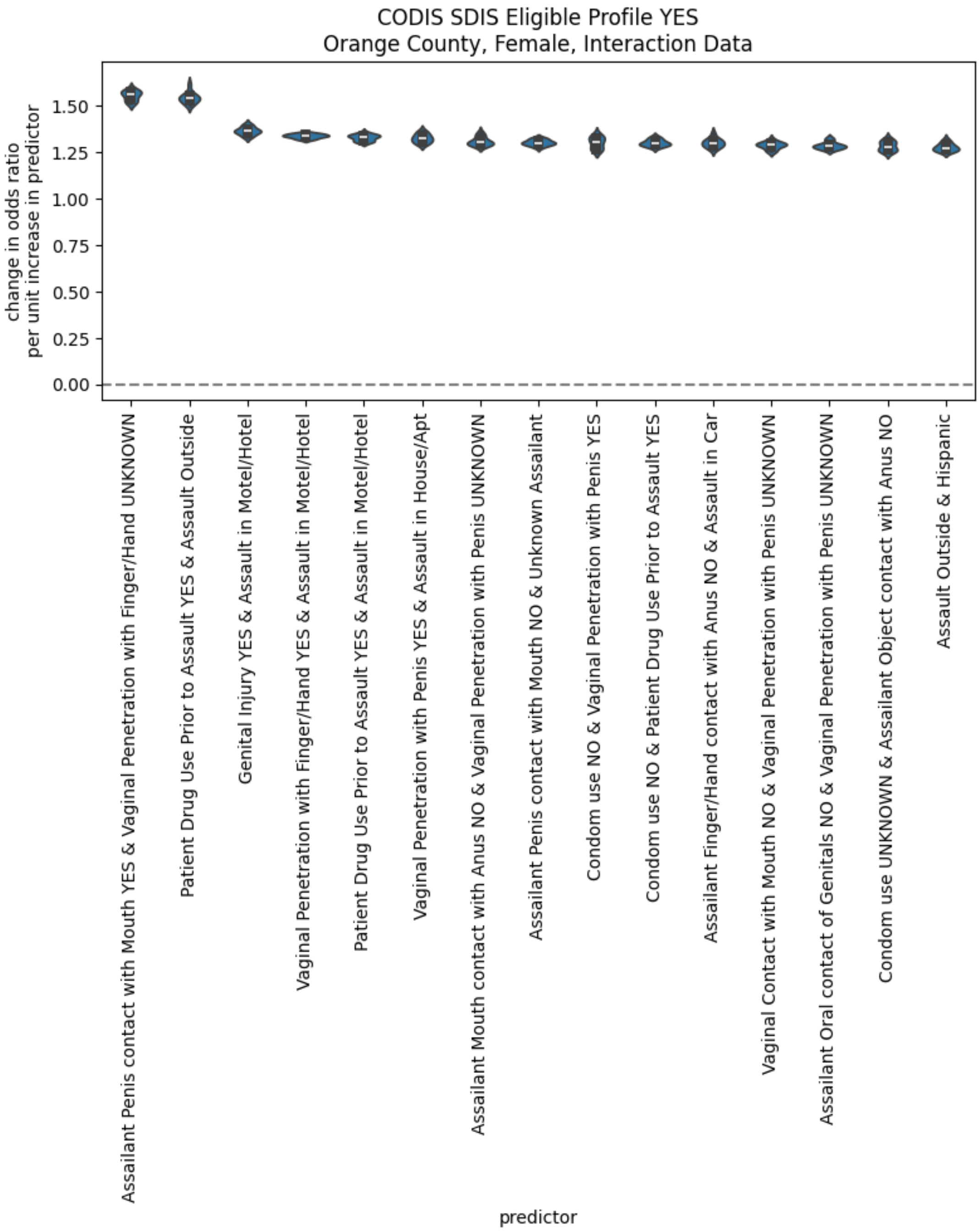


Figure 8. Orange County Female Not Normalized, Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile



Idaho/ISP Data on Females

Figure 9: Idaho Female Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile

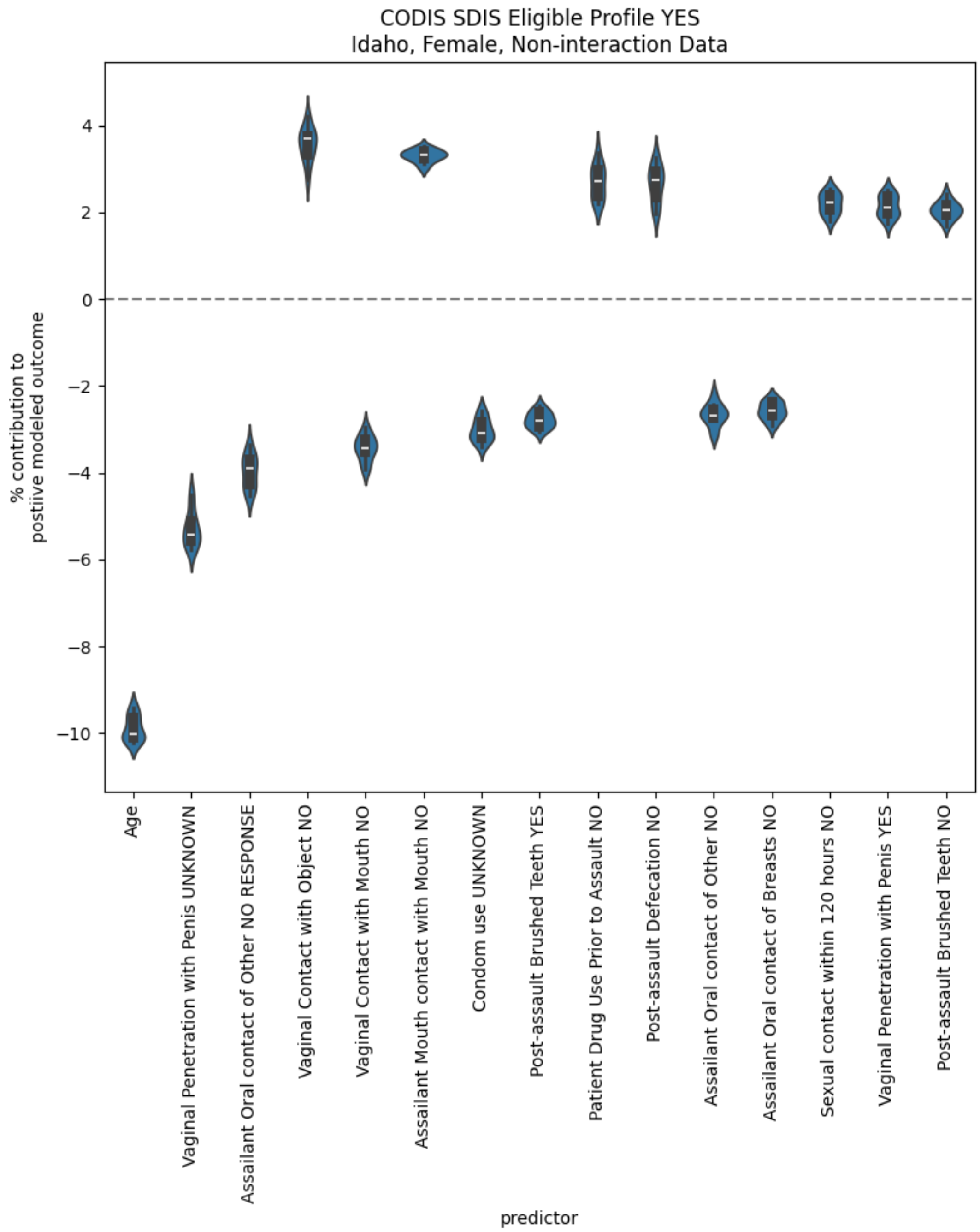


Figure 10. Idaho Female Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile

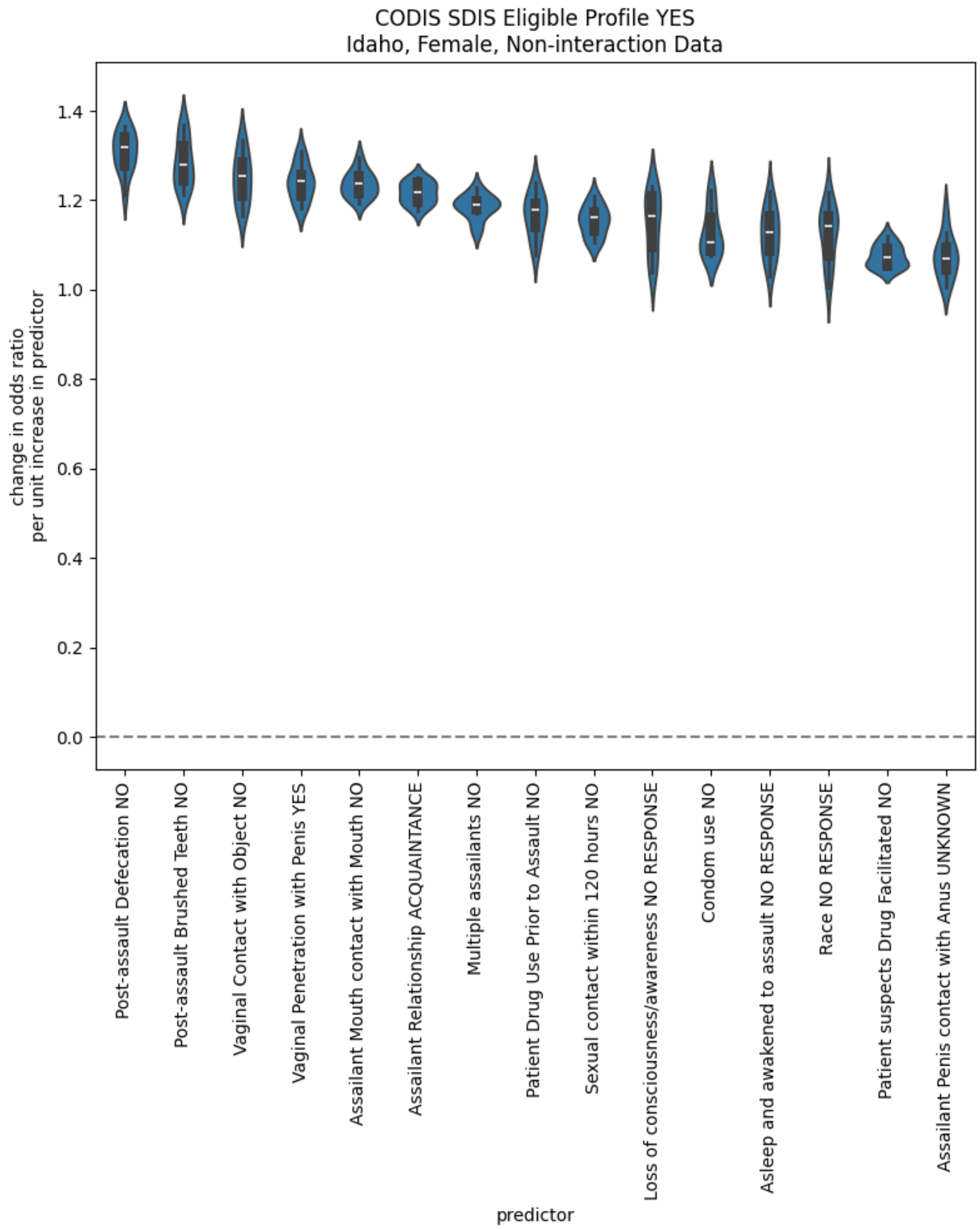


Figure 11: Idaho Female Normalized with Interactions Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile

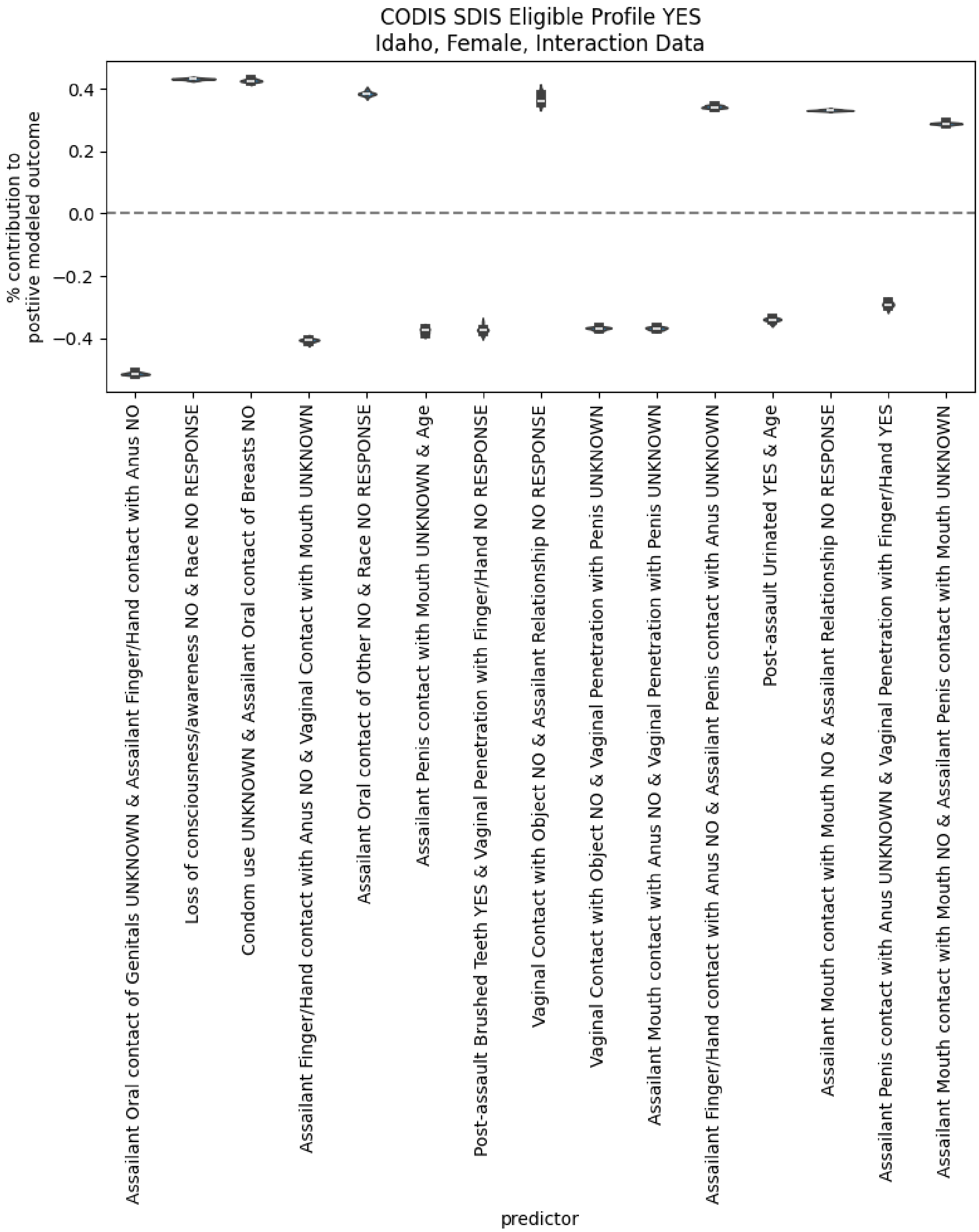
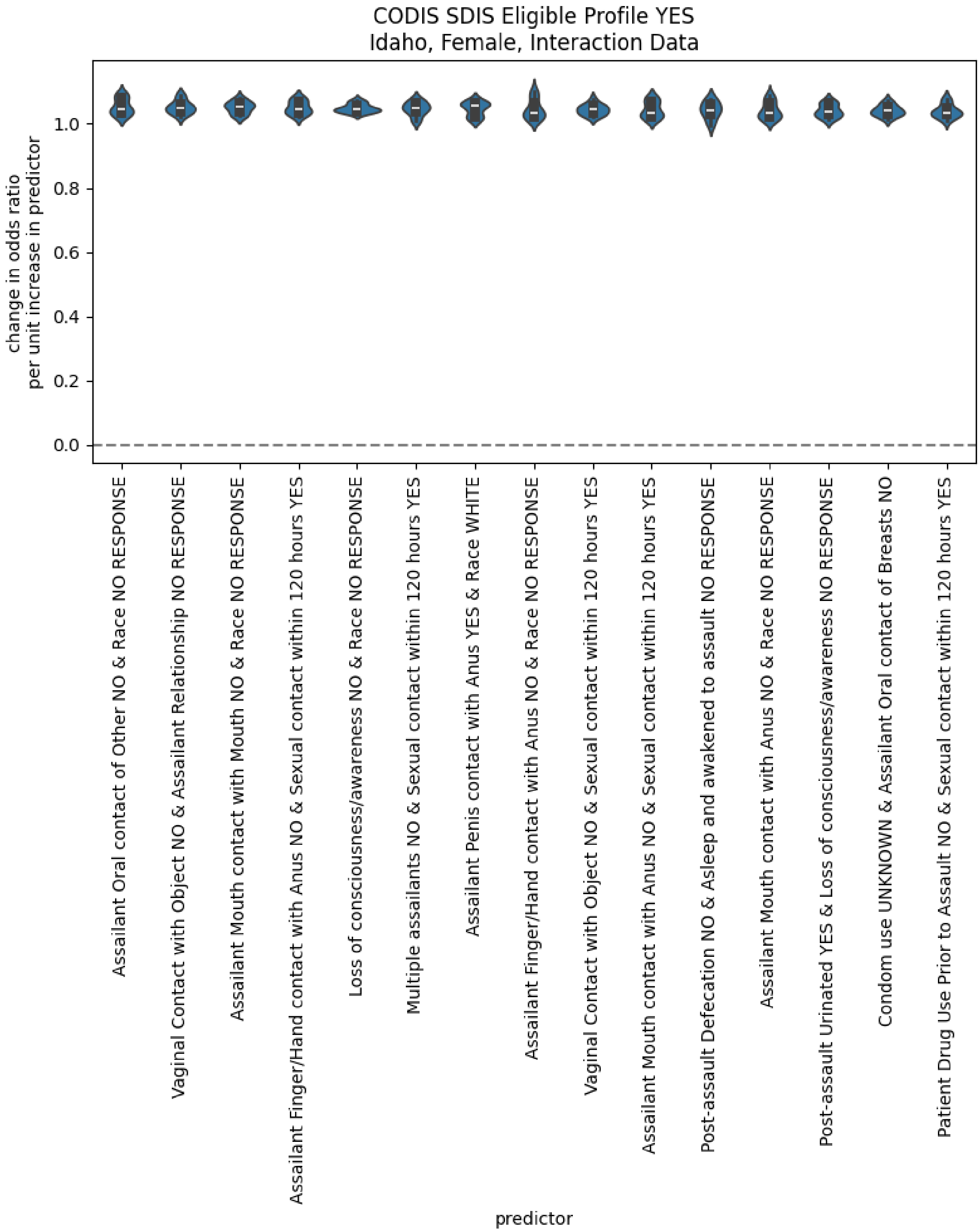


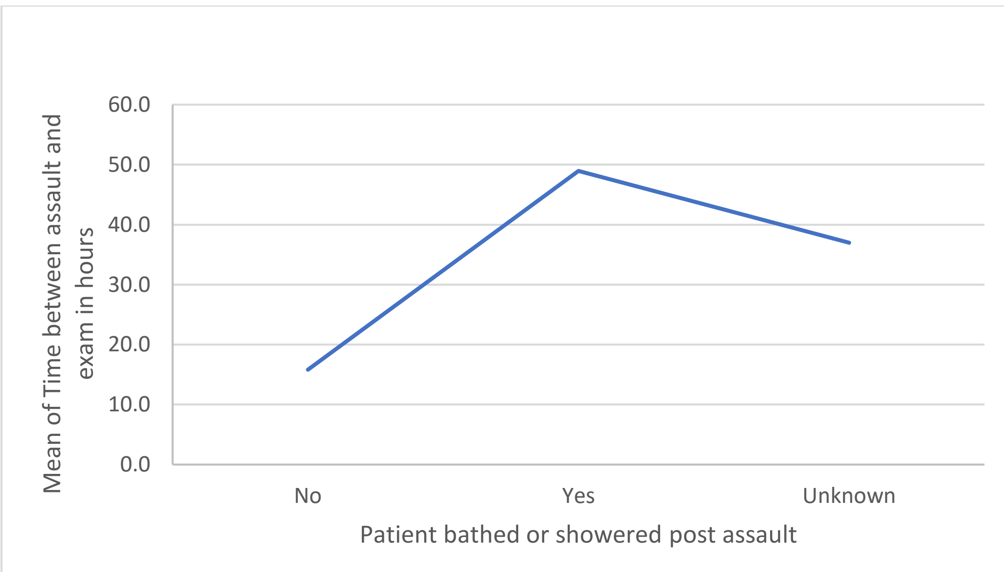
Figure 12. Idaho Female Not Normalized, Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile



<u>Summary of Key Findings on Female Models on Development of CODIS/SDIS Uploaded Profiles:</u>

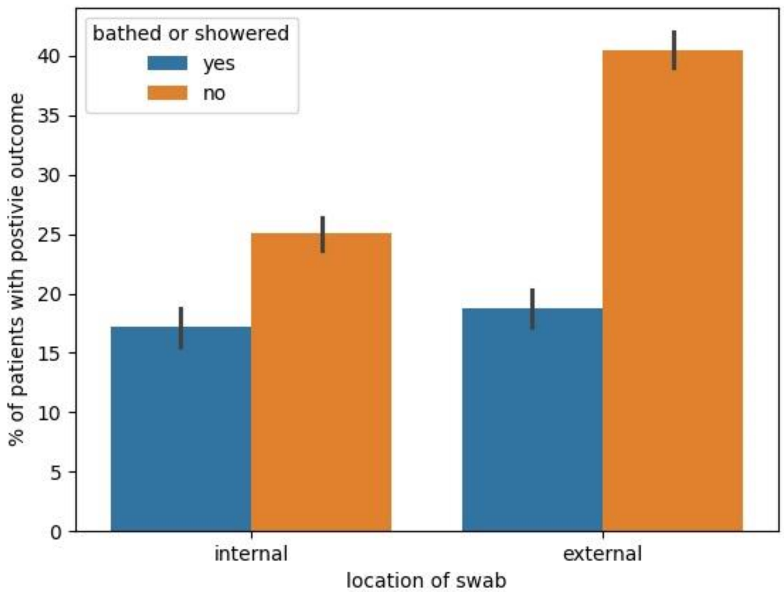
The variable of victim/patient bathing or showering post-assault was evident in all of the models. We explored the relationship between bathing/showering with time between assault and examination as we theorized that the longer between assault and SAMFE, the increased likelihood of bathing/showering. One-way ANOVA calculation was completed on bathing/showering and time between assault and SAMFE and found to be highly significant [$F(2,8772) = 971.398$, $p < .001$].

Figure 13. Time between Assault and Examination and Patient Bathed/Showered Post-Assault



We also evaluated the impact of bathing/showering on development of full or partial STR DNA profiles of foreign contributors on internal and external swabs (Figure 14). Bathing or showering decreases the likelihood of developing full or partial foreign contributors' profiles from external swabs substantially more than from internal swabs. A key take-away is that regardless of bathing/showering status, full or partial STR DNA profiles of foreign contributors can be developed. We found in the Utah data that 25% of CODIS uploaded profiles were obtained from patients who reported bathing or showering post-assault.

Figure 14. Impact of Bathing/Showering on Internal and External Swabs



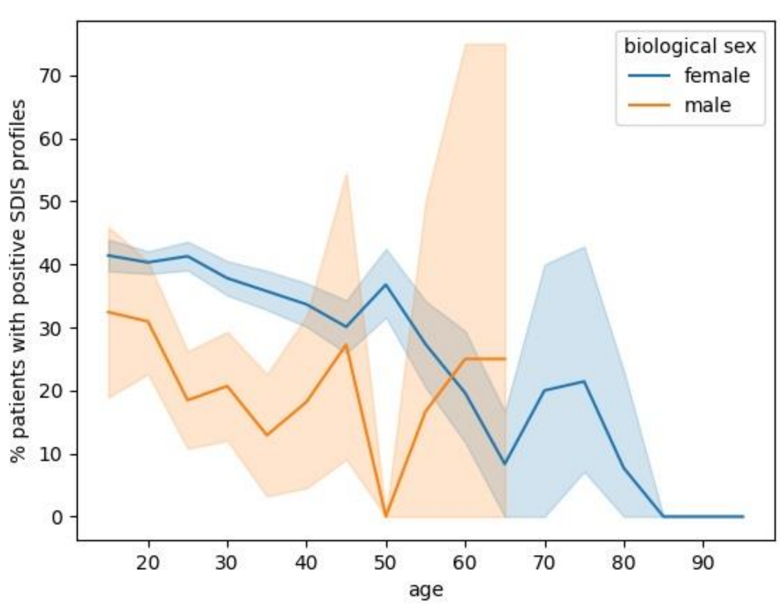
The variable of ejaculation was also prevalent in all of the models as being significant in predicting development of uploaded CODIS profile of foreign contributor. Notably, the most common victim response to the question if ejaculation occurred was “unknown” (52% UBFS, 57.8% OCCL, and 51.5% ISPFS) which could be interpreted to support the encounter as a non-consensual sex act. Ejaculation site, particularly if known by patient to be in vagina, was also significant in models. Penile penetration in vagina, with or without known ejaculation, was significant.

Oral contact by assailant on victims’ bodies including breasts, neck, mouth, and other body parts was highly significant across models. This is supported by the high number of external swabs from breasts and neck that resulted in full or partial STR DNA profiles of foreign contributor(s).

Several variables with the response of “unknown” were found to be significant in the models including ejaculation, vaginal penetration, condom use, and hand or oral contact on body sites. Almost half of the victims reported some degree of loss of consciousness or awareness during the sexual assault (UBFS 47.8%, OCCL 46.1%, and ISPFS 47.4%). If victims are unable to answer questions regarding what happened during the assault and what portions of their bodies were touched, the SANE would not have as much information to guide evidence collection.

Victim's age was noted in some models. Additional analysis on age found a significant association between age and development of uploaded CODIS profiles. As females age, the development of uploaded CODIS profiles dramatically decreases especially after the age of 50 years. When women reach menopause age, the estrogen levels decrease resulting in changes to anogenital tissues and decreased secretions. We theorize that these changes result in a decreased ability of tissues/secretions to maintain foreign contributor's cells or DNA. The association in male patients of age and development of uploaded CODIS DNA profile decreased with a low point at age 50 years.

Figure 15. Age and Development of Uploaded CODIS DNA Profiles



Condom use was noted as being significant in some models. Overall, condom use by assailants was low (5.2%-7.1%) which also supports non-consensual sexual activity. In consensual sex, most partners will discuss STI and pregnancy prevention. A national poll of university students found that approximately 40% used condoms when engaging in consensual vaginal-penile intercourse (American College Health Association, 2022). While condom use was significant in some models in decreasing the odds of developing uploaded CODIS SDIS profiles, many cases with condom use still resulted in uploaded CODIS profiles. In evaluating data from Utah, we found that 31.1% of cases in which the victim reported condom use during the assault (n= 641) developed uploaded CODIS SDIS profiles.

Findings from Male Victims

We explored "Question 4A. What victim and sexual assault (SA) variables were statistically significant in predicting an uploaded CODIS (SDIS) profile?" on data from male victims from Utah (n=430). We did not complete logistic regression analyses on male victims in the Orange County (n=48) and Idaho data (n=48) due to the low case numbers.

Figure 16. Utah Male Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile

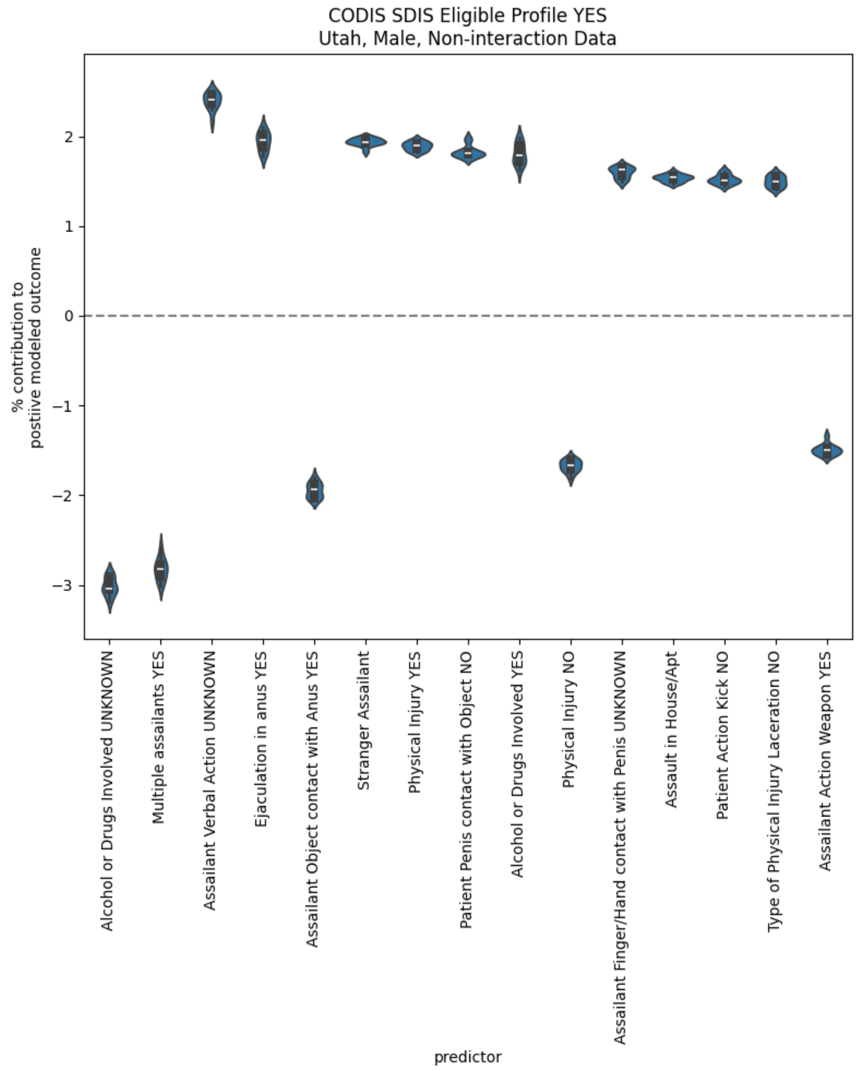


Figure 17. Utah Male Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile

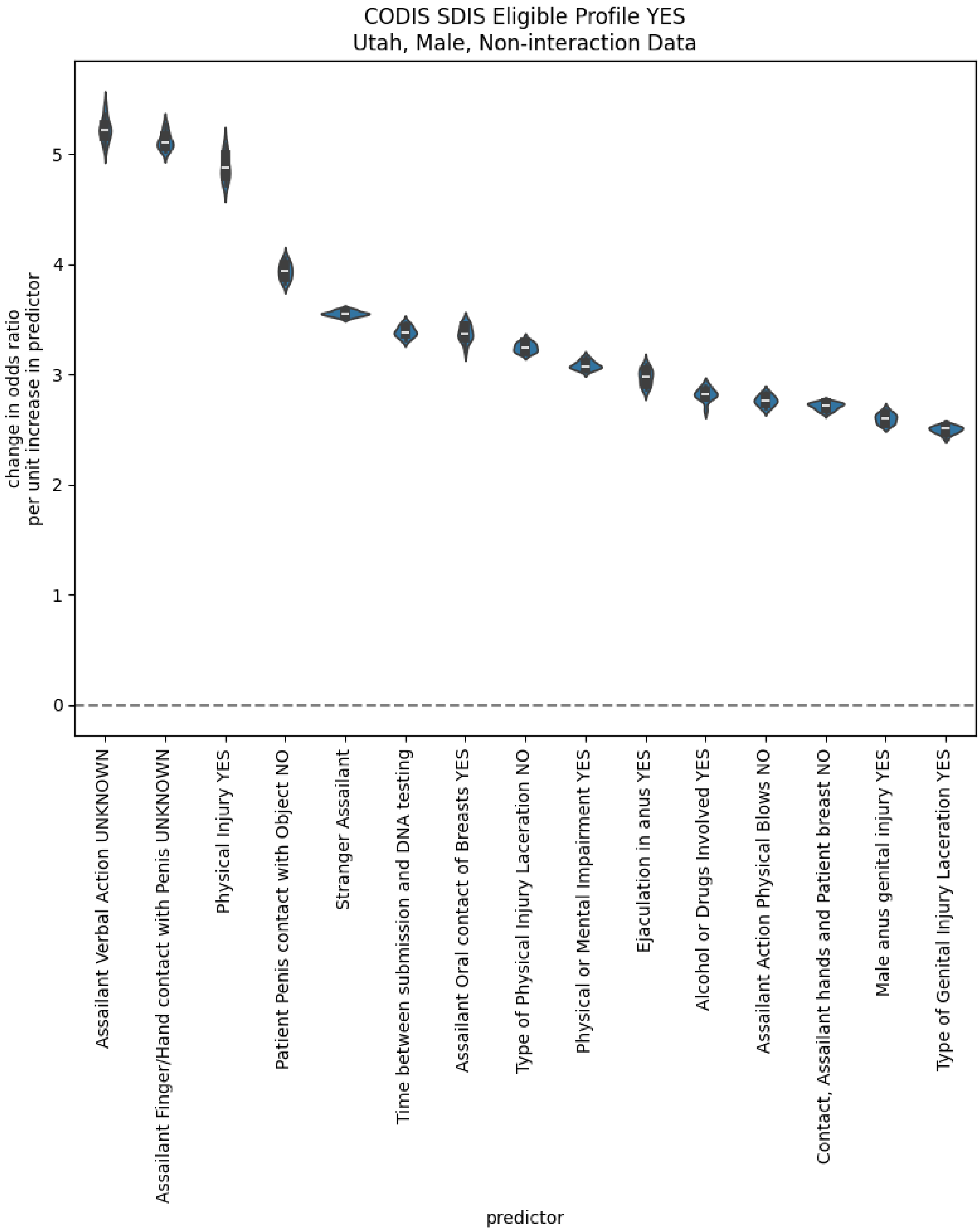


Figure 18. Utah Male Normalized with Interactions Percent Contribution to the Model Decision-Making of Development of Uploaded CODIS/SDIS Profile

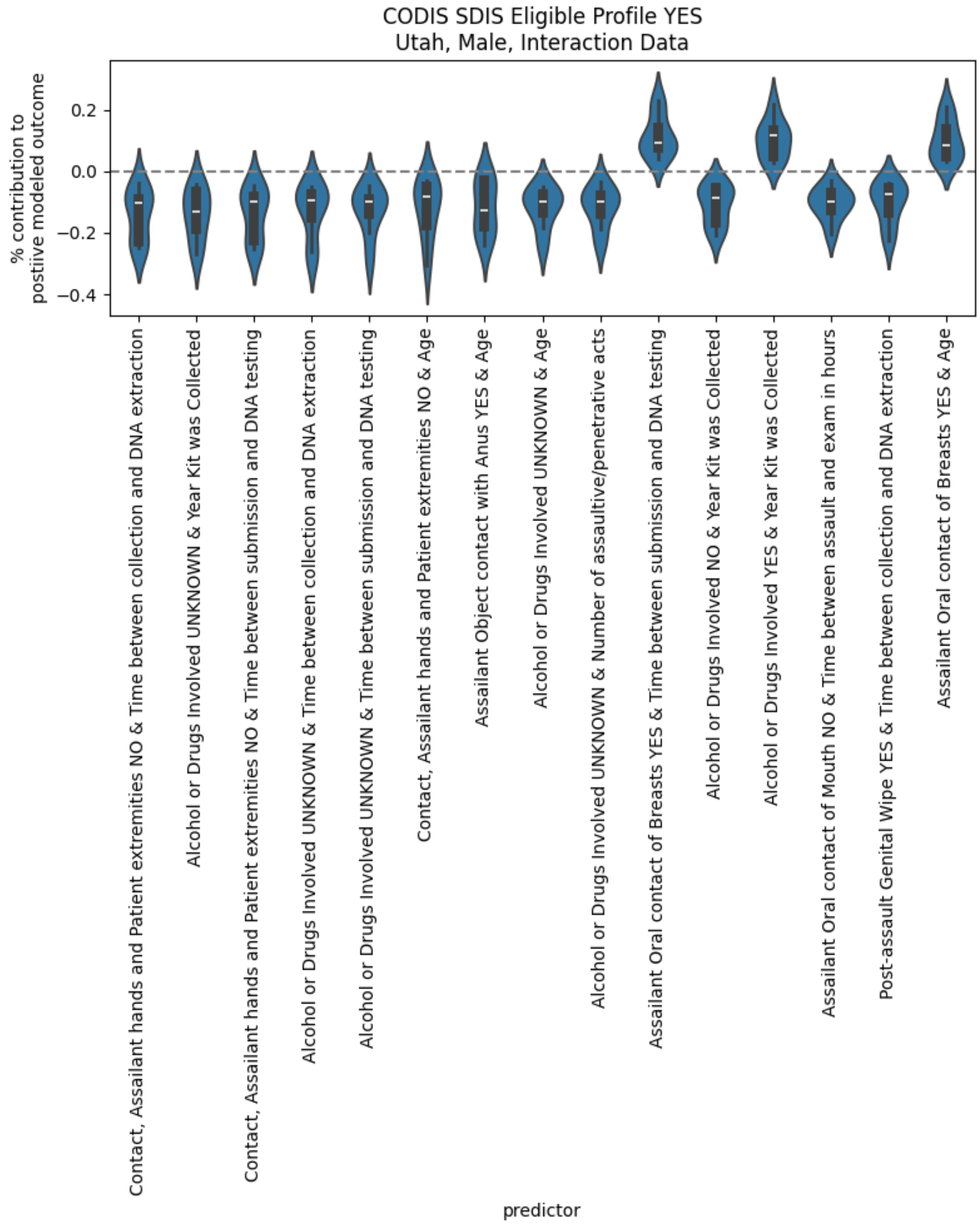
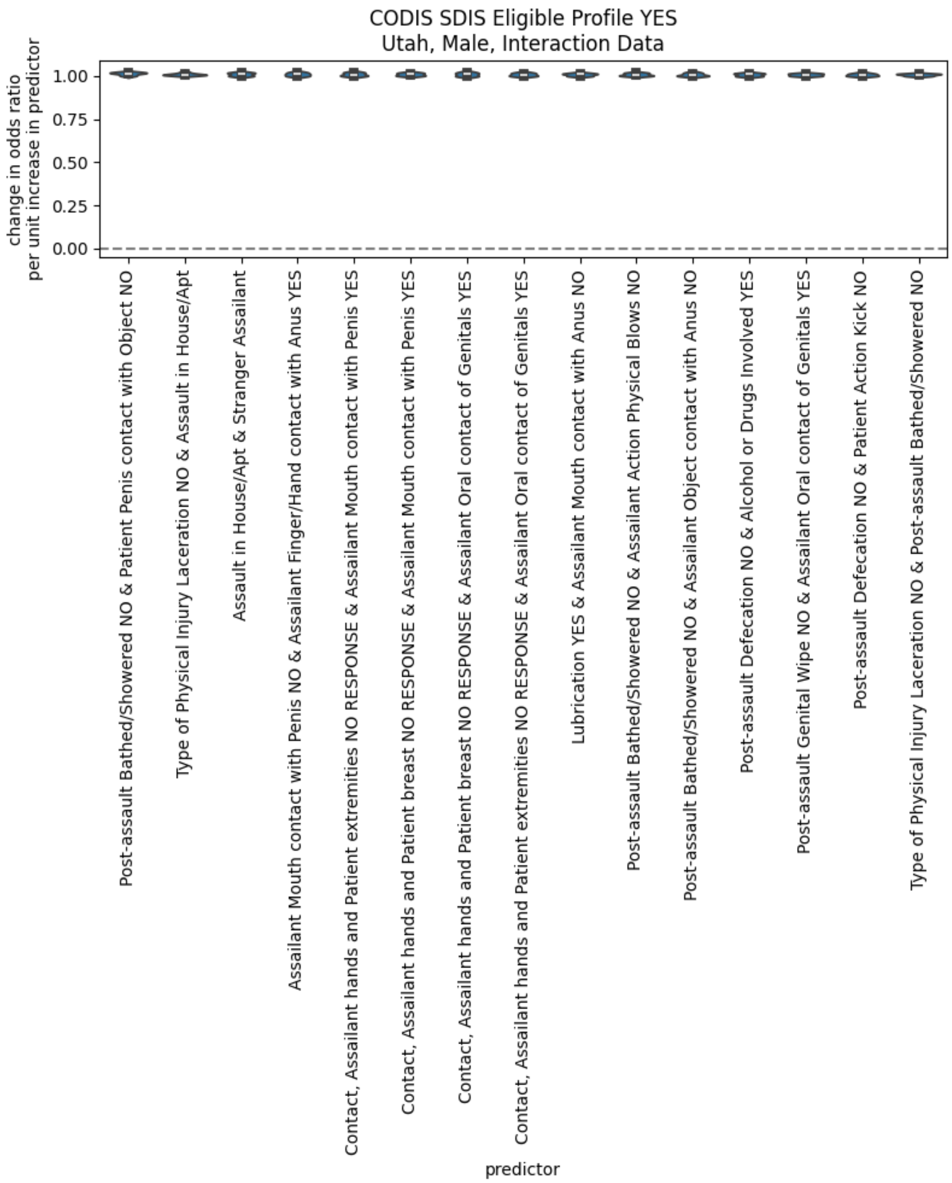


Figure 19. Utah Male Not Normalized, Interaction Change in Odds Ratio of Predictors for Uploaded CODIS/SDIS Profile



<u>Summary of Key Findings on Male Models on Development of CODIS/SDIS Uploaded Profiles:</u>

Several variables found to be significant in predicting development of CODIS/SDIS uploaded profiles in female victims were also significant in male victims: known ejaculation, oral contact of body parts, anogenital injury, and penetration of body orifice (anus). Statistically significant variables in the male patients on the development of uploaded CODIS profiles included multiple assailants, stranger assailant, and alcohol or drug use. Further exploration into findings from SAKs from male victims will be reported in upcoming publication (See Products).

Findings on Question 4B

Question 4B "What victim and SA variables were statistically significant in predicting the development of full or partial STR DNA profiles of foreign contributors based upon swab location?" was explored in the Utah/UBFS female and male data. This question was not explored in the Orange County/OCCL and Idaho/ISPFS data as missing many data points related to victim and SA variables. The same interpretation methods apply for these models with the outcome variable of development of full or partial STR DNA profiles of foreign contributors based upon body swab location. The findings and figures are presented in order of internal swabs, female and male, and then external swabs, female and male. A short summary of key findings is provided for each swab site, female and male.

Vaginal Swabs (n=3273):

Figure 20. Vaginal Swab Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

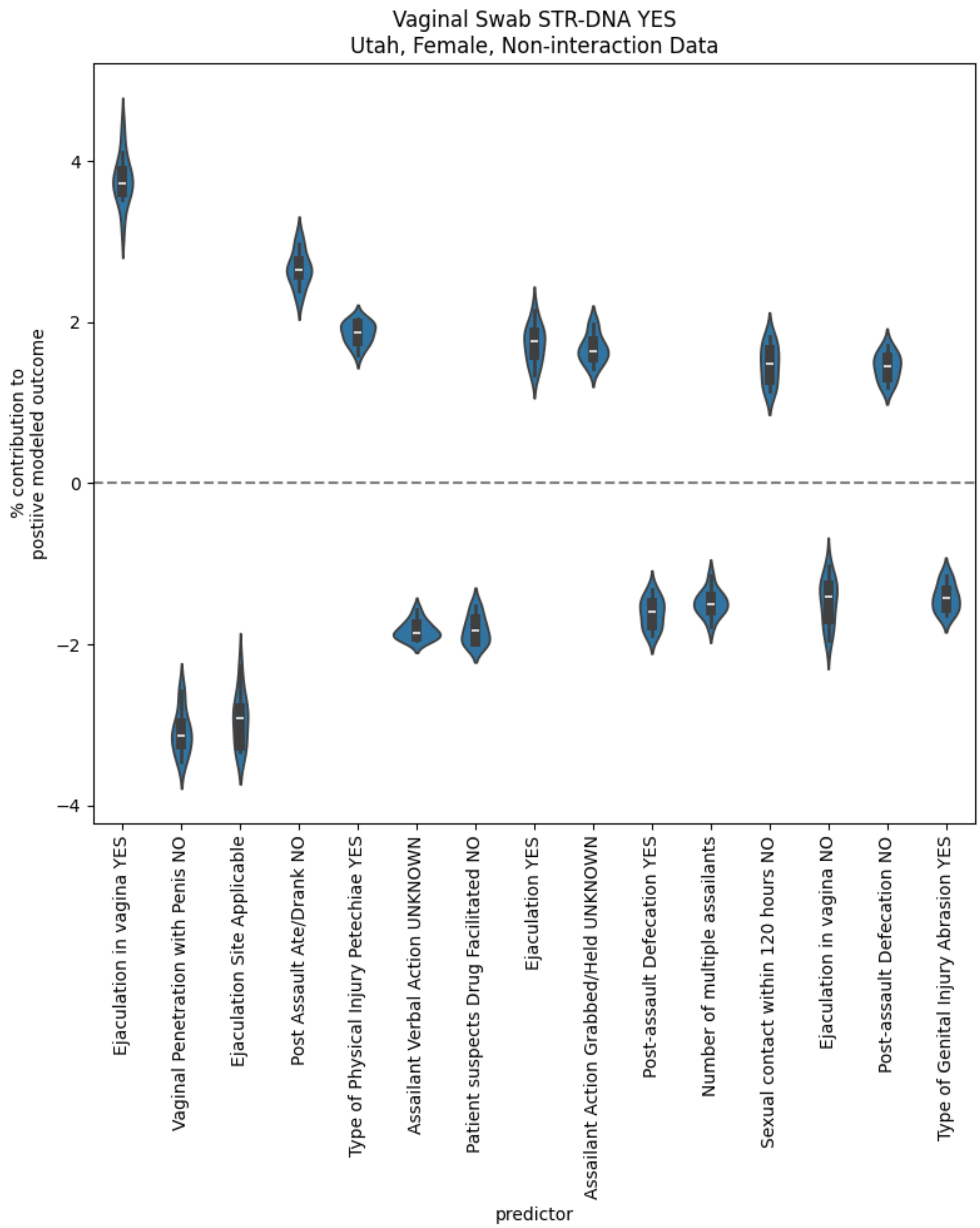


Figure 21. Vaginal Swab Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

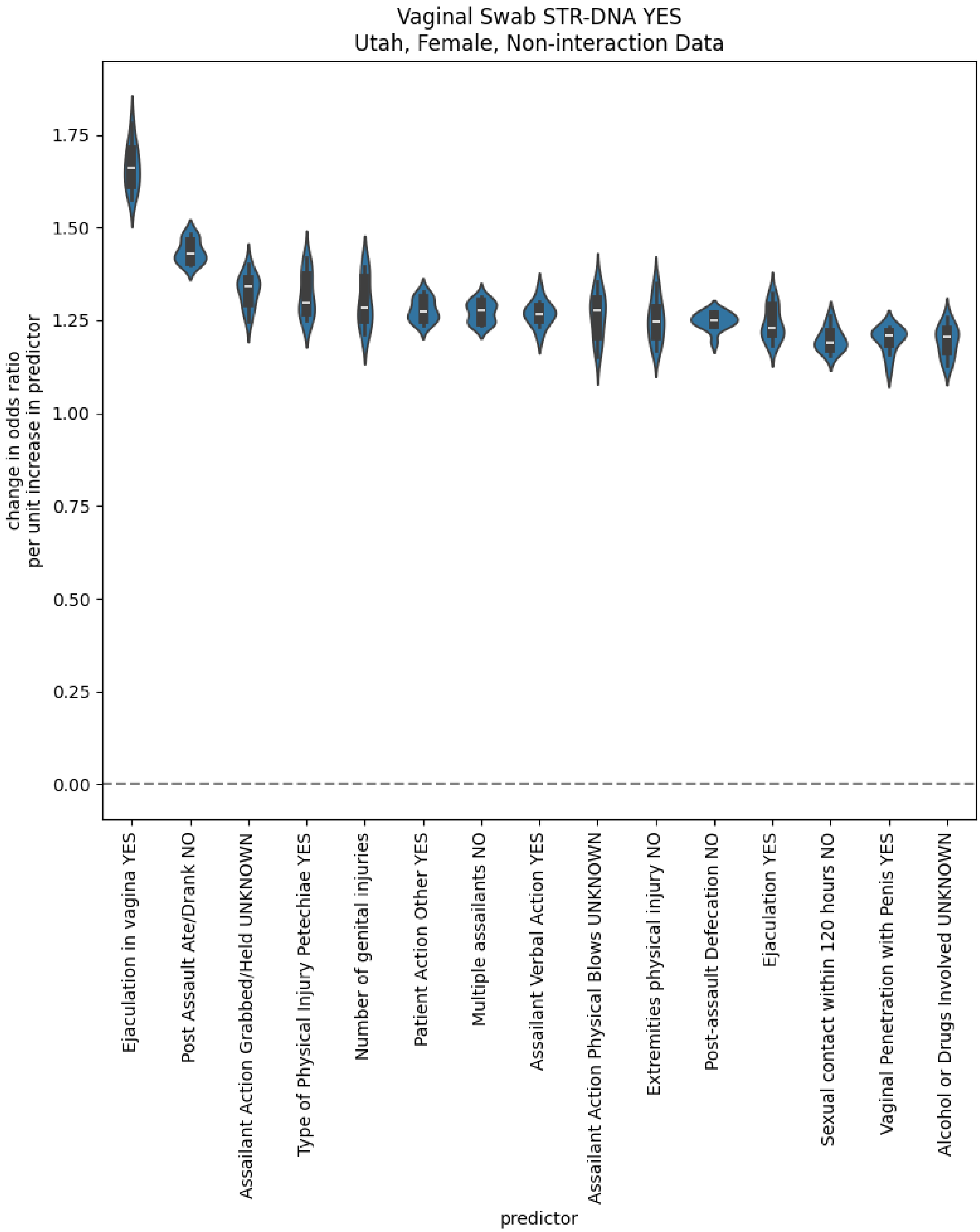


Figure 22. Vaginal Swab Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

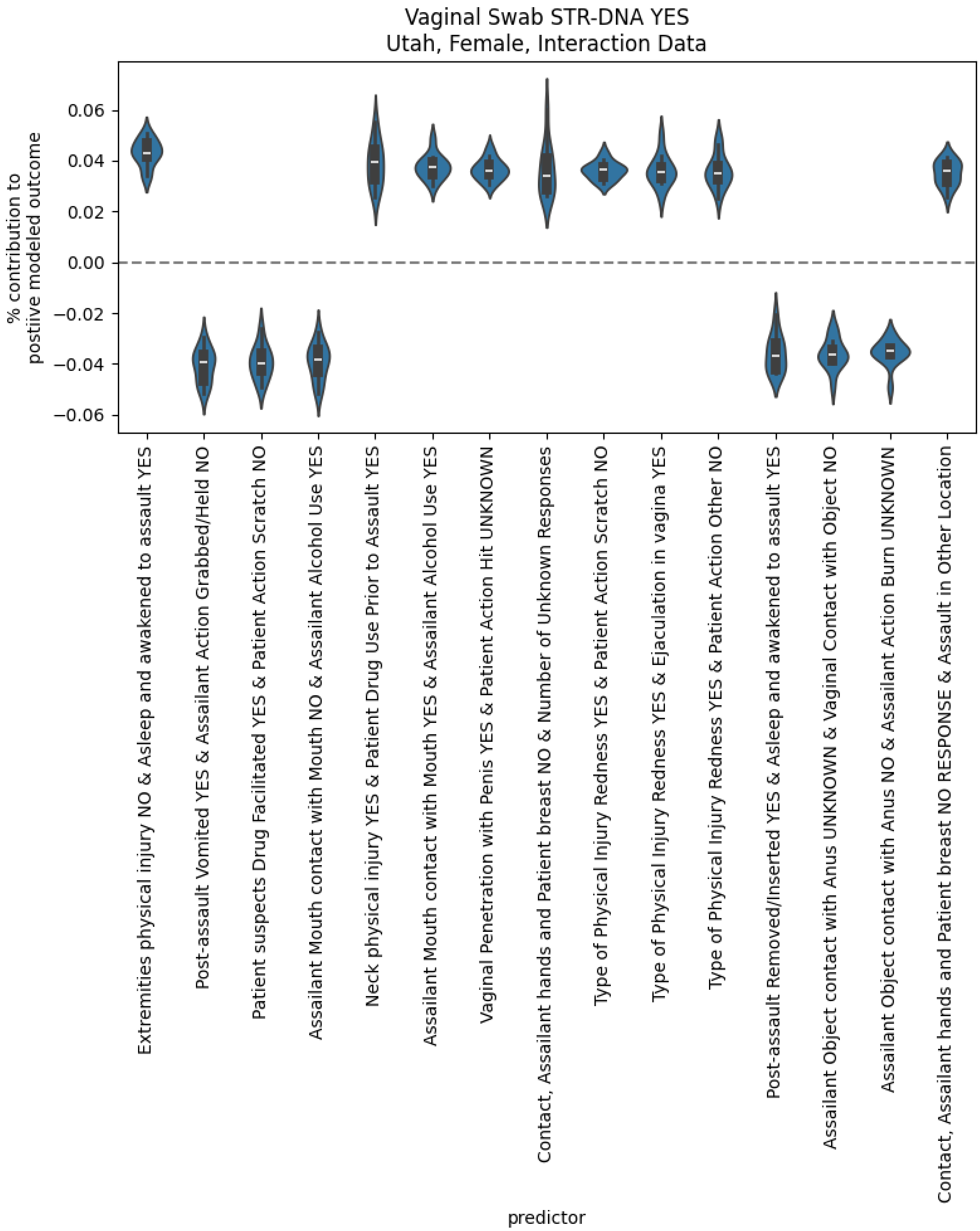
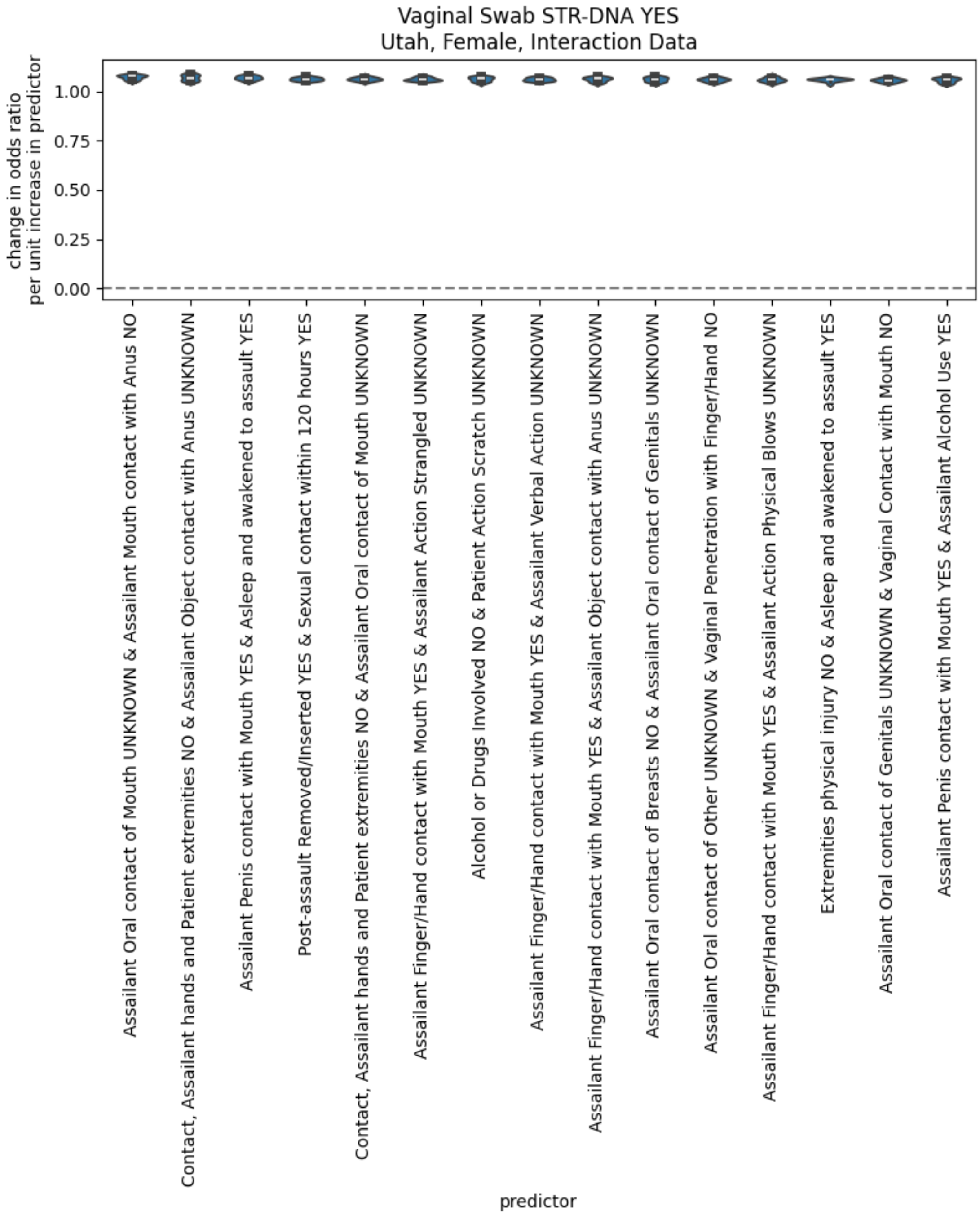


Figure 23. Vaginal Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



<u>Summary of Key Findings from Vaginal Swab</u>

In summarizing the non-interaction models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the vaginal swab include ejaculation in vagina, penetration of penis in vagina, lack of post-assault defecation, single assailant rather than multiple assailants, lack of eating/drinking post-assault (correlated with time between assault and SAMFE), and genital injuries.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Cervical Swabs (n=772):

Figure 24. Cervical Swab Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

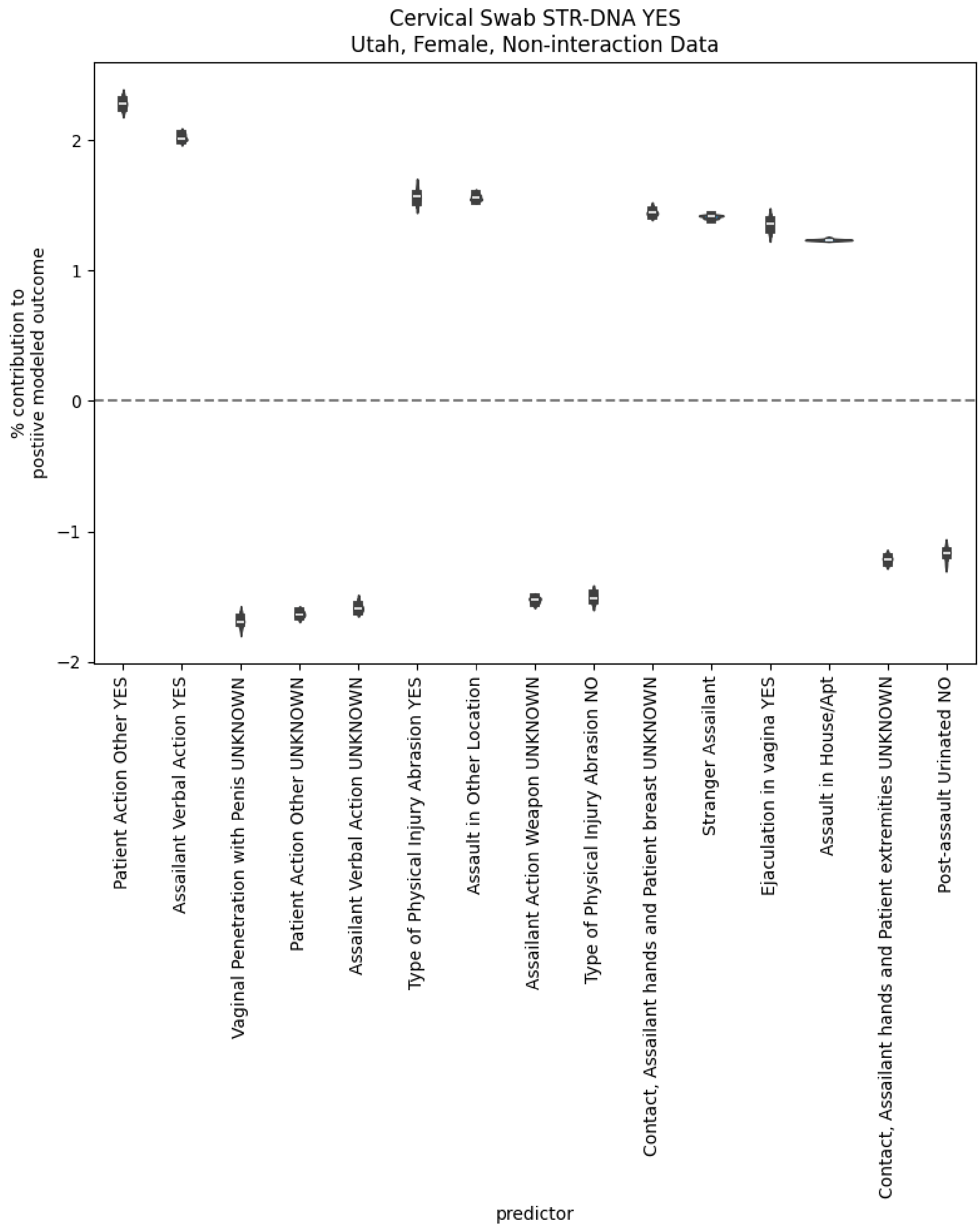


Figure 25. Cervical Swab Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

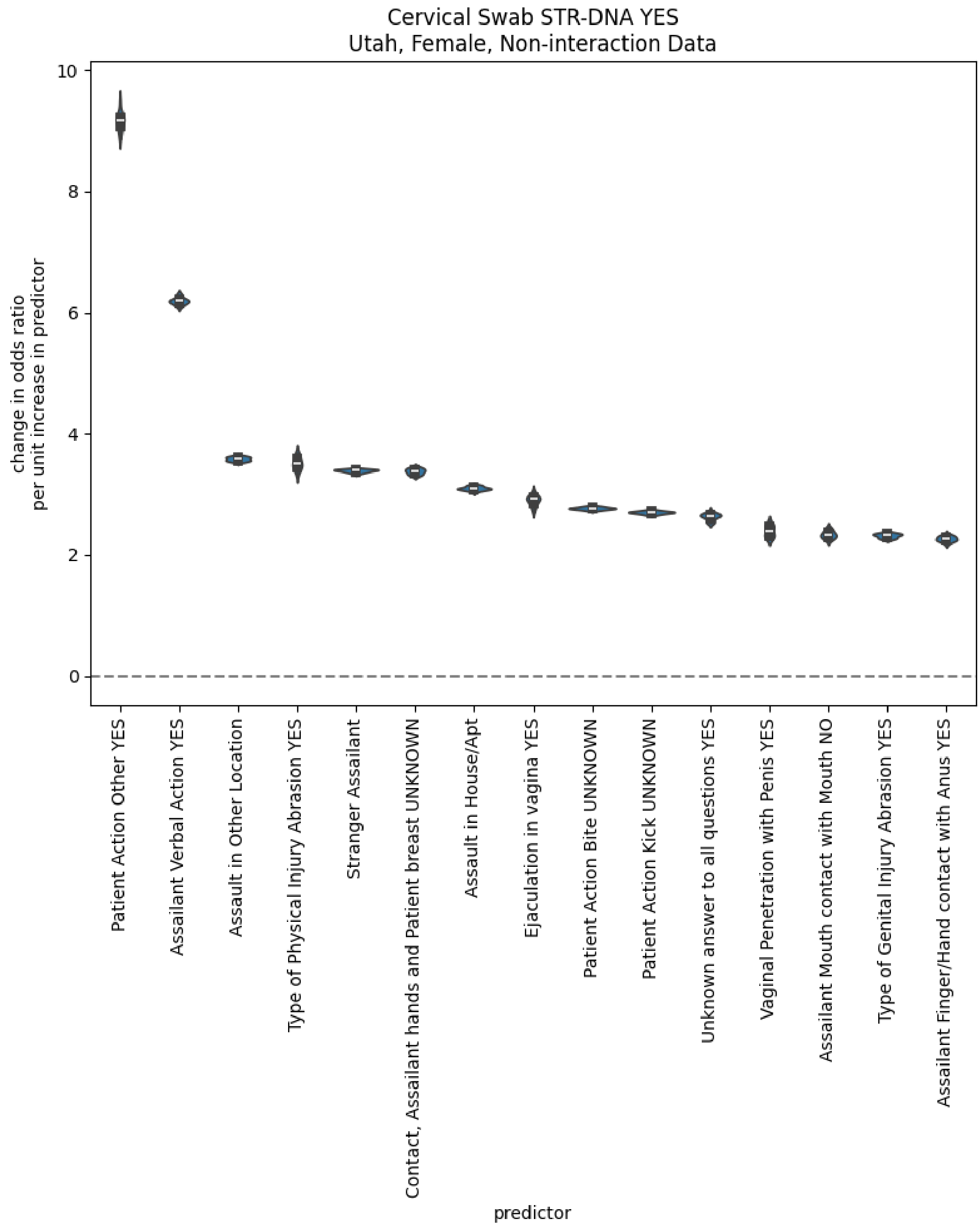


Figure 26. Cervical Swab Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

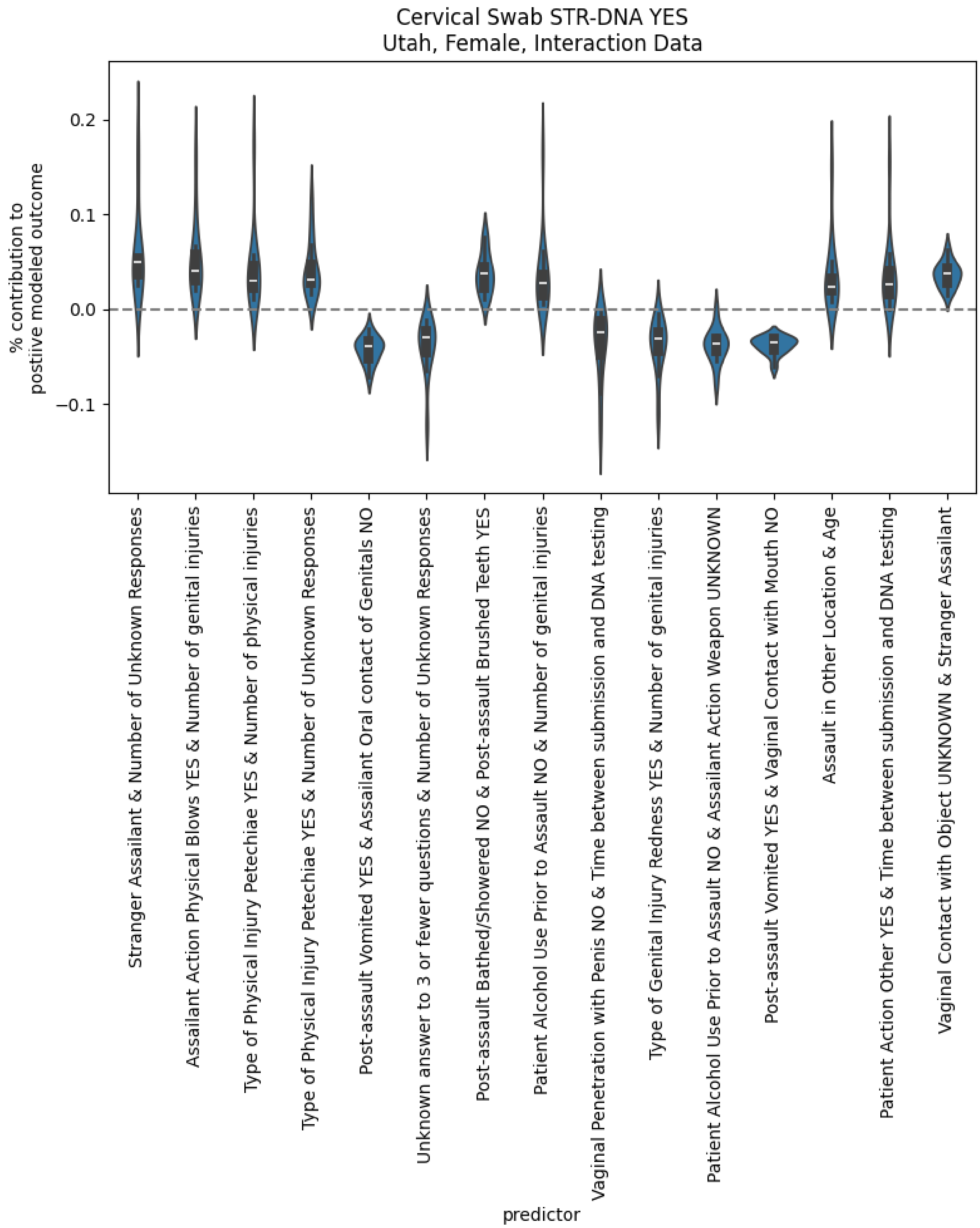
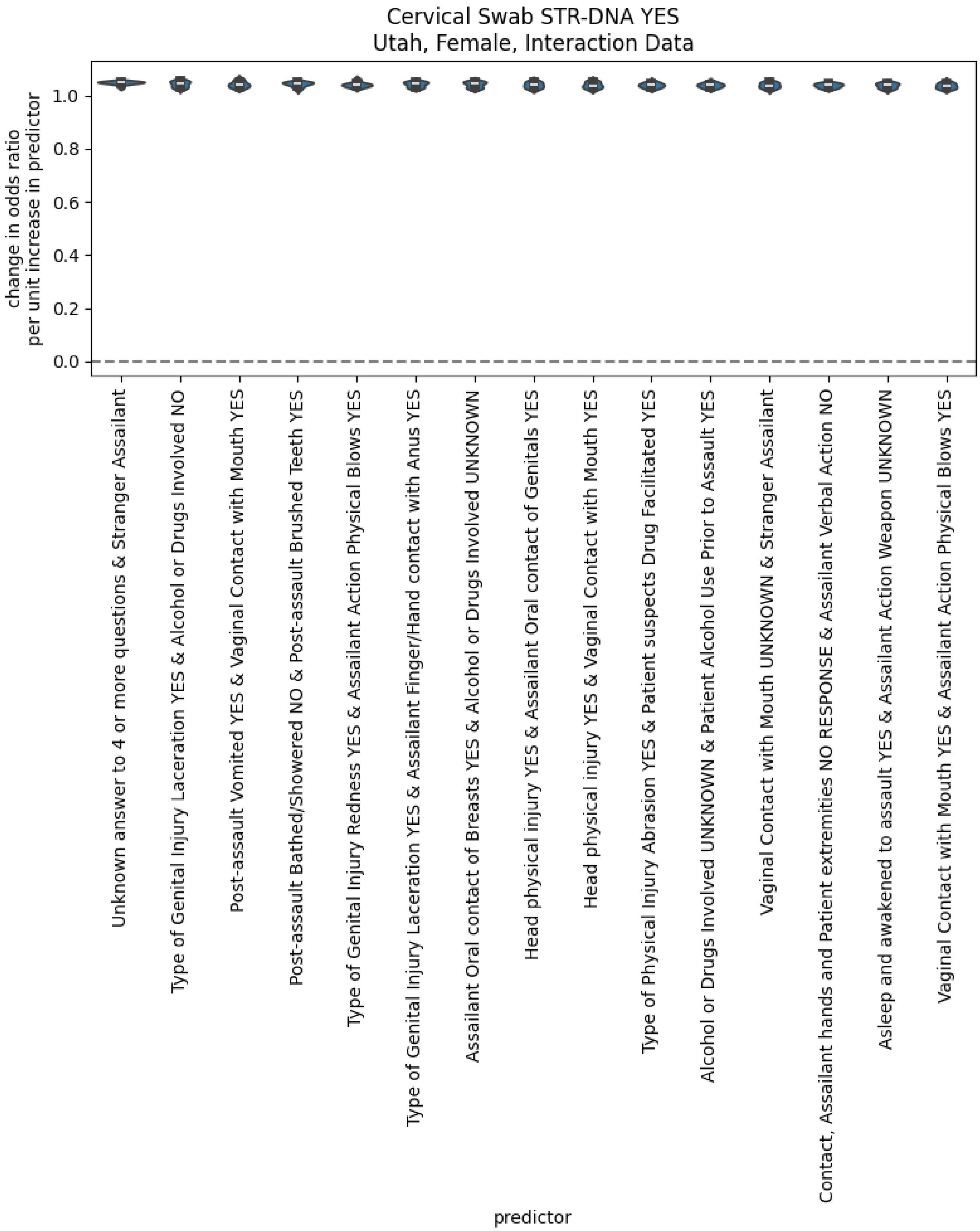


Figure 27. Cervical Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Cervical Swab

In summarizing the non-interaction models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the cervical swabs include patient action of “other” (generally indicates victim shoved or pushed assailant), verbal threats or coercion by assailant, assault locations, physical injury of abrasion, and stranger assailant. Interestingly, ejaculation in vagina and penetration of penis in vagina had lower odds ratio of predicting positive results.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Rectal Swabs (n=1031)

Females:

Figure 28. Rectal Swab, Female, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

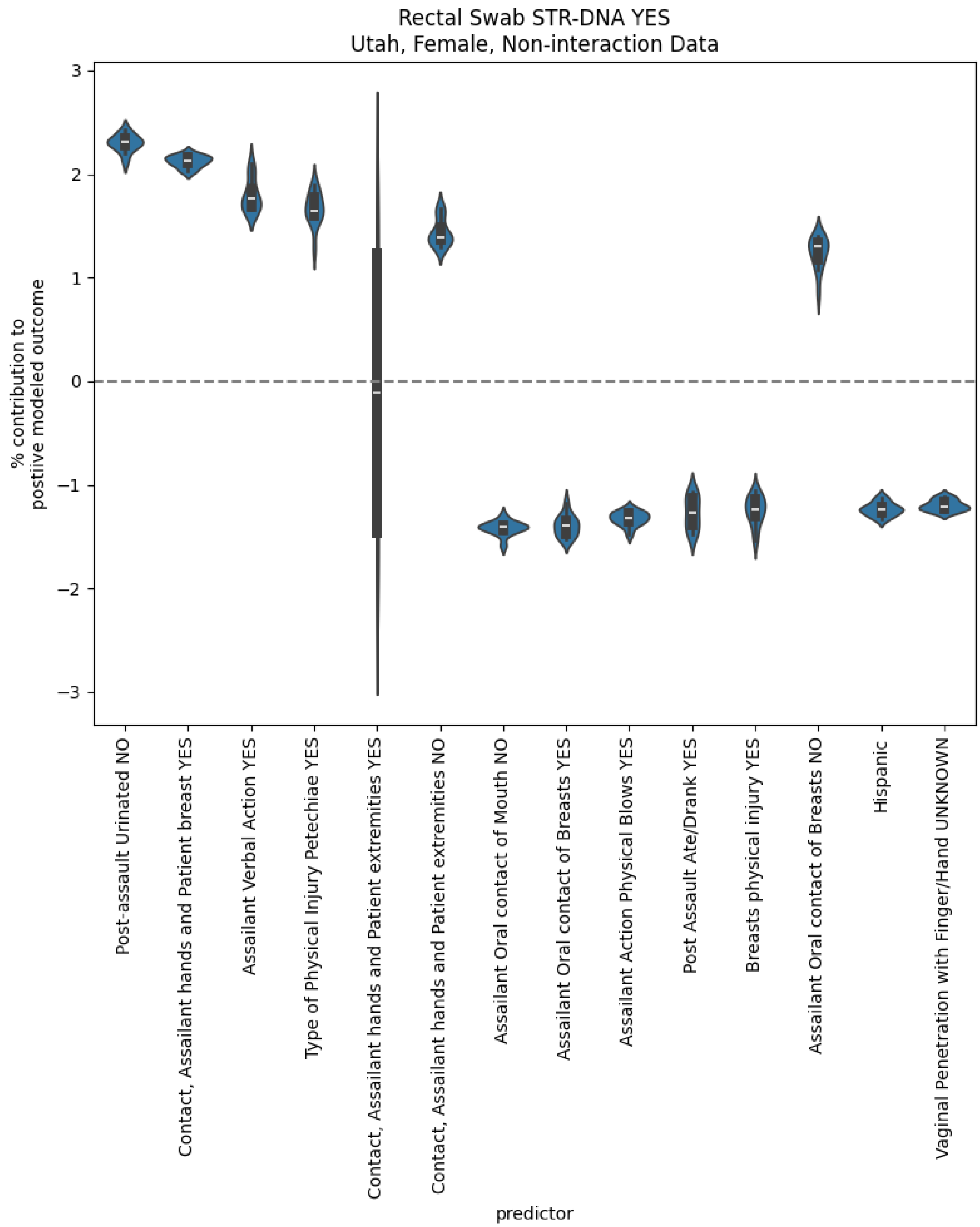


Figure 29. Rectal Swab, Female, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

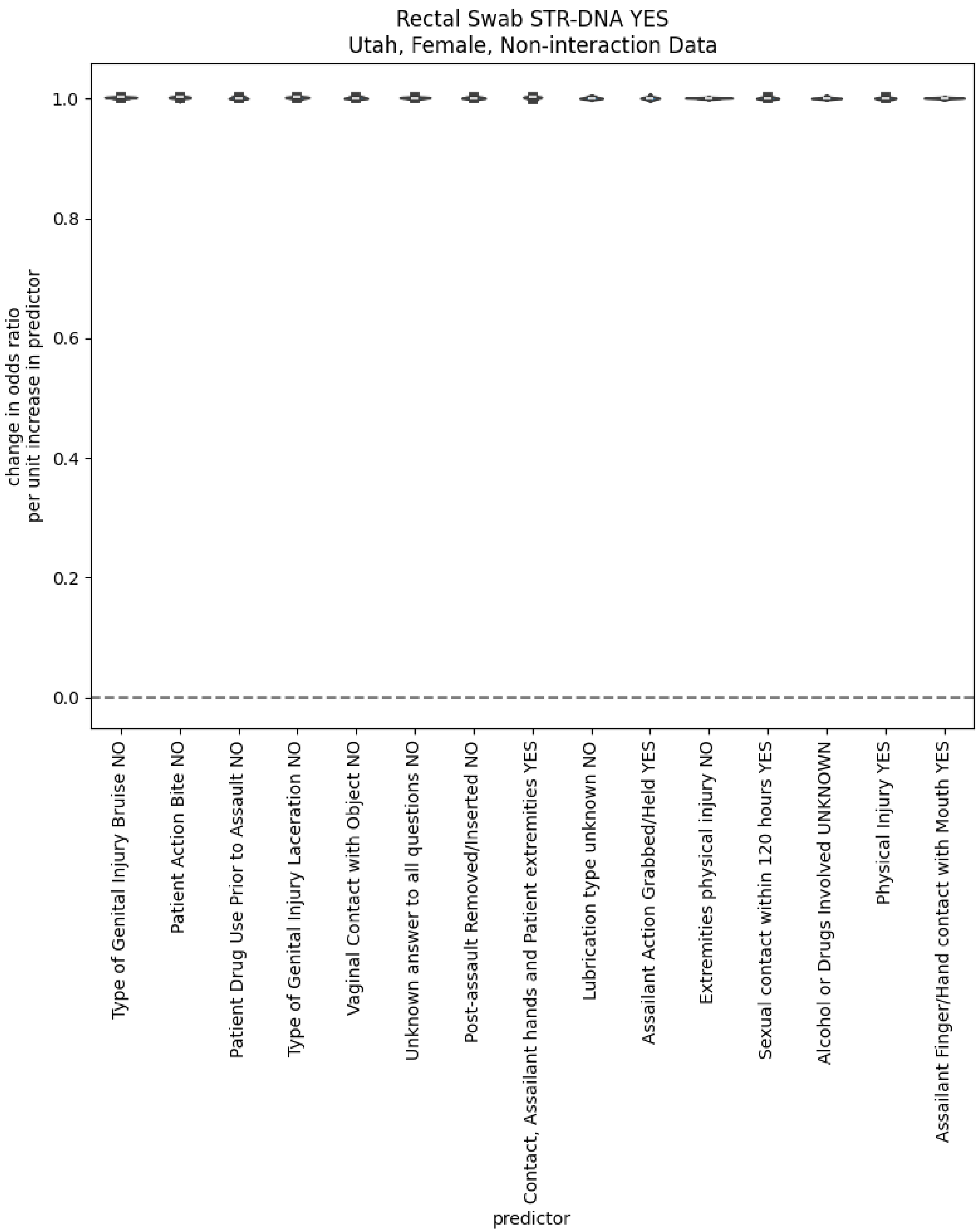


Figure 30. Rectal Swab, Female, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

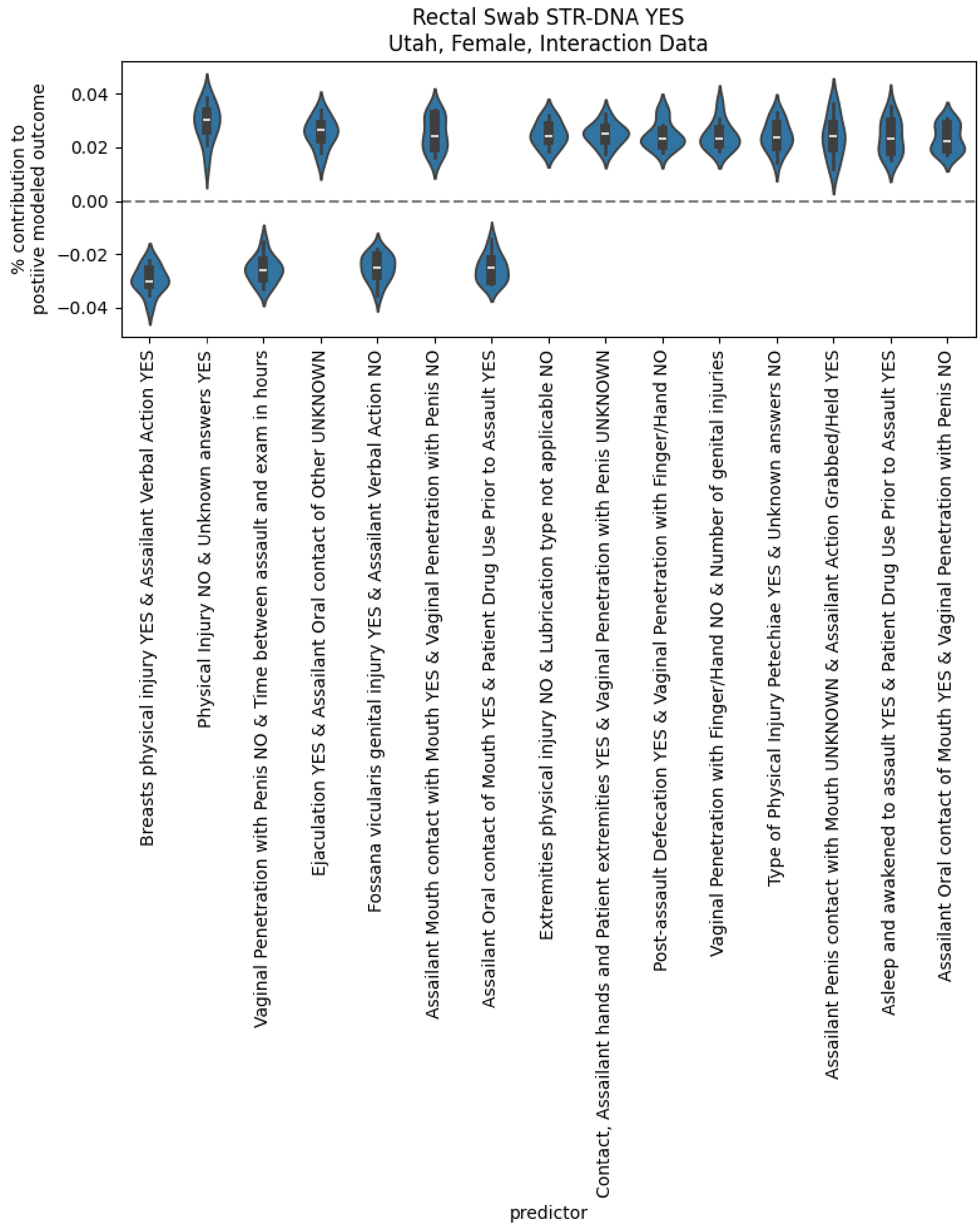
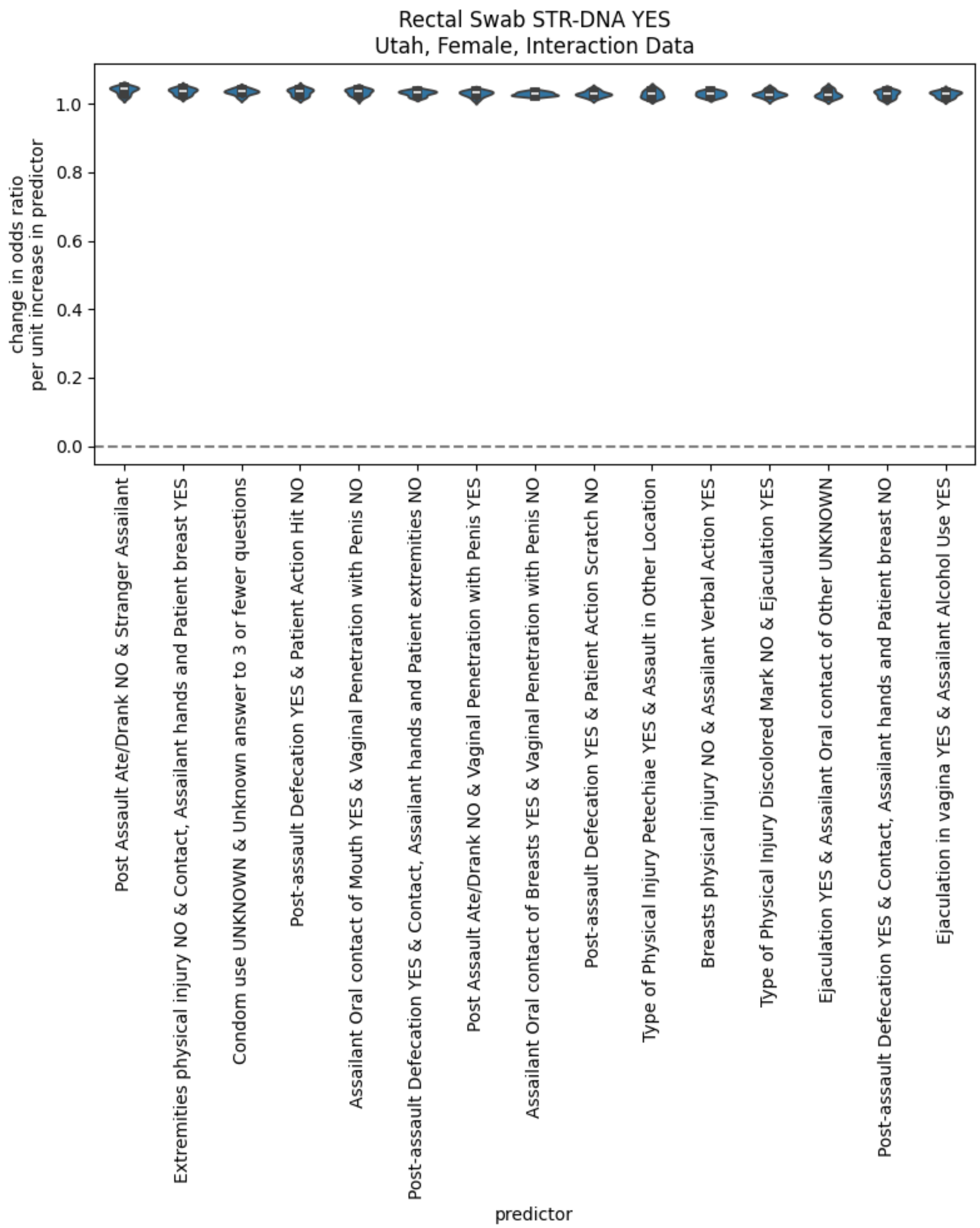


Figure 31. Rectal Swab, Female, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



<u>Summary of Key Findings from Rectal Swab from Females</u>

The models exploring variables for development of full or partial STR DNA profiles of foreign contributors from rectal swabs indicated that no variables were found to significantly predict positive outcomes.

<u>Males:</u>

Figure 32. Rectal Swab, Male, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

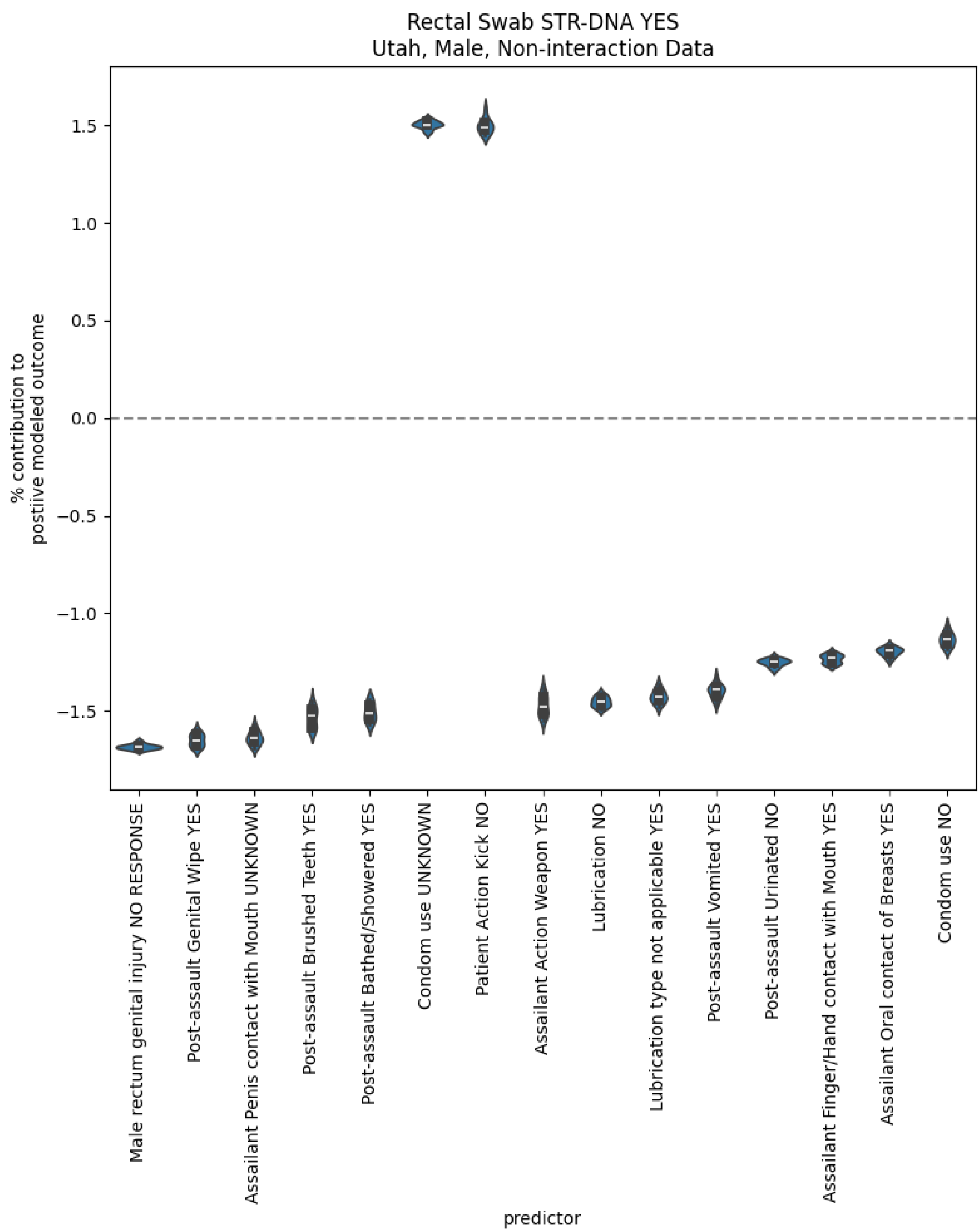


Figure 33. Rectal Swab, Male, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

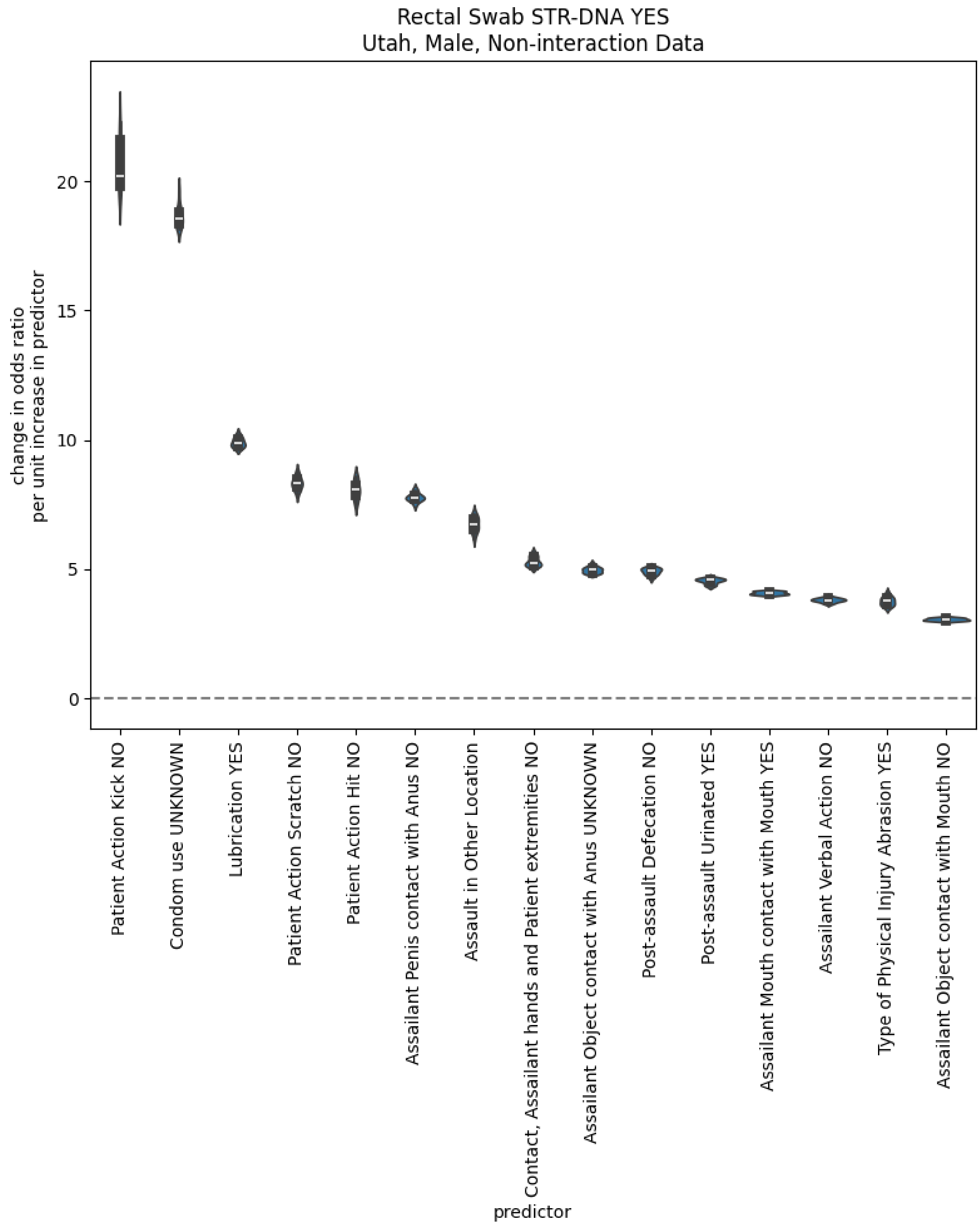


Figure 34. Rectal Swab, Male, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

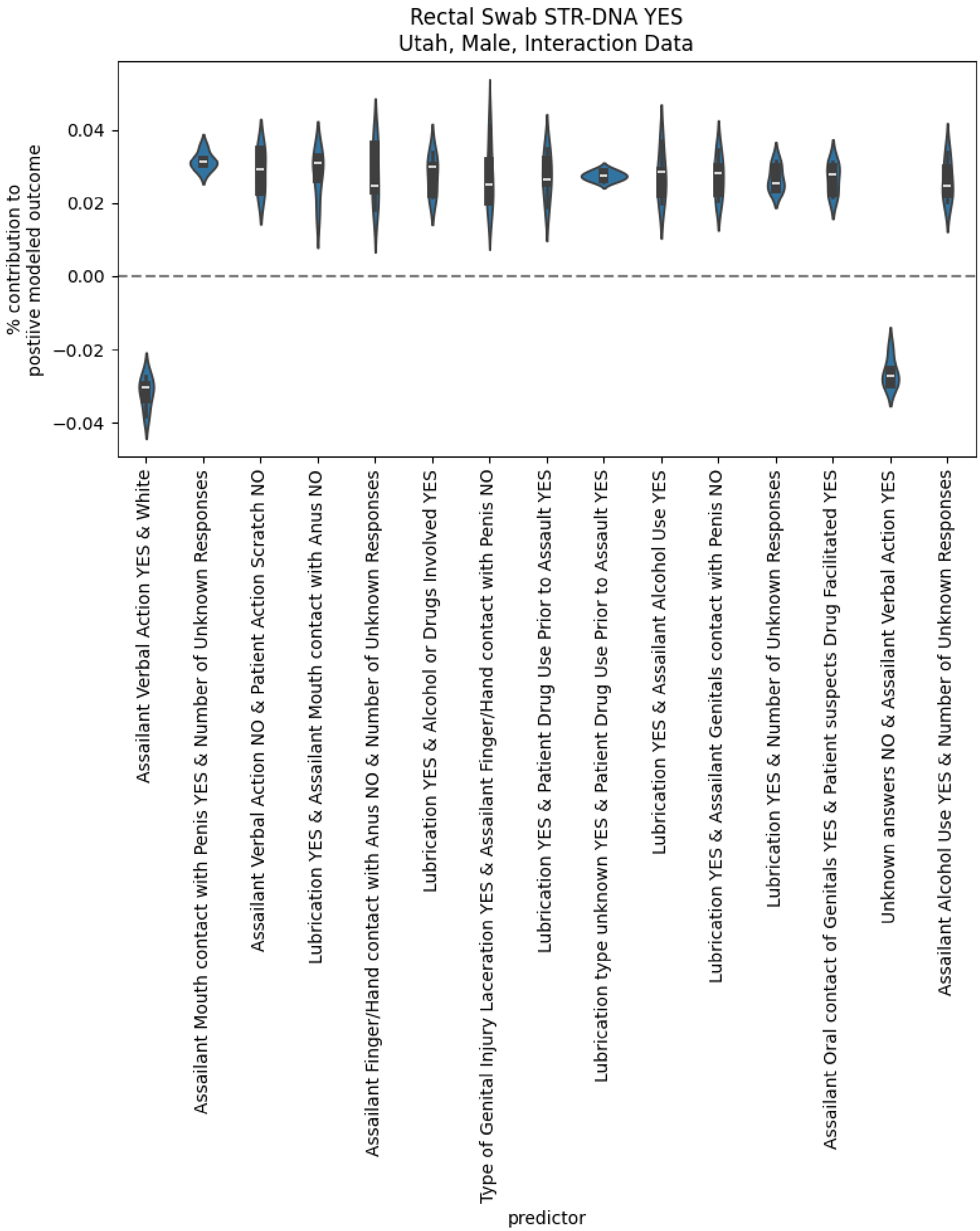
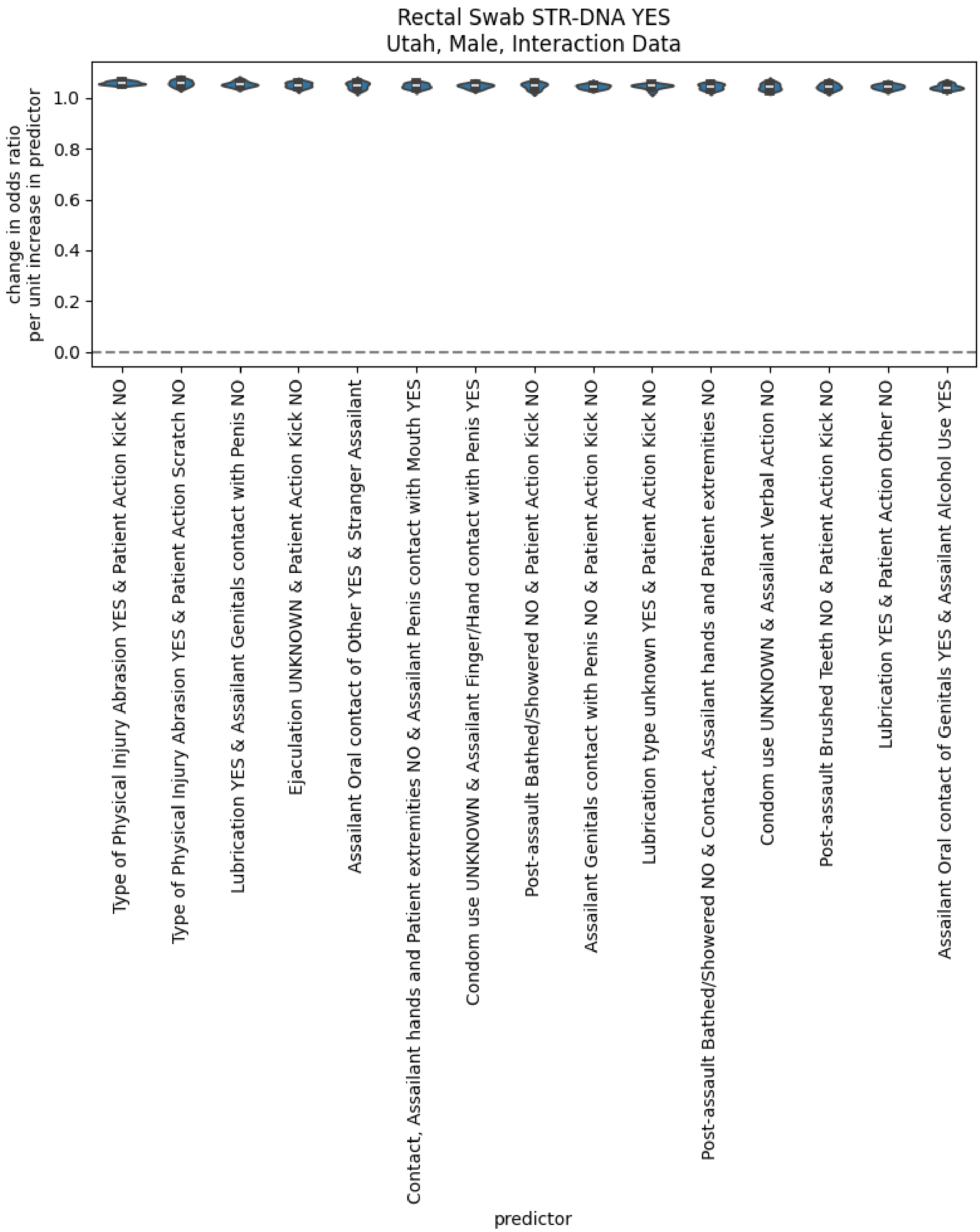


Figure 35. Rectal Swab, Male, Not Normalized with Interactions Change in Odds Ratio of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Rectal Swabs from Males

In summarizing the non-interaction models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the rectal swabs of males include if victim did not kick, scratch, or hit the assailant (indicating victim did not physically resist during the assault); unknown condom use, positive lubrication use, and lack of post-assault defection. A finding requiring further investigation is that if the victim reported the assailant's penis did *not* contact the anus, the odds of developing full or partial profile of foreign contributor(s) increased.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Oral Swabs (n=442)

Females:

Figure 36. Oral Swab, Female, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

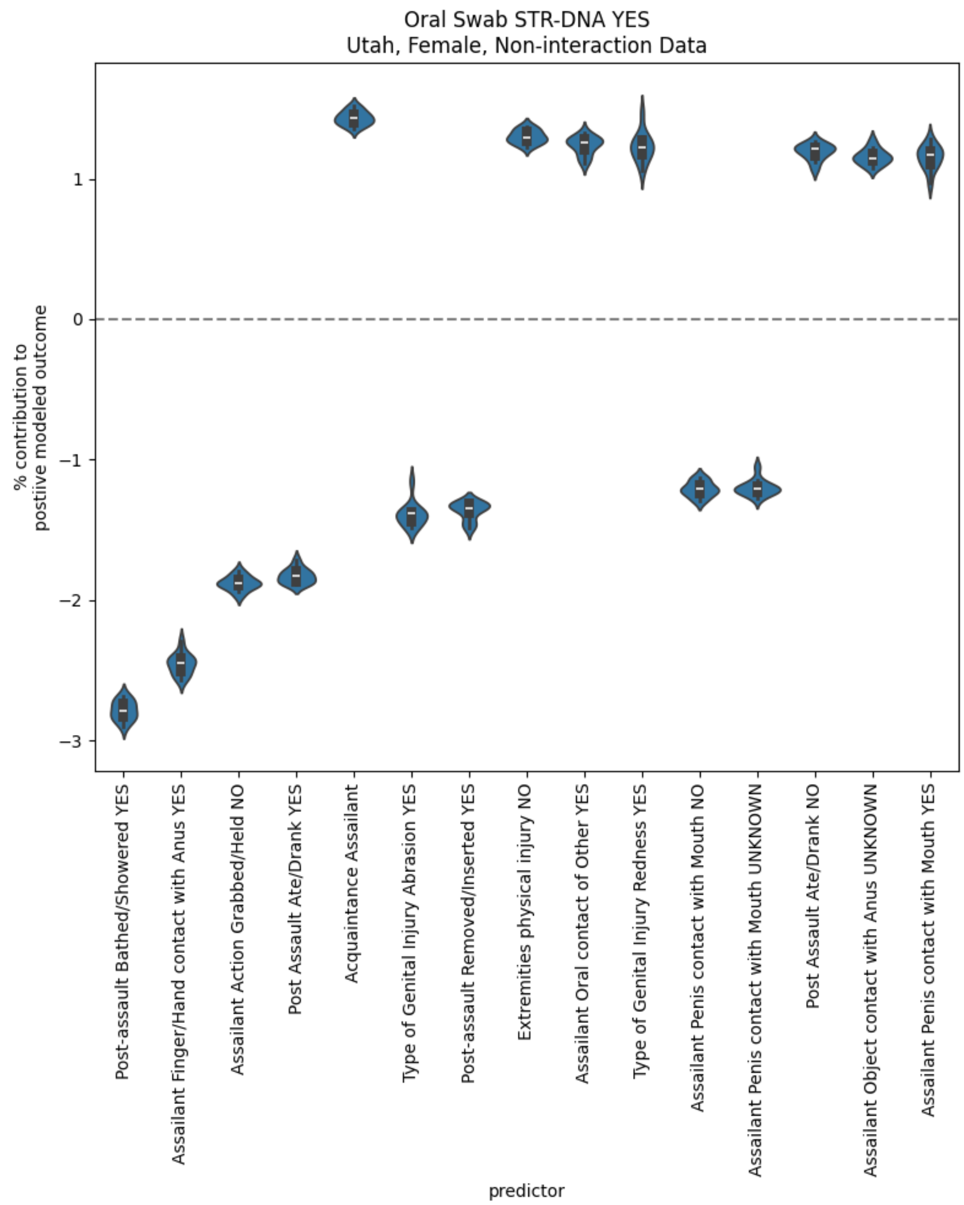


Figure 37. Oral Swab, Female, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

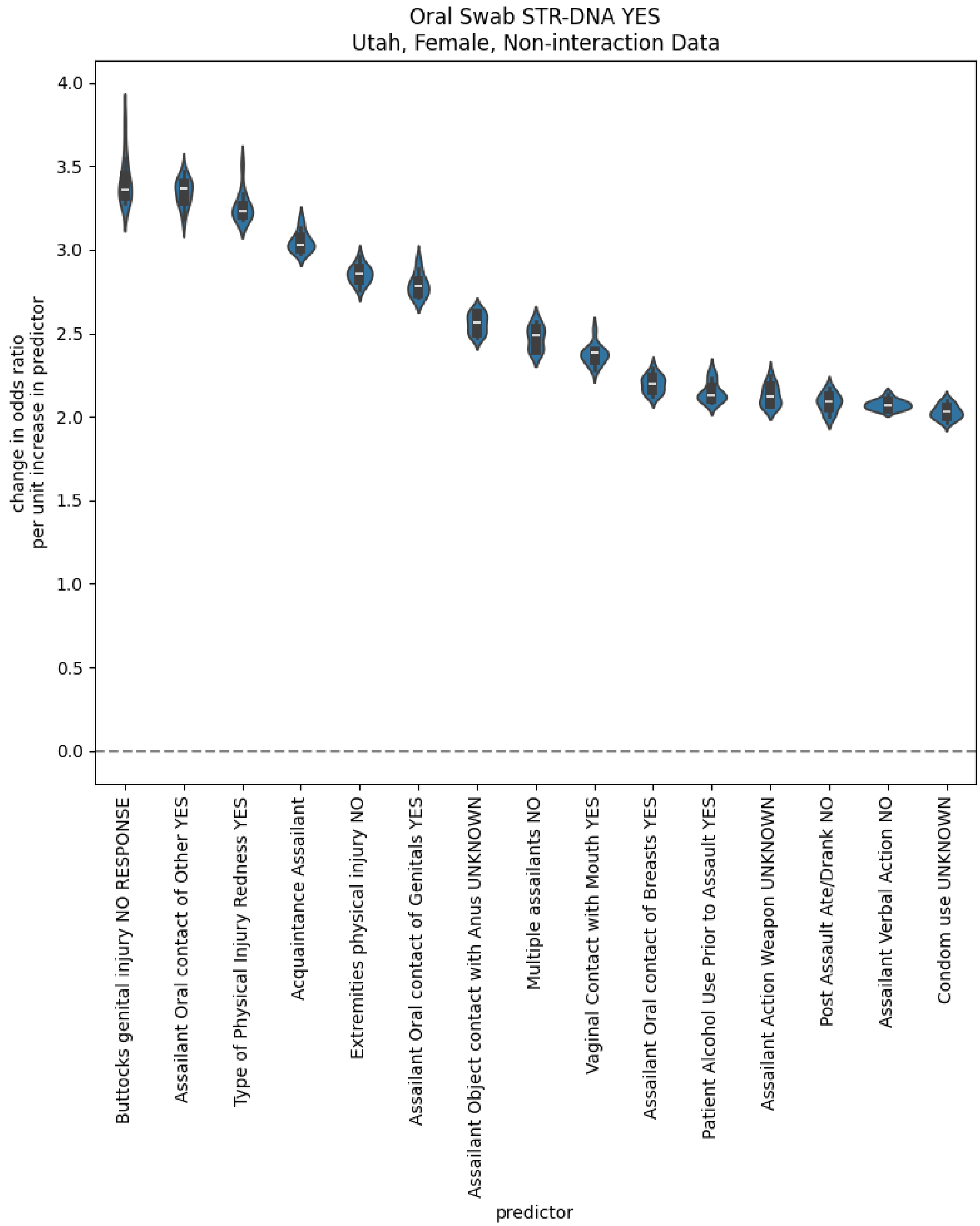


Figure 38. Oral Swab, Female, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

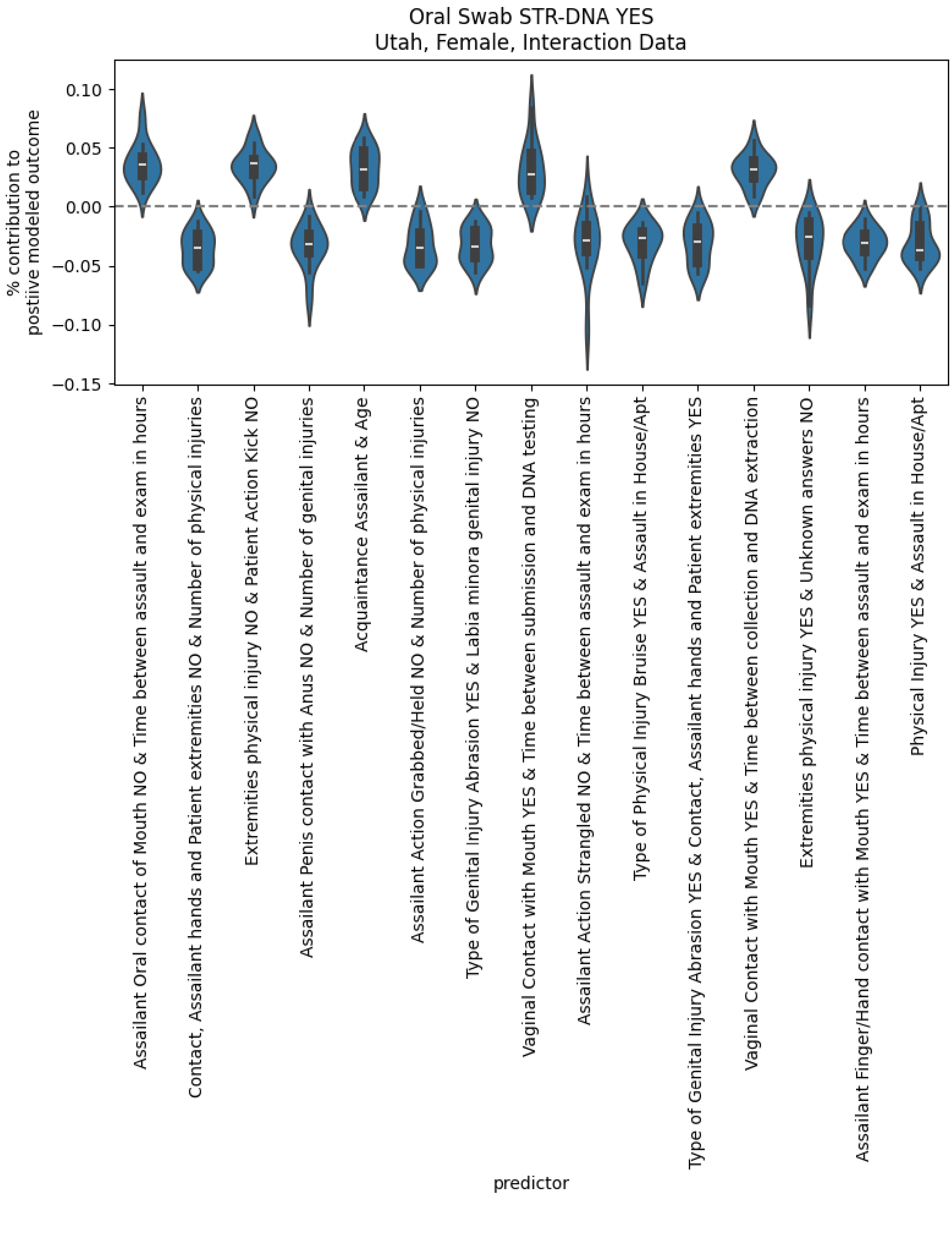
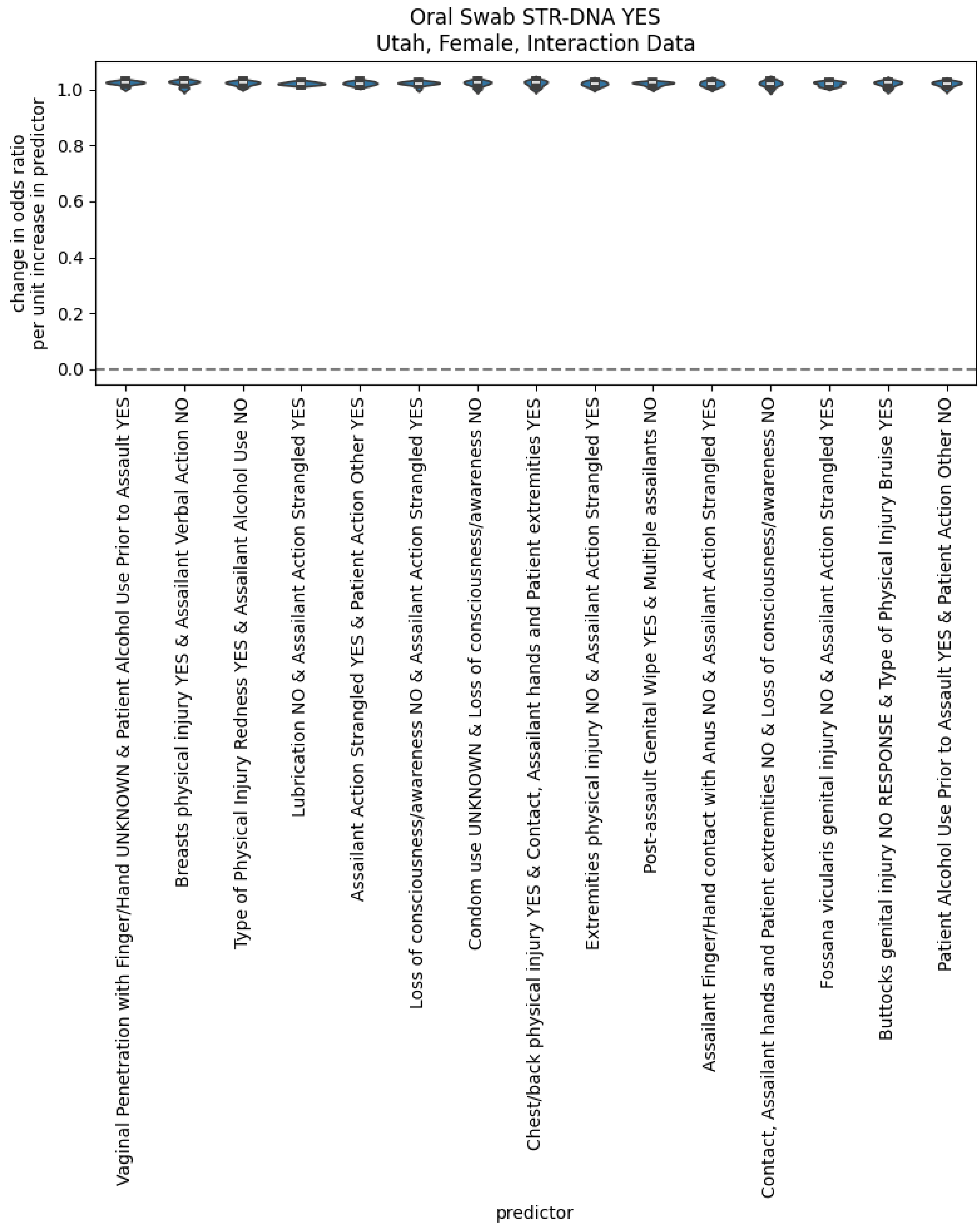


Figure 39. Oral Swab, Female, Not Normalized with Interactions Change in Odds Ratio of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Oral Swabs from Females

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the oral swabs of females include if oral contact by assailant of genitals, breasts, and other body locations; acquaintance relationship; lack of multiple assailants; assailant penis contact of mouth; and no post-assault eating or drinking prior to SAMFE. Mouth-to-mouth contact, "kissing," between assailant and victim (48.5% of cases) was not a predictor in the non-interaction models.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Males:

Figure 40. Oral Swab, Male, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

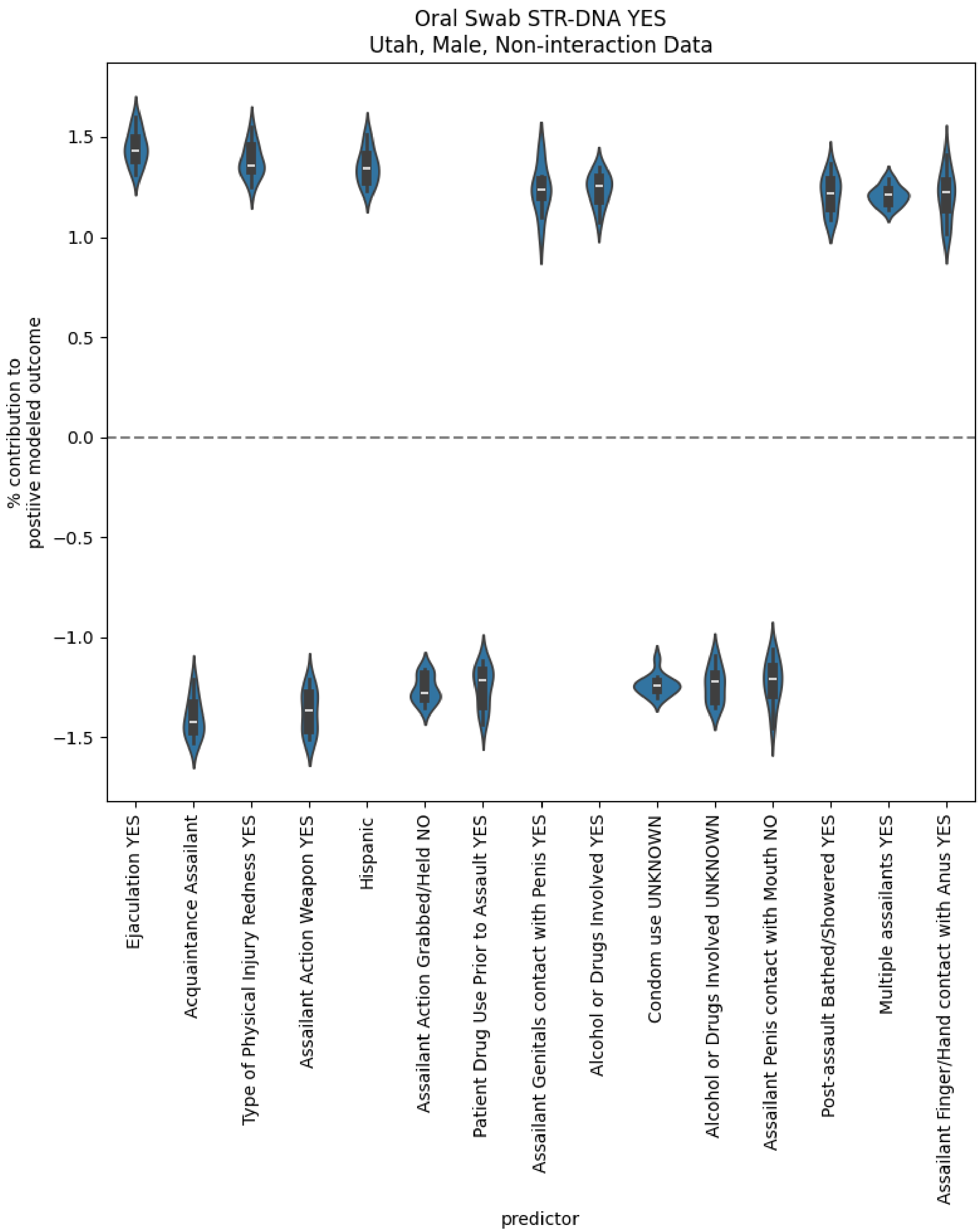


Figure 41. Oral Swab, Male, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

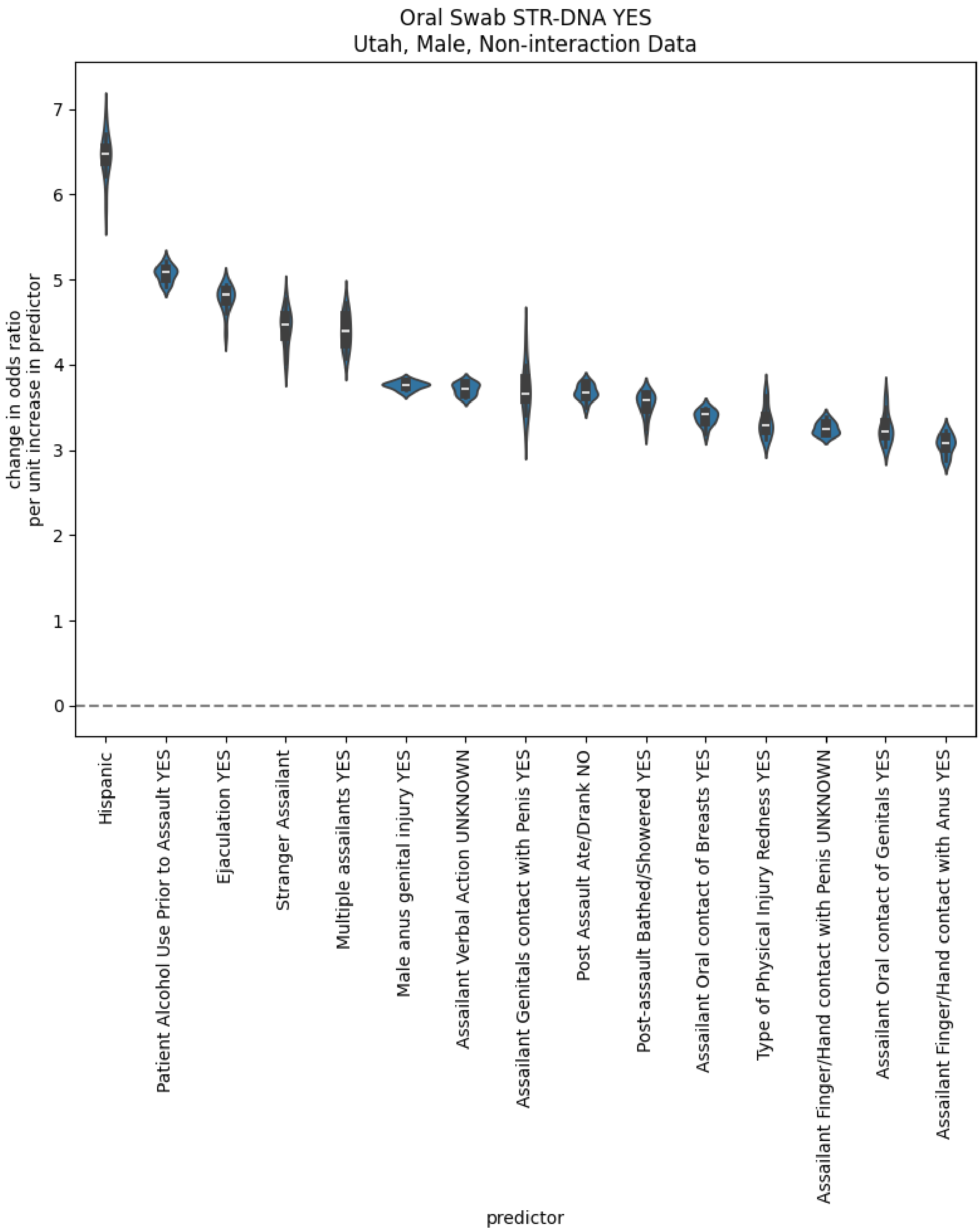


Figure 42. Oral Swab, Male, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

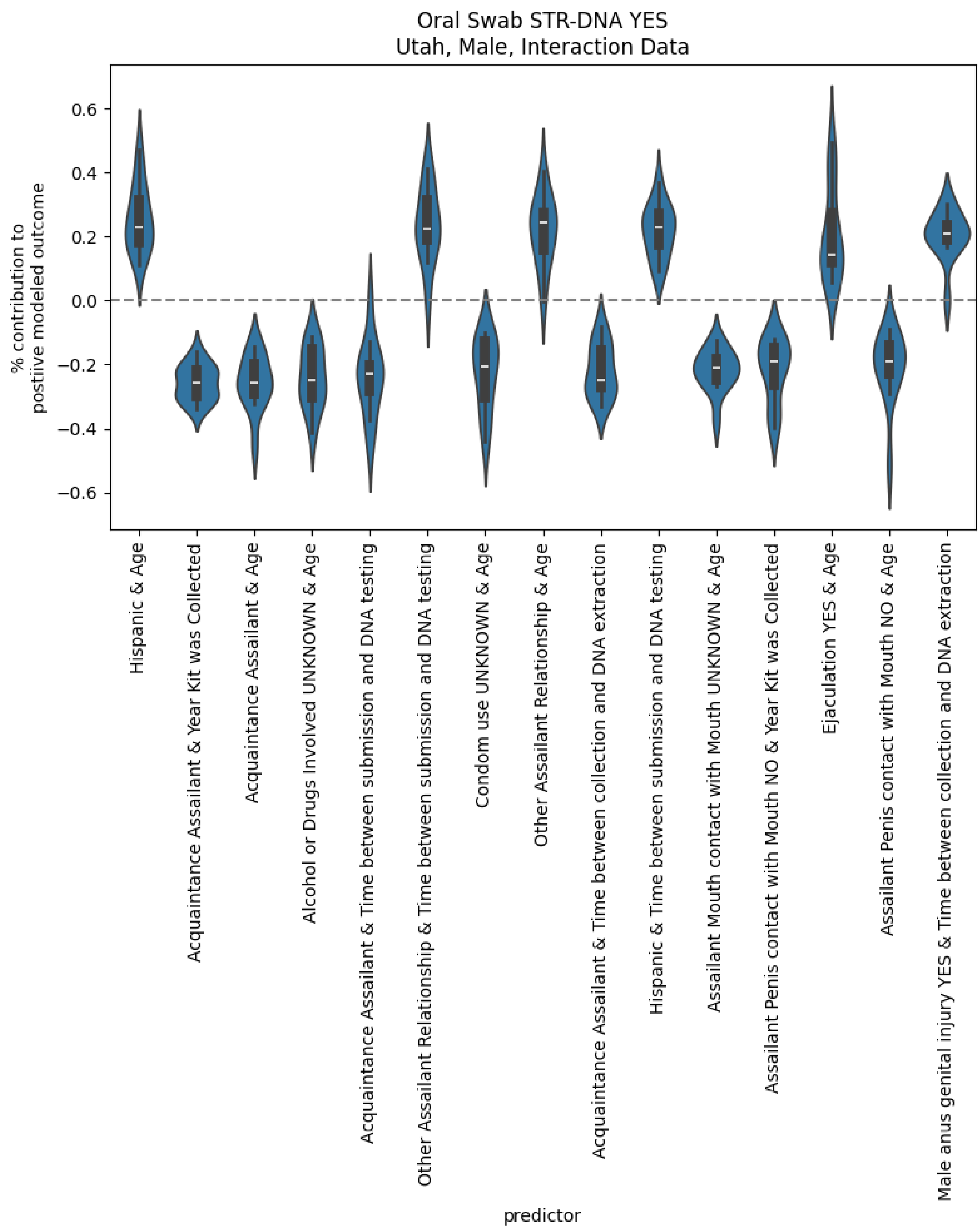
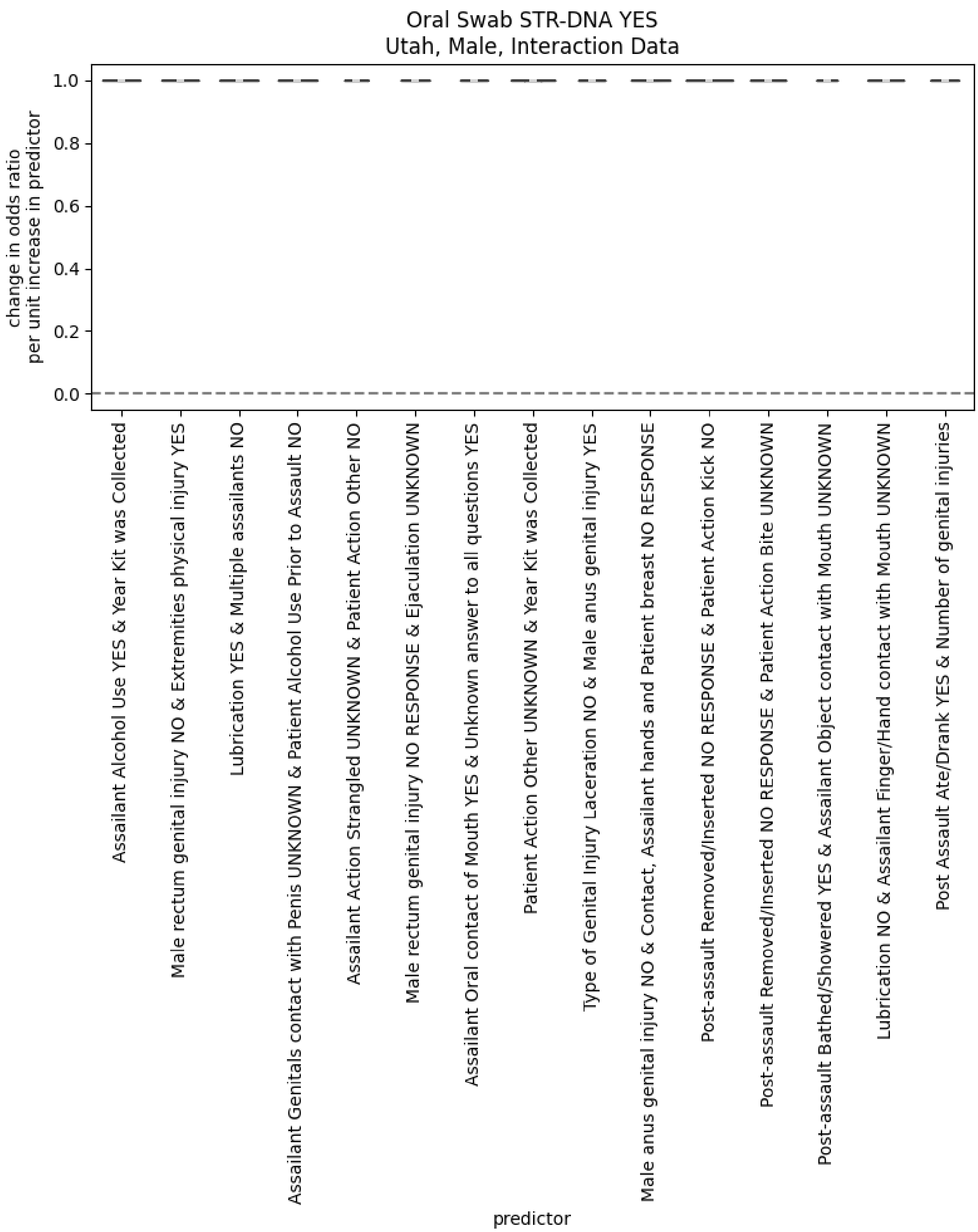


Figure 43. Oral Swab, Male, Not Normalized with Interactions Change in Odds Ratio of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Oral Swabs from Males

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the oral swabs of males include Hispanic race, victims' use of alcohol, ejaculation occurred, multiple assailants, genital injury, and lack of eating or drinking post-assault and prior to SAMFE. The inclusion of race as a significant variable requires further investigation.

The variable of victims' age was significant in several of the interaction features suggesting further investigation of the impact of age on outcomes of oral swabs.

Perianal Swabs (n=3574)

Females

Figure 44. Perianal Swab, Female, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

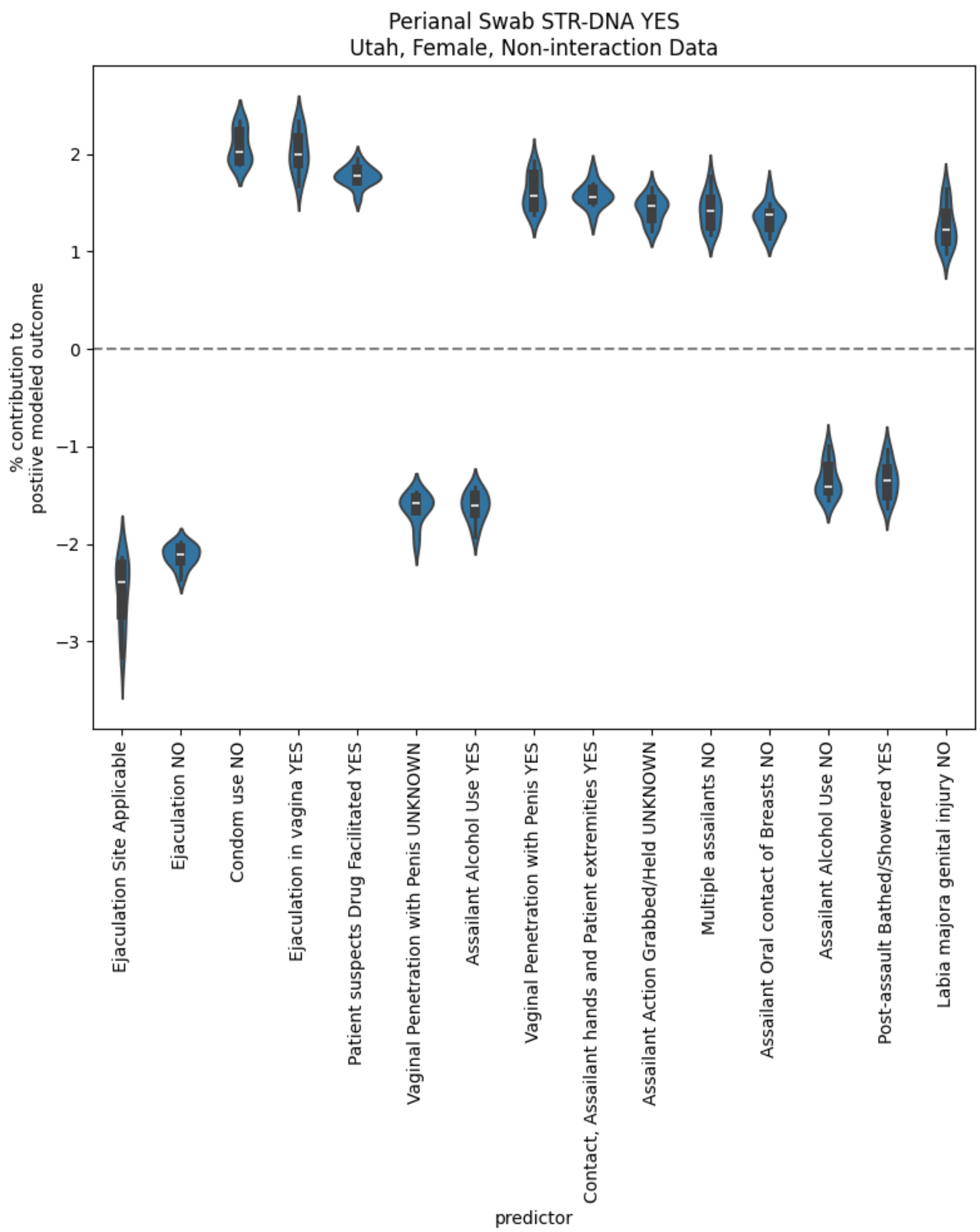


Figure 45. Perianal Swab, Female, Swab Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

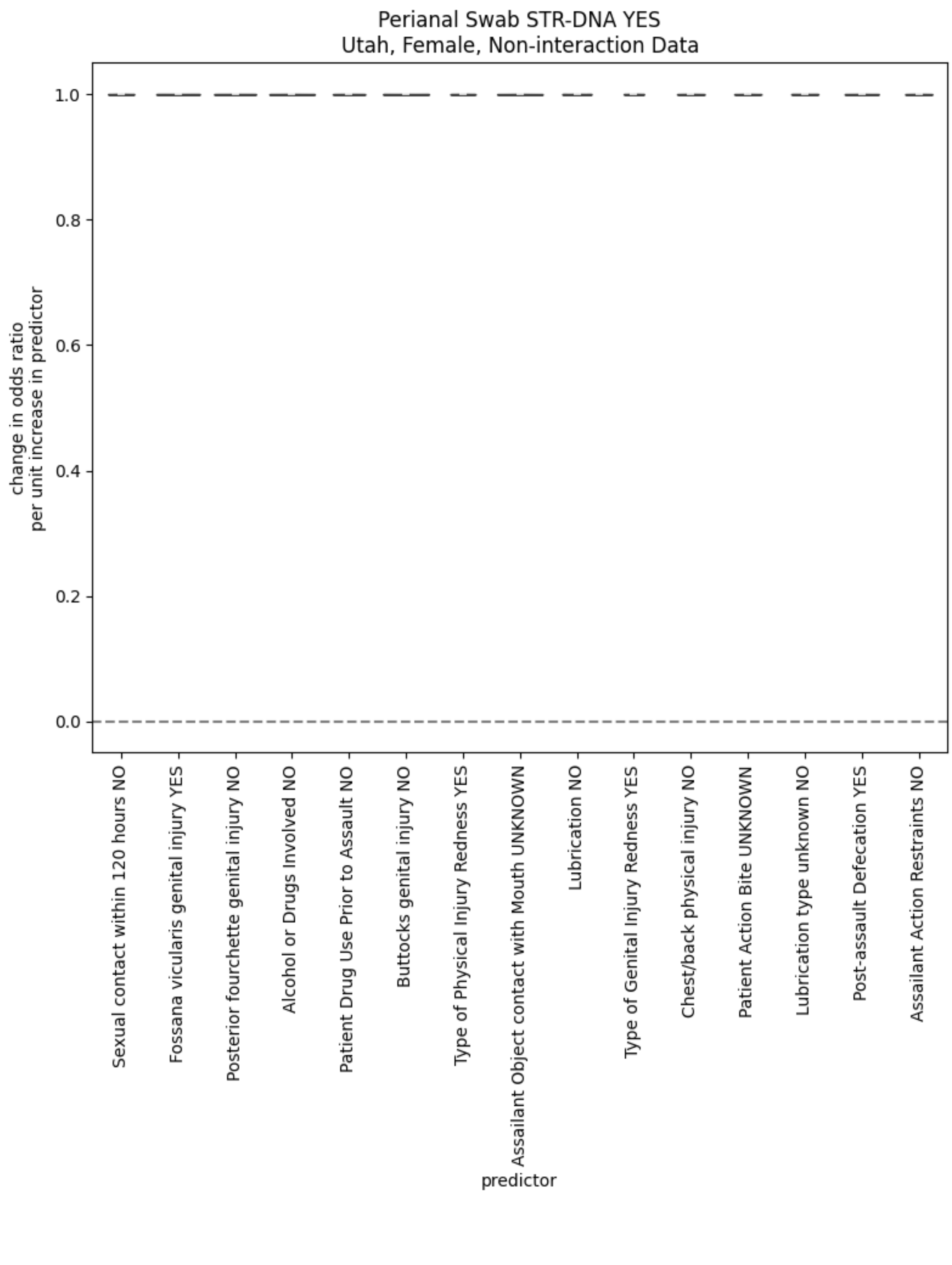


Figure 46. Perianal Swab, Female, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

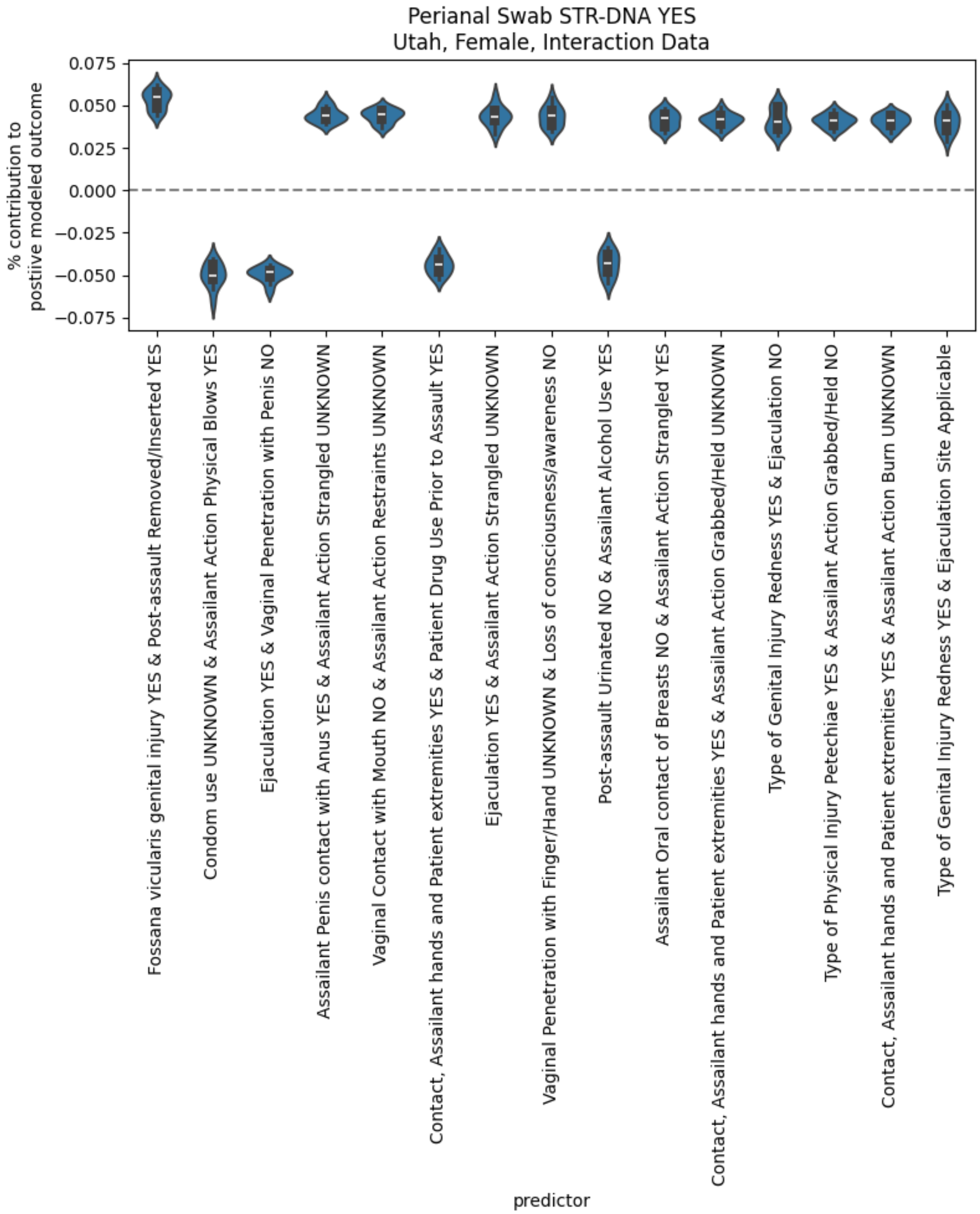
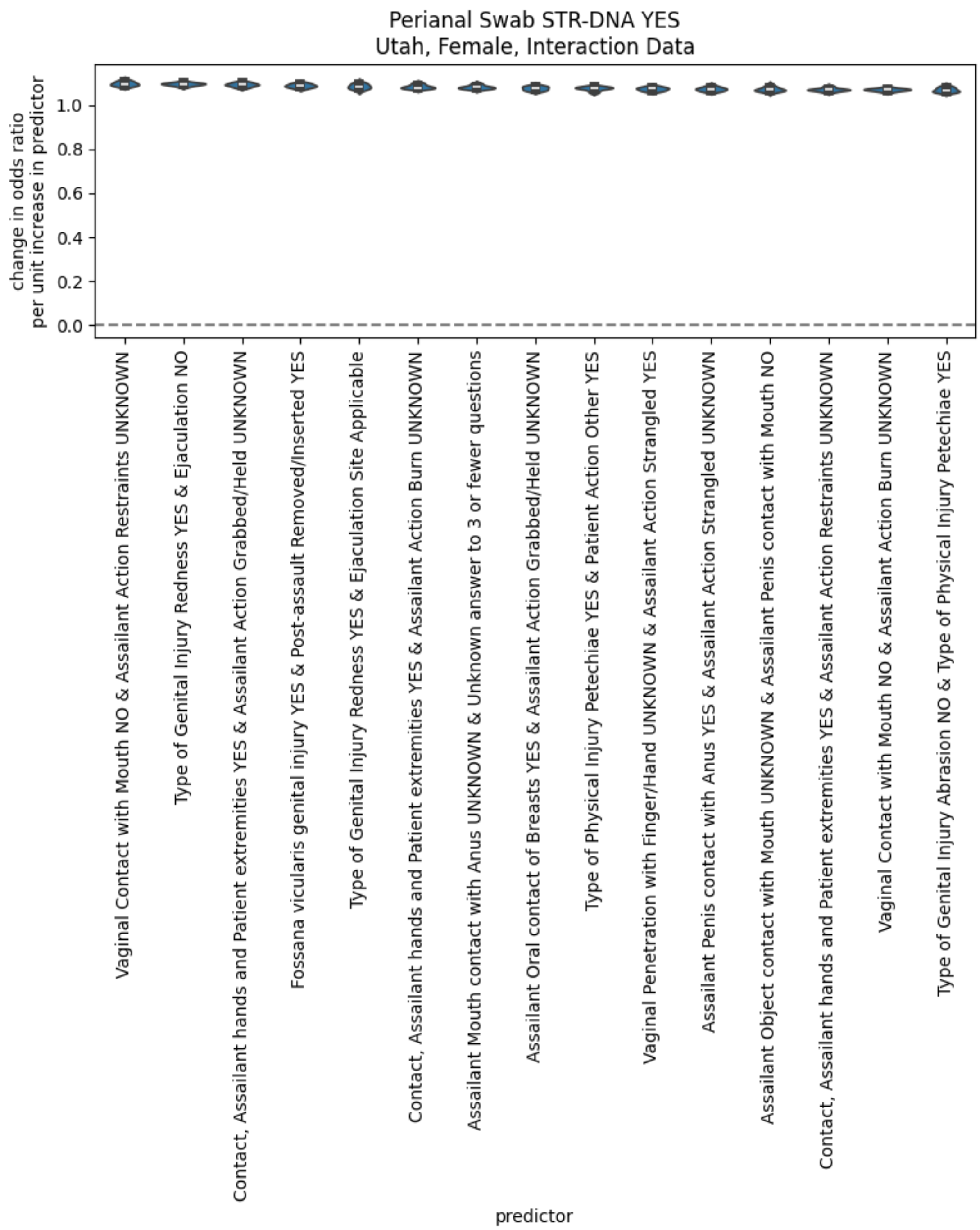


Figure 47. Perianal Swab, Female, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Perianal Swabs from Females

The coefficients in this model indicate that the variables do not meaningfully influence model output. The models were less accurate than random guessing.

Males:

Figure 48. Perianal Swab, Males, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

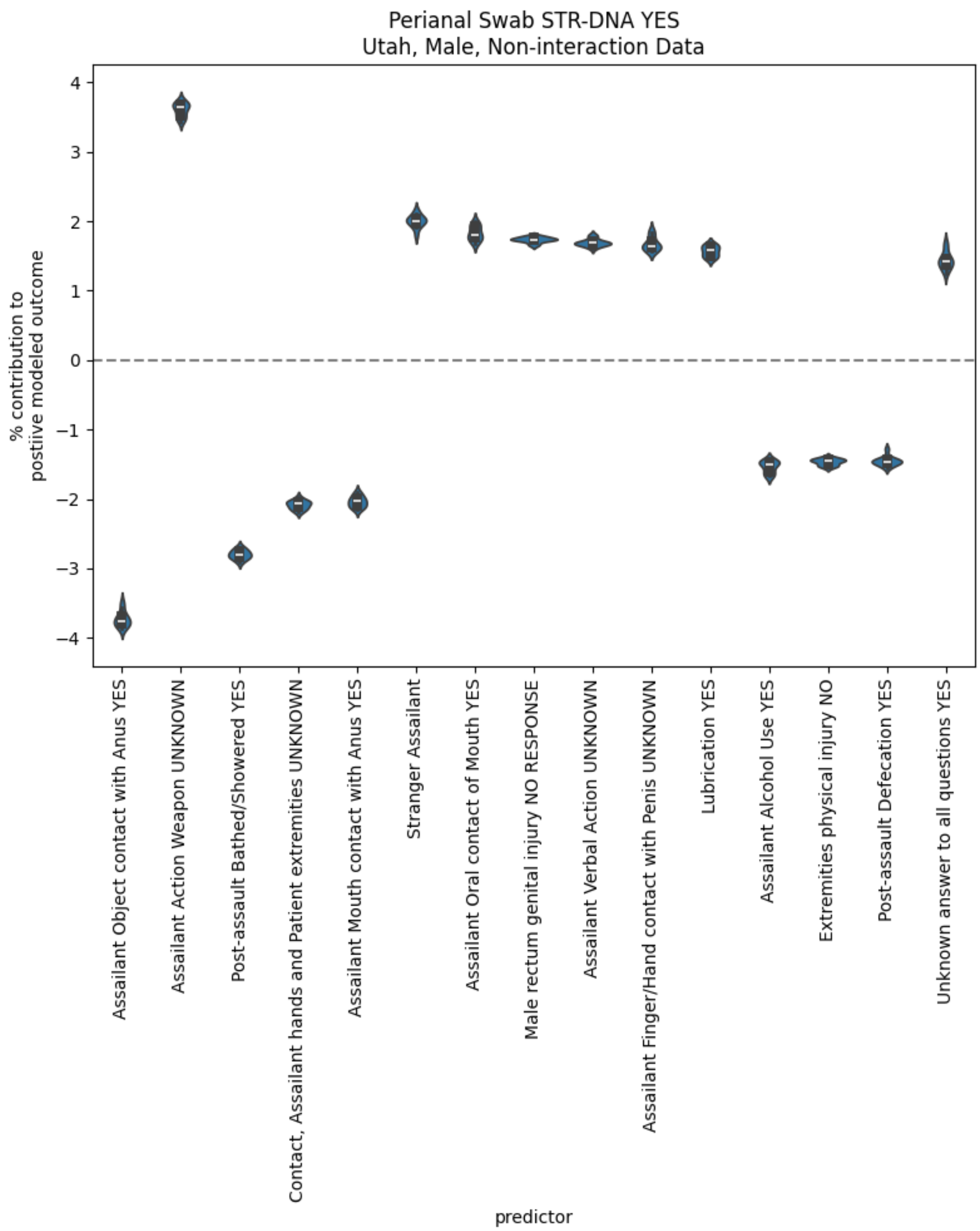


Figure 49. Perianal Swab, Males, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

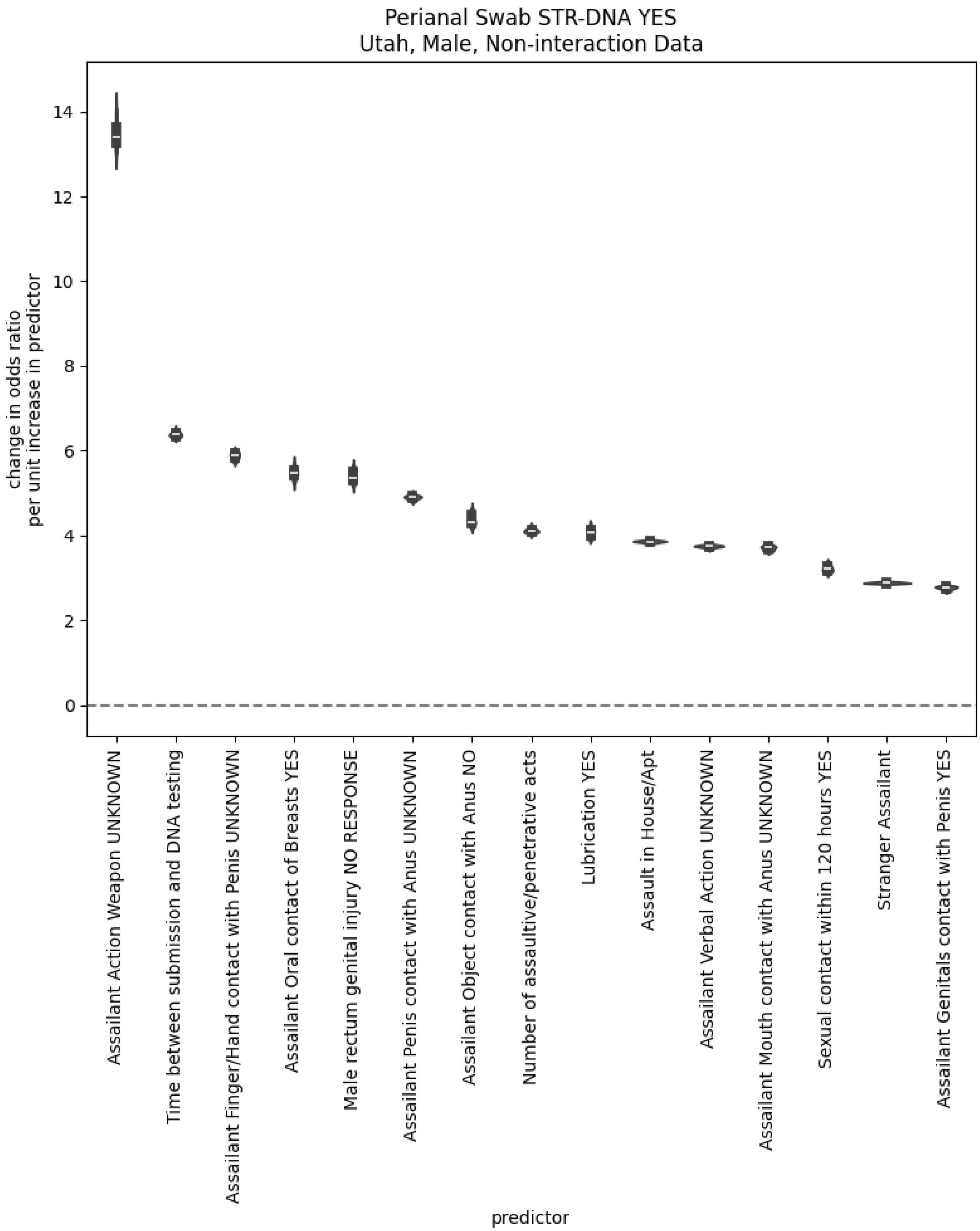


Figure 50. Perianal Swab, Males, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

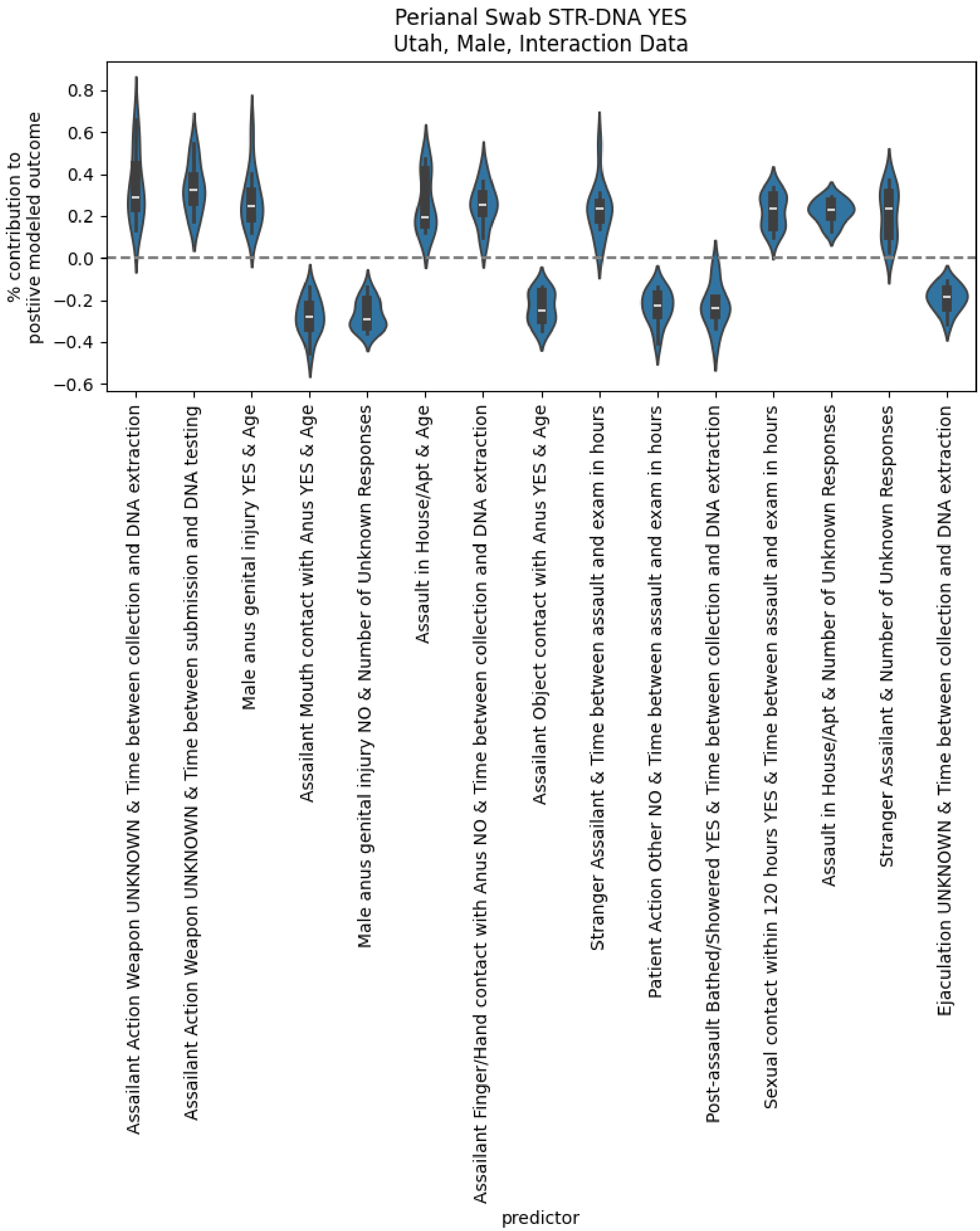
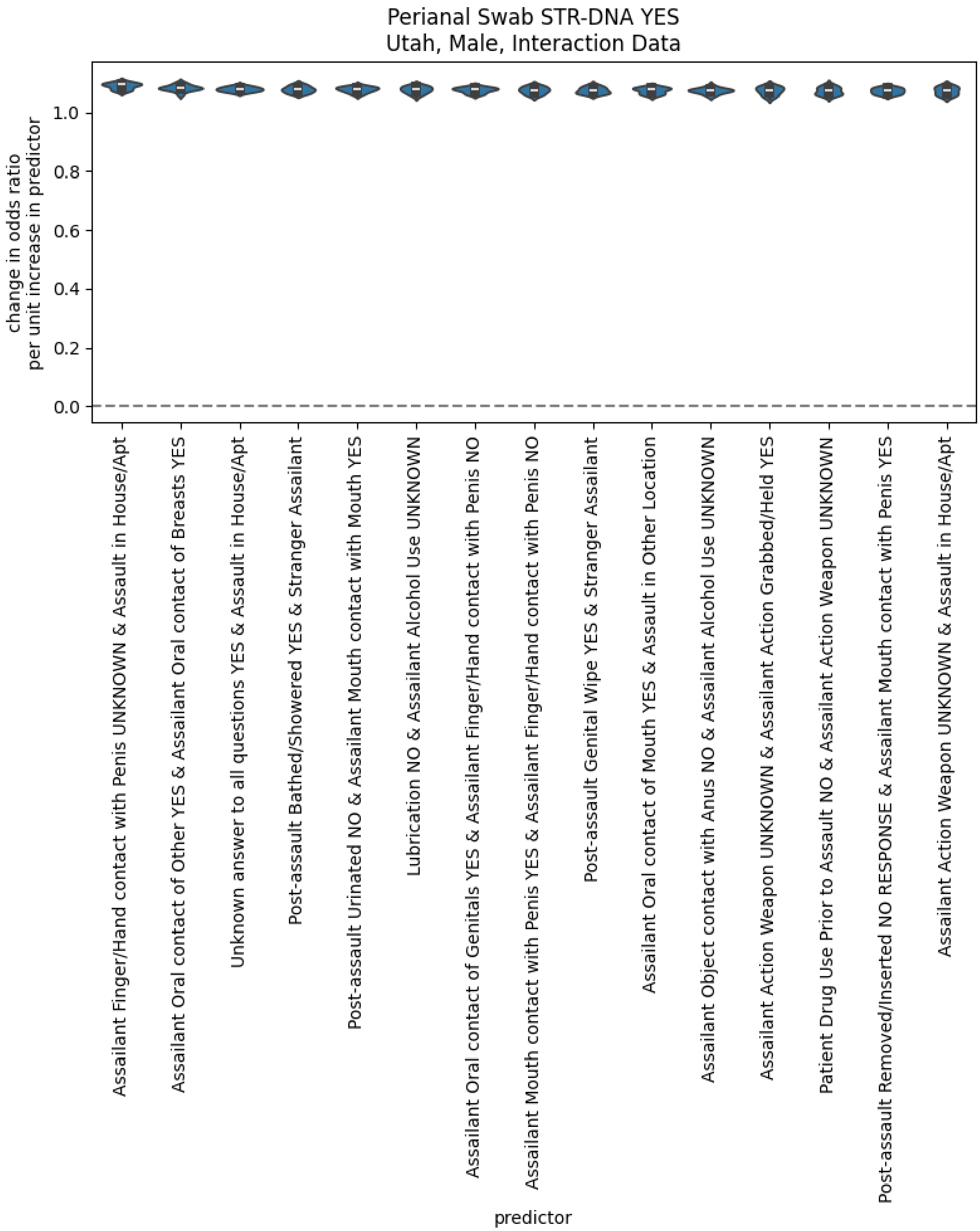


Figure 51. Perianal Swab, Males, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Perianal Swabs from Males

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the perianal swabs of males include the answer of "unknown" to several questions including weapon use, assailant finger/hand contact with penis, and contact with assailant penis on genitals; higher number of assaultive/penetrative acts; and use of lubrication.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Breast(s) Swabs (n=1063), Females only

Figure 52. Breast(s) Swab Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

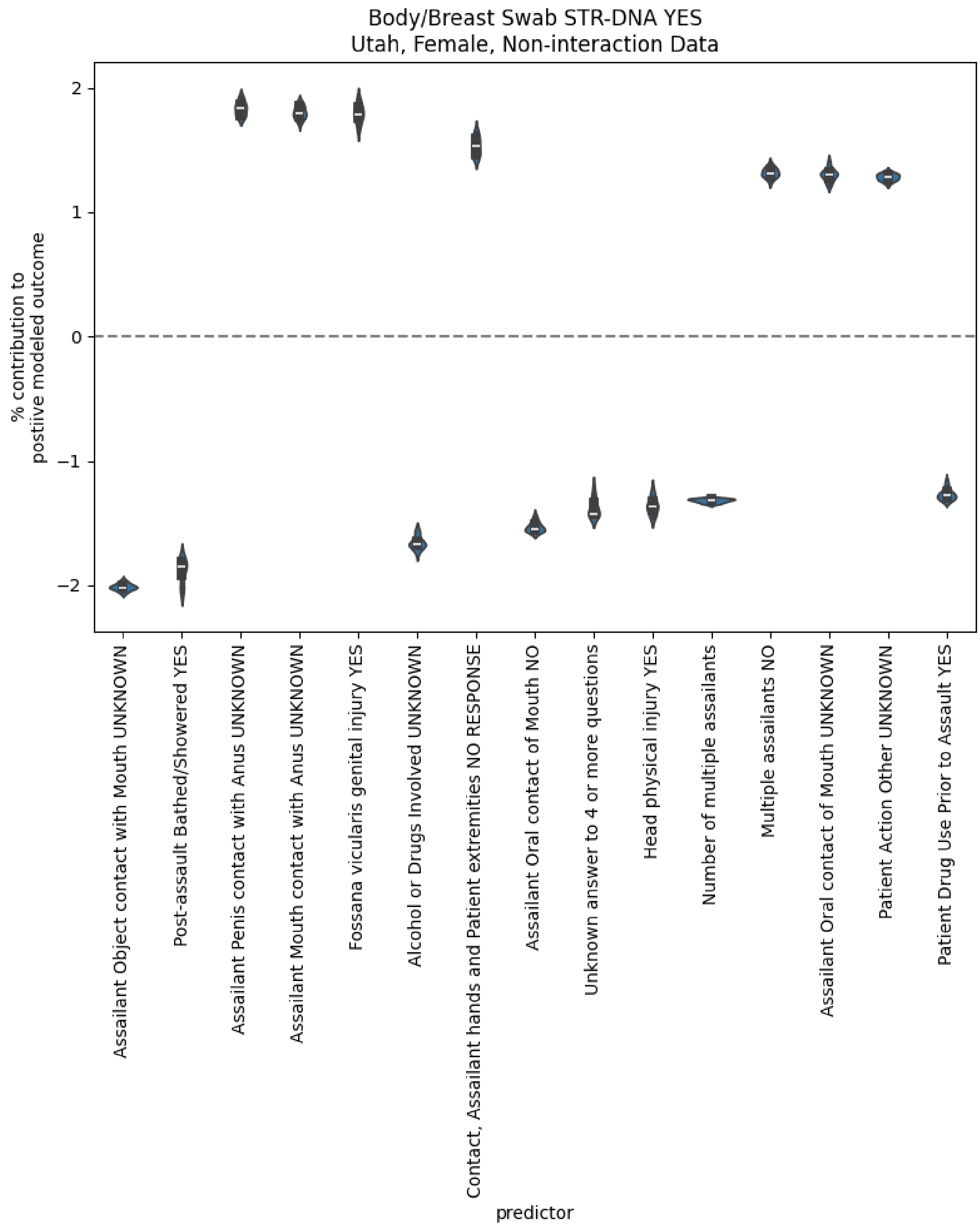


Figure 53. Breast(s) Swab Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

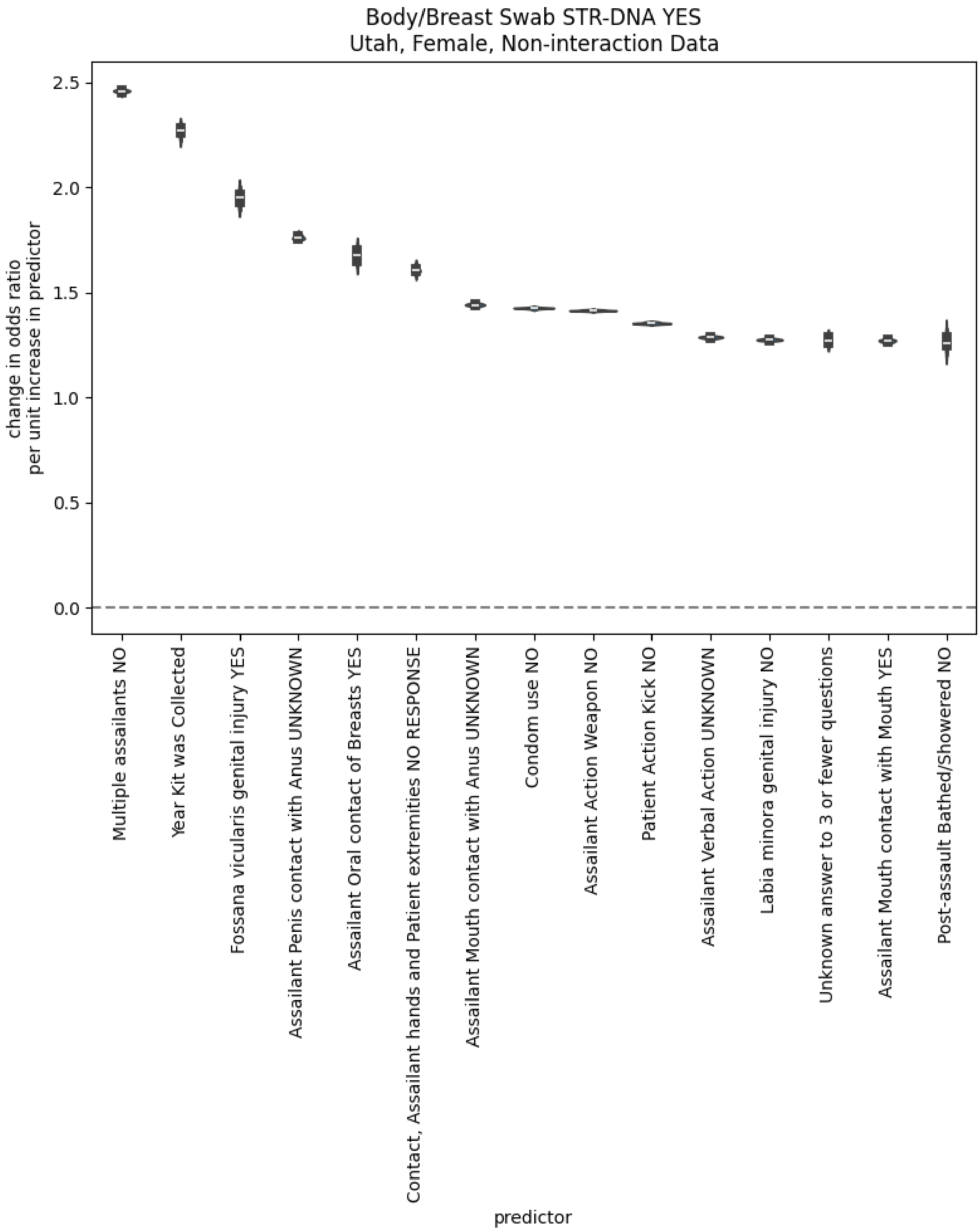


Figure 54. Breast(s) Swab Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

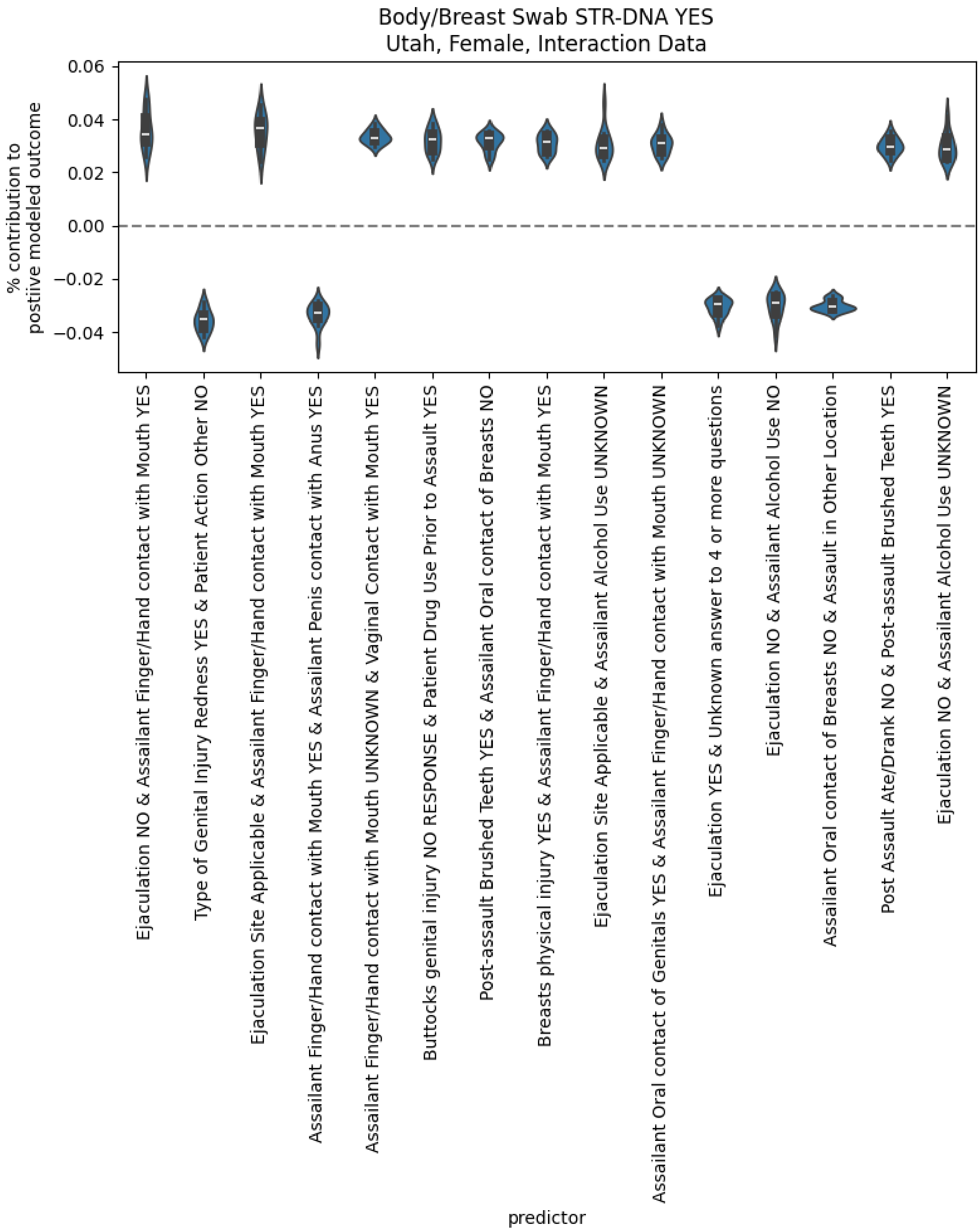
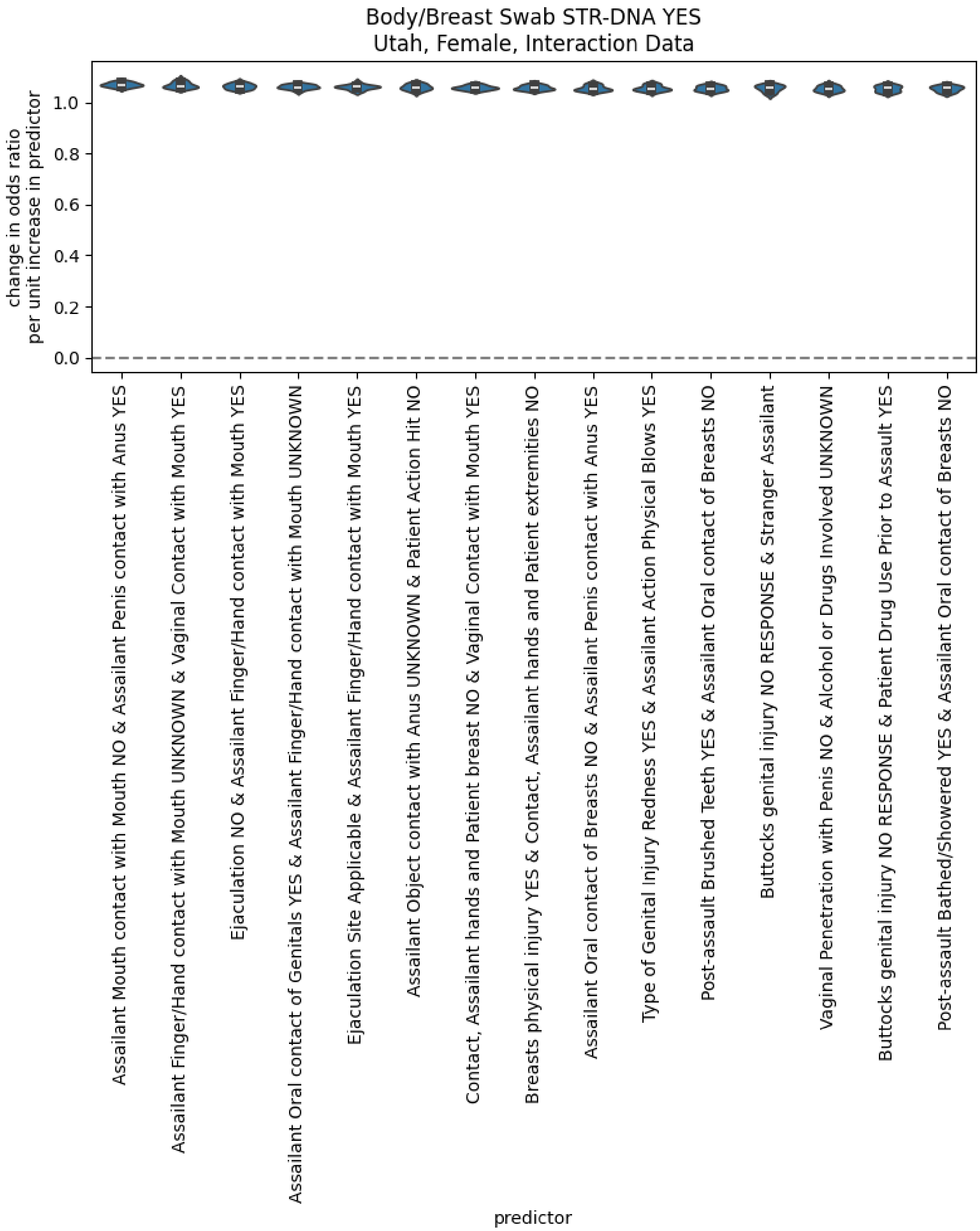


Figure 55. Breast(s) Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Breast(s) Swabs from Females

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the breast(s) swabs of females include not multiple assailants (indicating single assailant), year kit was collected, genital injury (fossa navicularis and labia minora), and assailant oral contact of breasts. Lack of post-assault bathing or showering was a significant predictor but a lower predictor in the model.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Neck Swabs (n=714)

Female

Figure 56. Neck Swab, Females, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

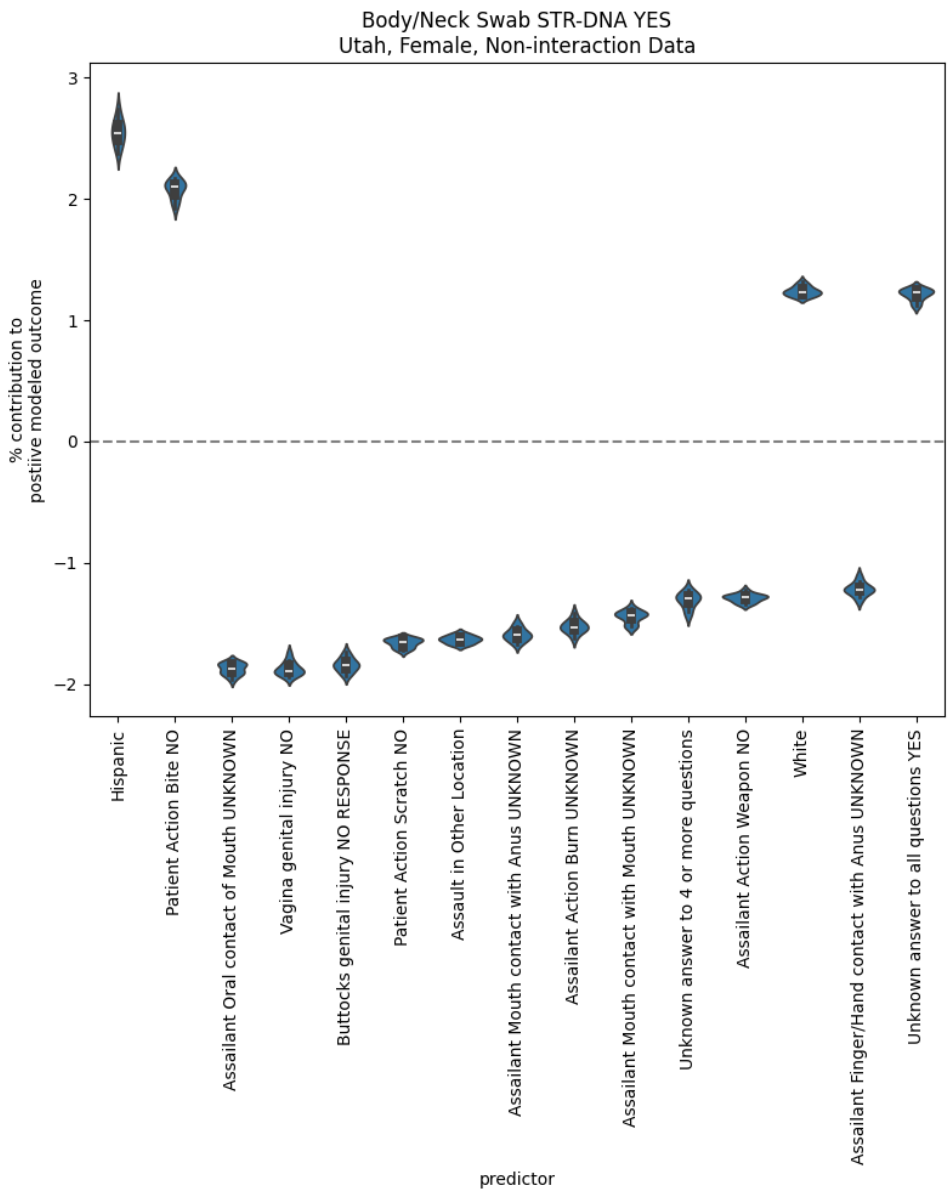


Figure 57. Neck Swab, Females, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

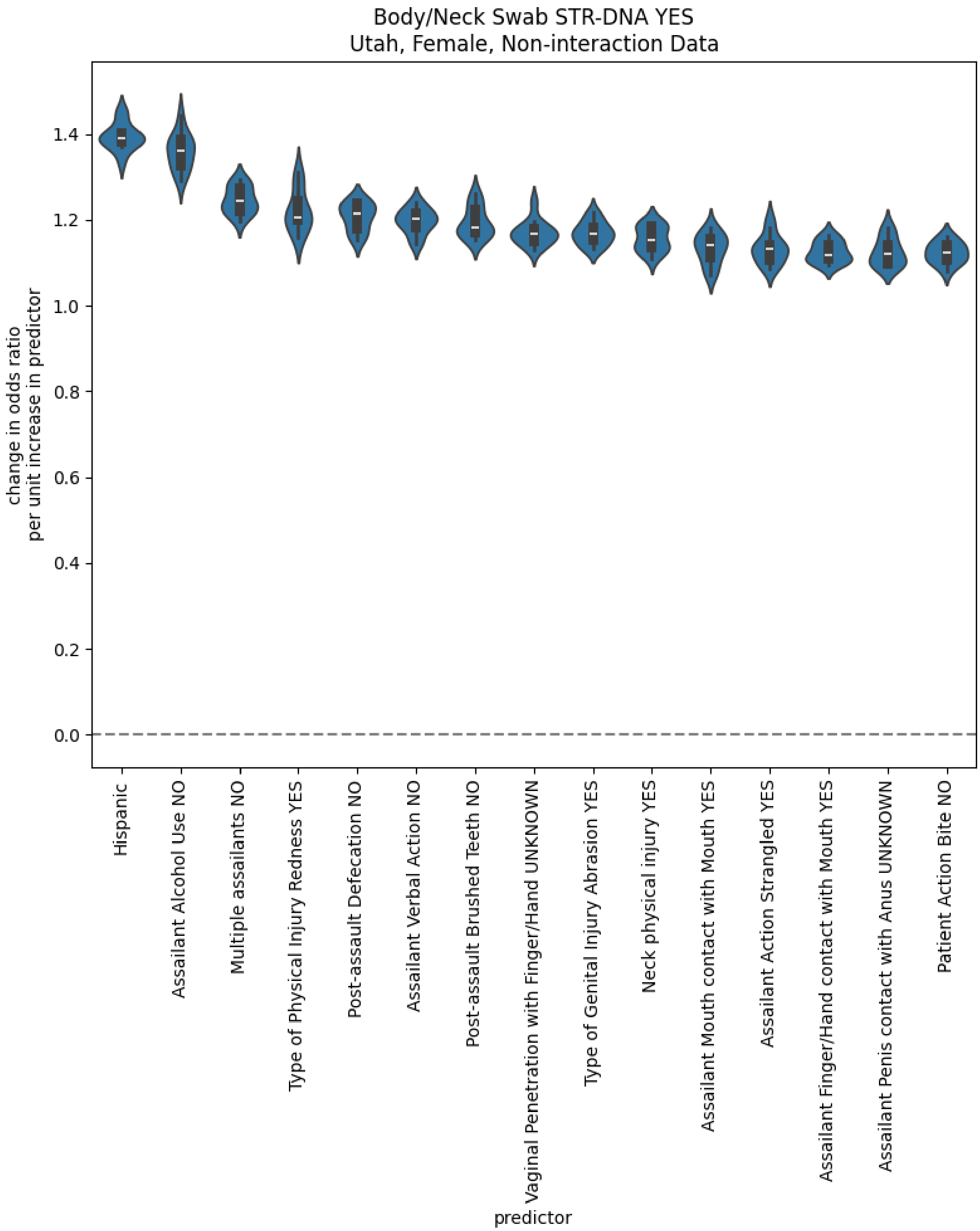


Figure 58. Neck Swab, Females, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

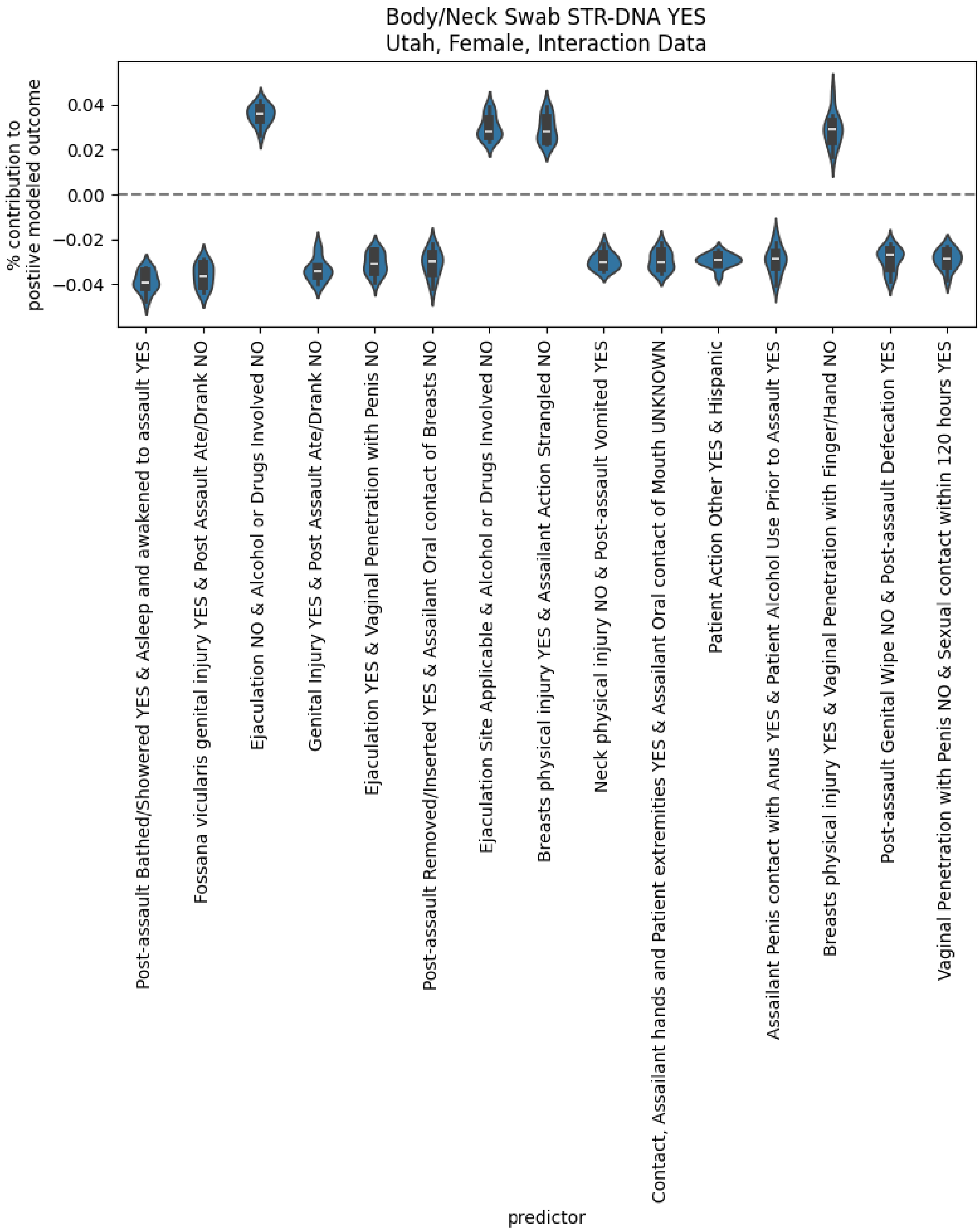
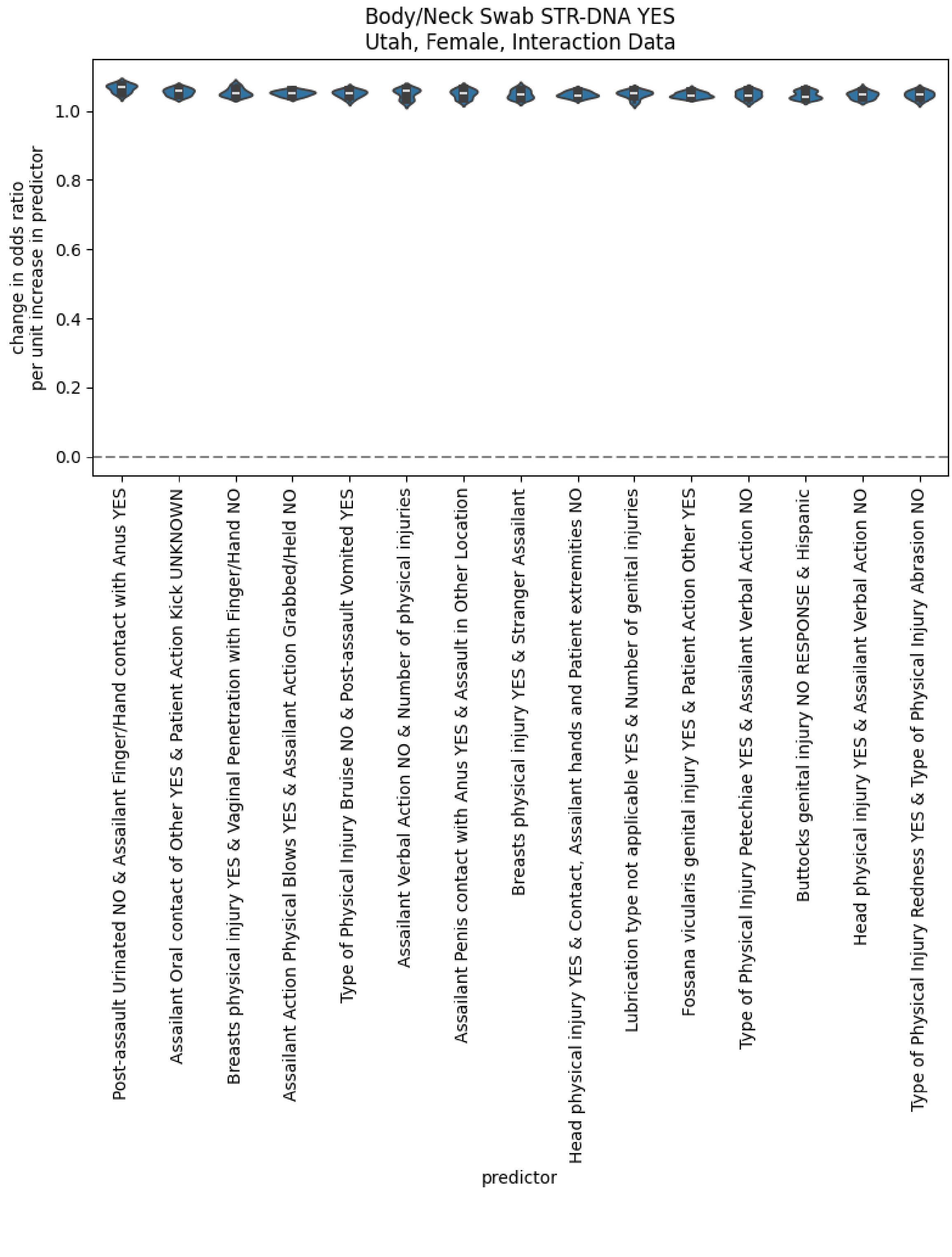


Figure 59. Neck Swab, Females, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Neck Swabs from Females

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the neck swabs of females include Hispanic race, lack of assailant drinking alcohol, single assailant, redness documented as a physical injury, lack of post-assault defecation, neck physical injury, mouth-to-mouth contact, not brushing teeth, and strangulation.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Males

Figure 60. Neck Swab, Males, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

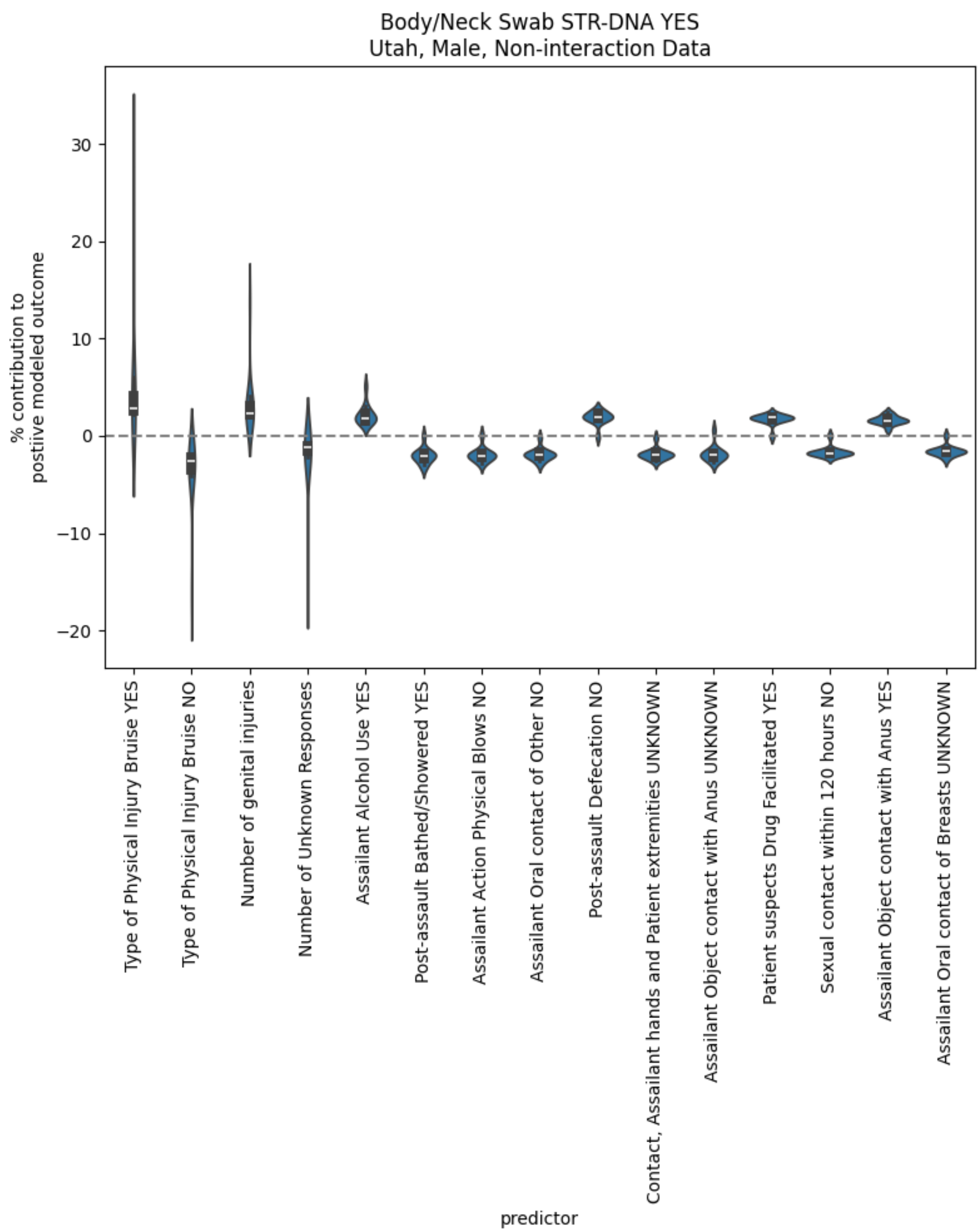


Figure 61. Neck Swab, Males, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

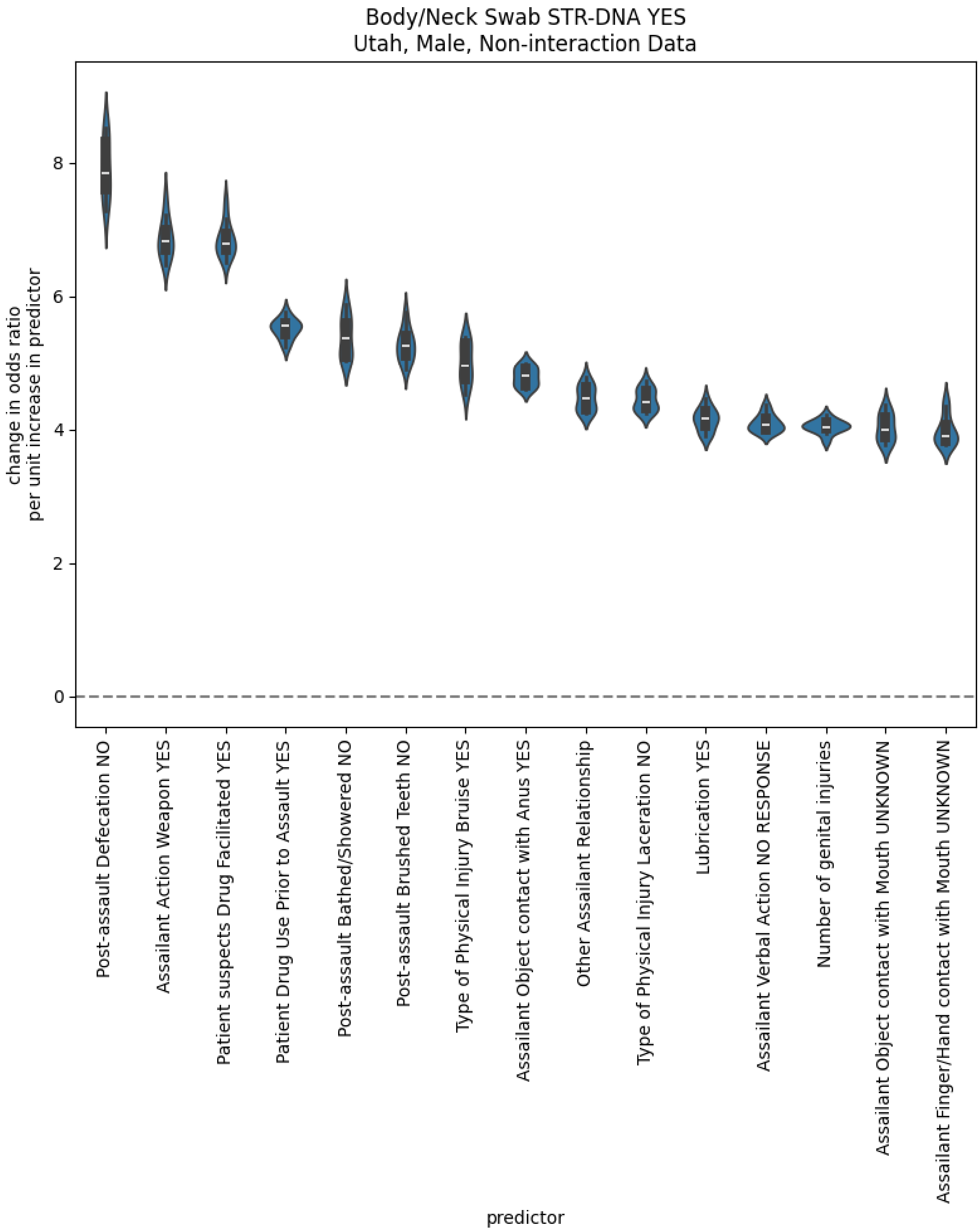


Figure 62. Neck Swab, Males, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

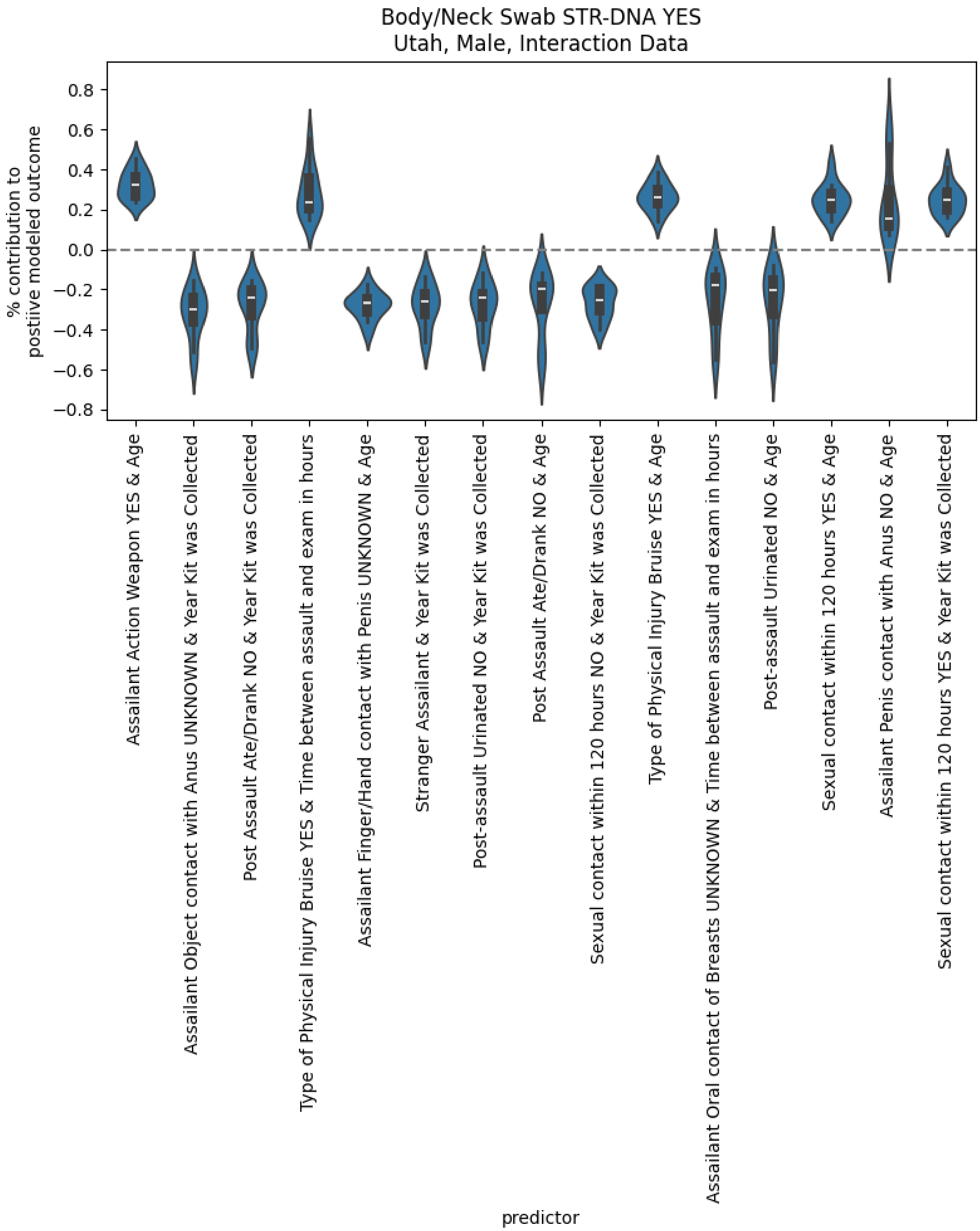
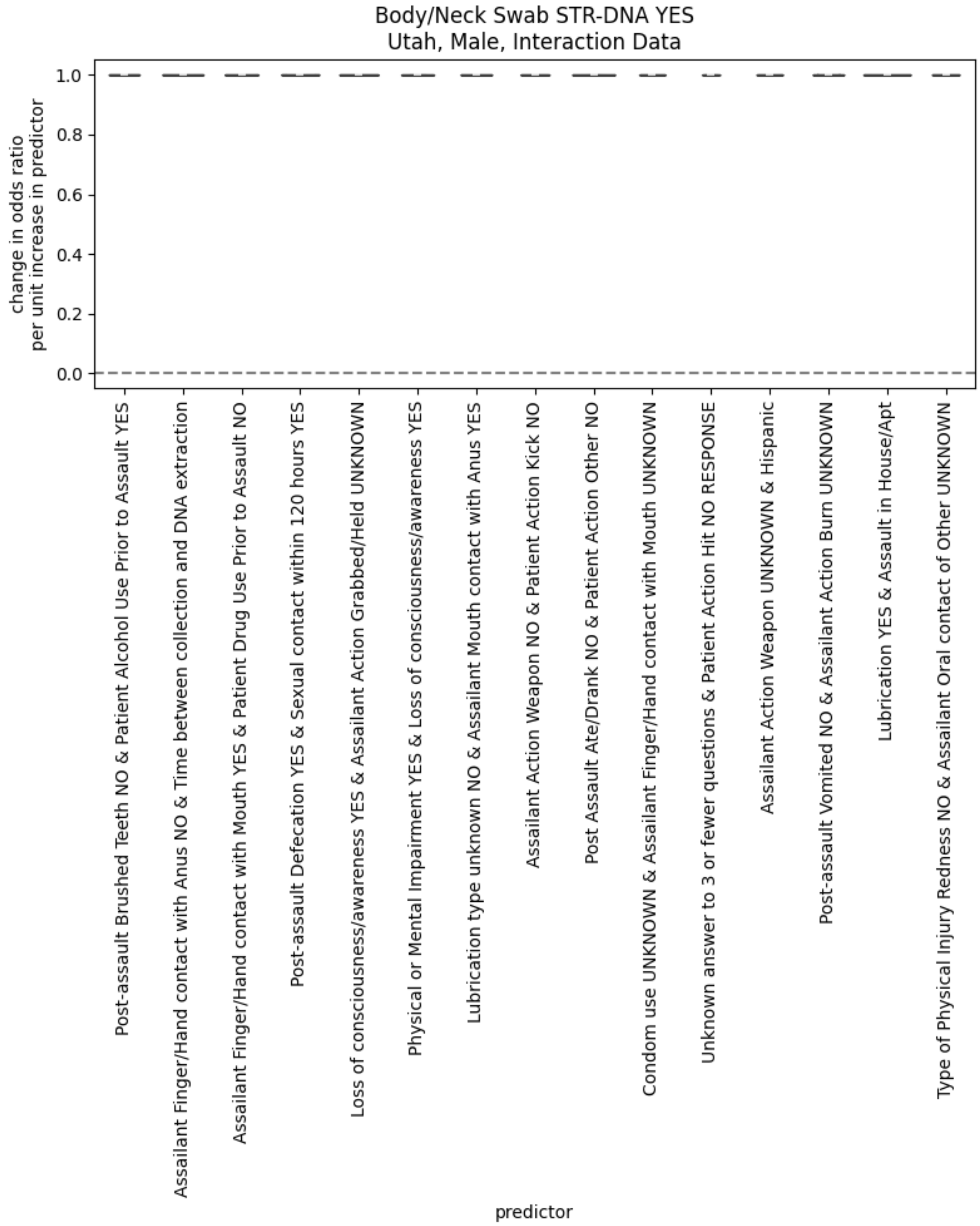


Figure 63. Neck Swab, Males, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Neck Swabs from Males

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the neck swabs of males include lack of post-assault defecation, weapon use in assault, suspected drug-facilitated sexual assault, victim drug use, lack of post-assault bathing/showering and brushing teeth, and bruise as a documented physical injury.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Body swabs, not including Neck or Breasts (n=623)

Female

Figure 64. Body Swab, Females, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

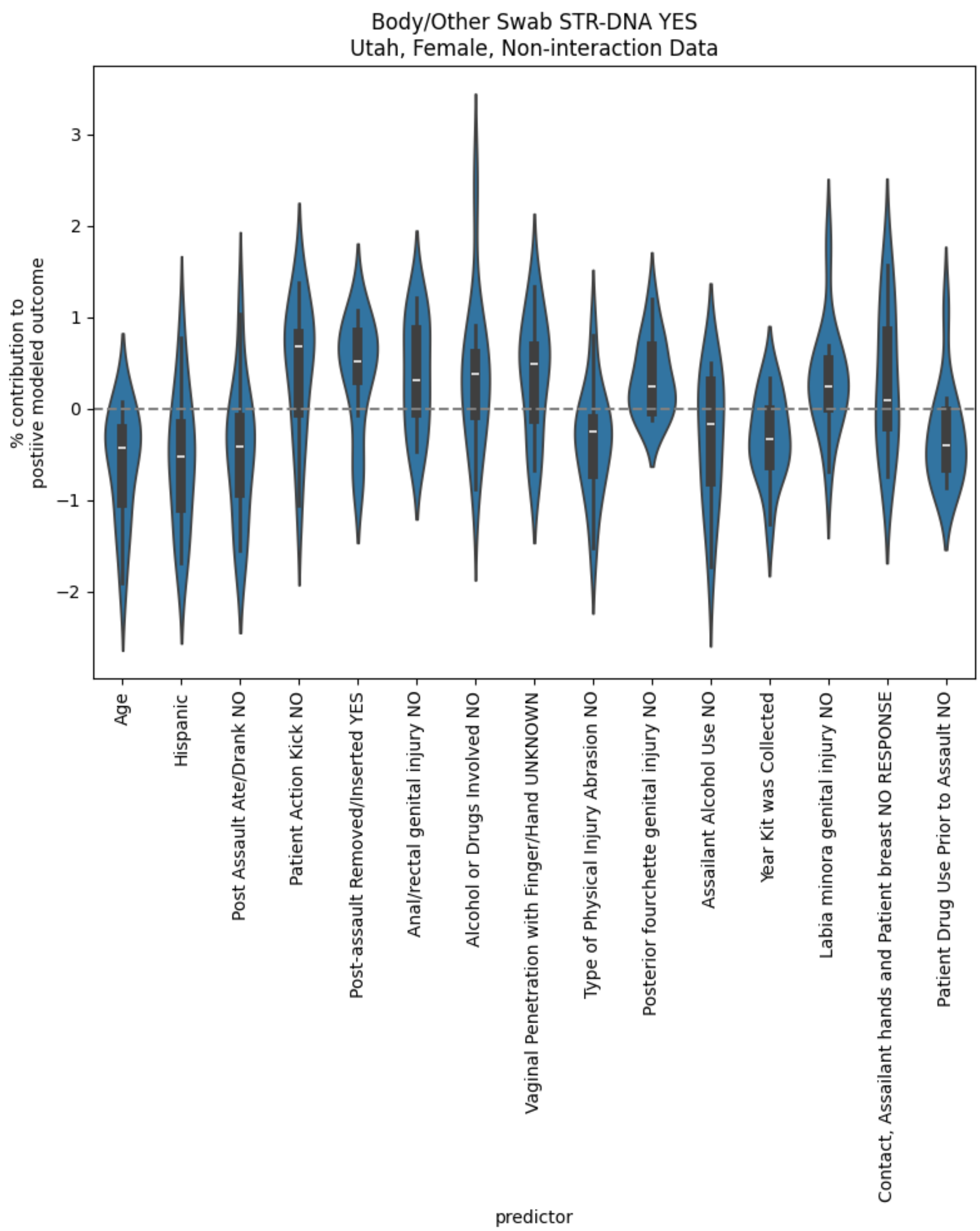


Figure 65. Body Swab, Females, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

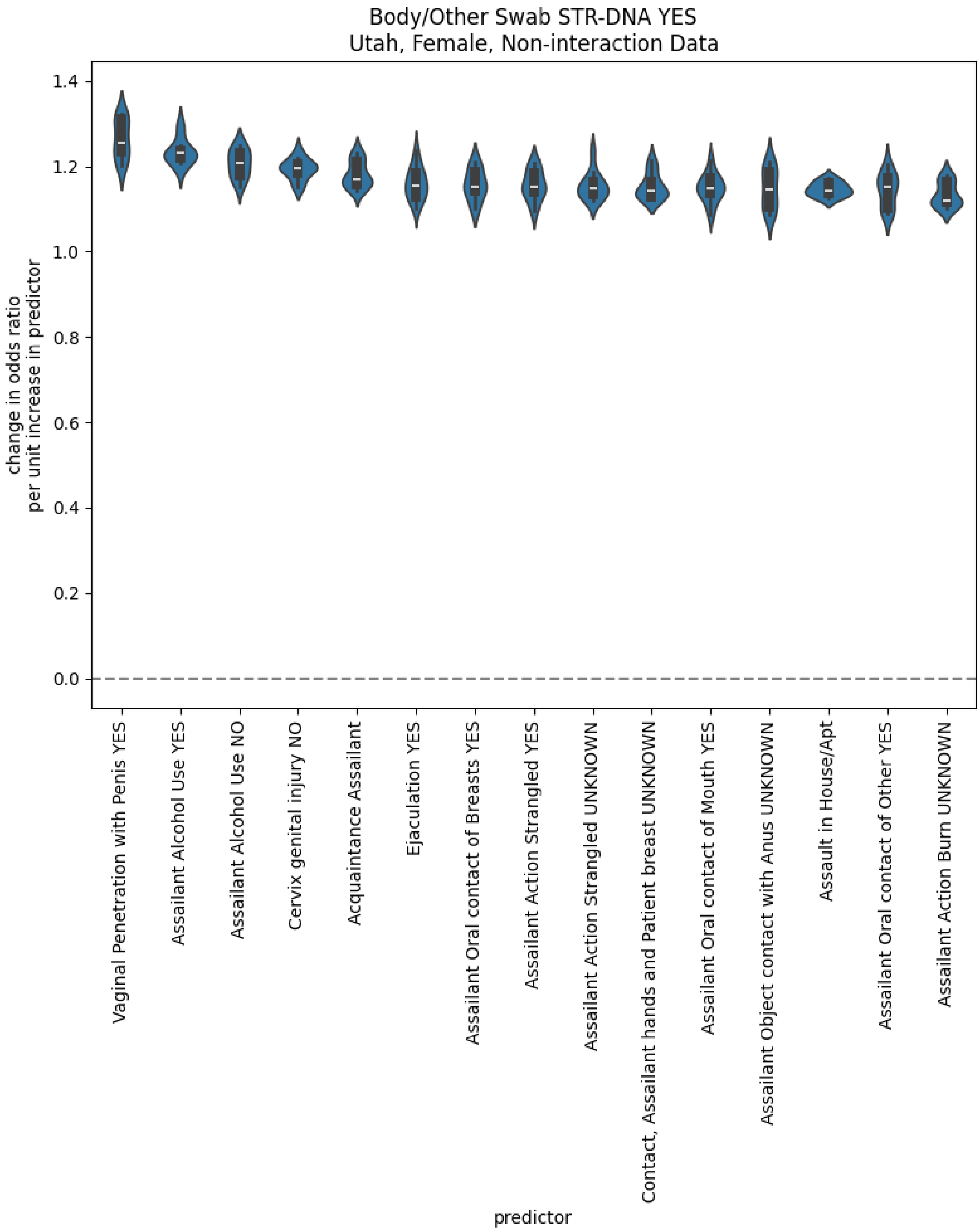


Figure 66. Body Swab, Females, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

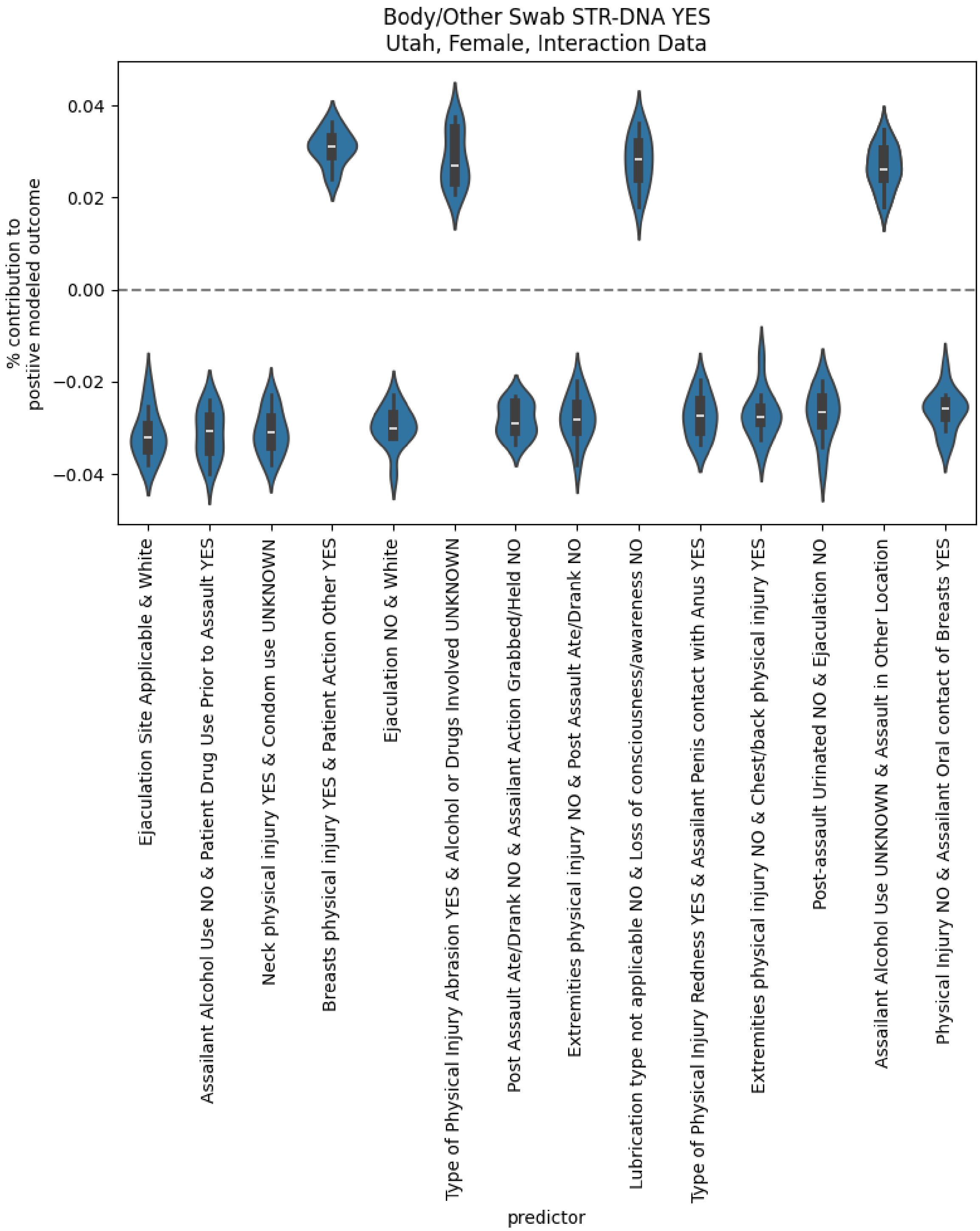
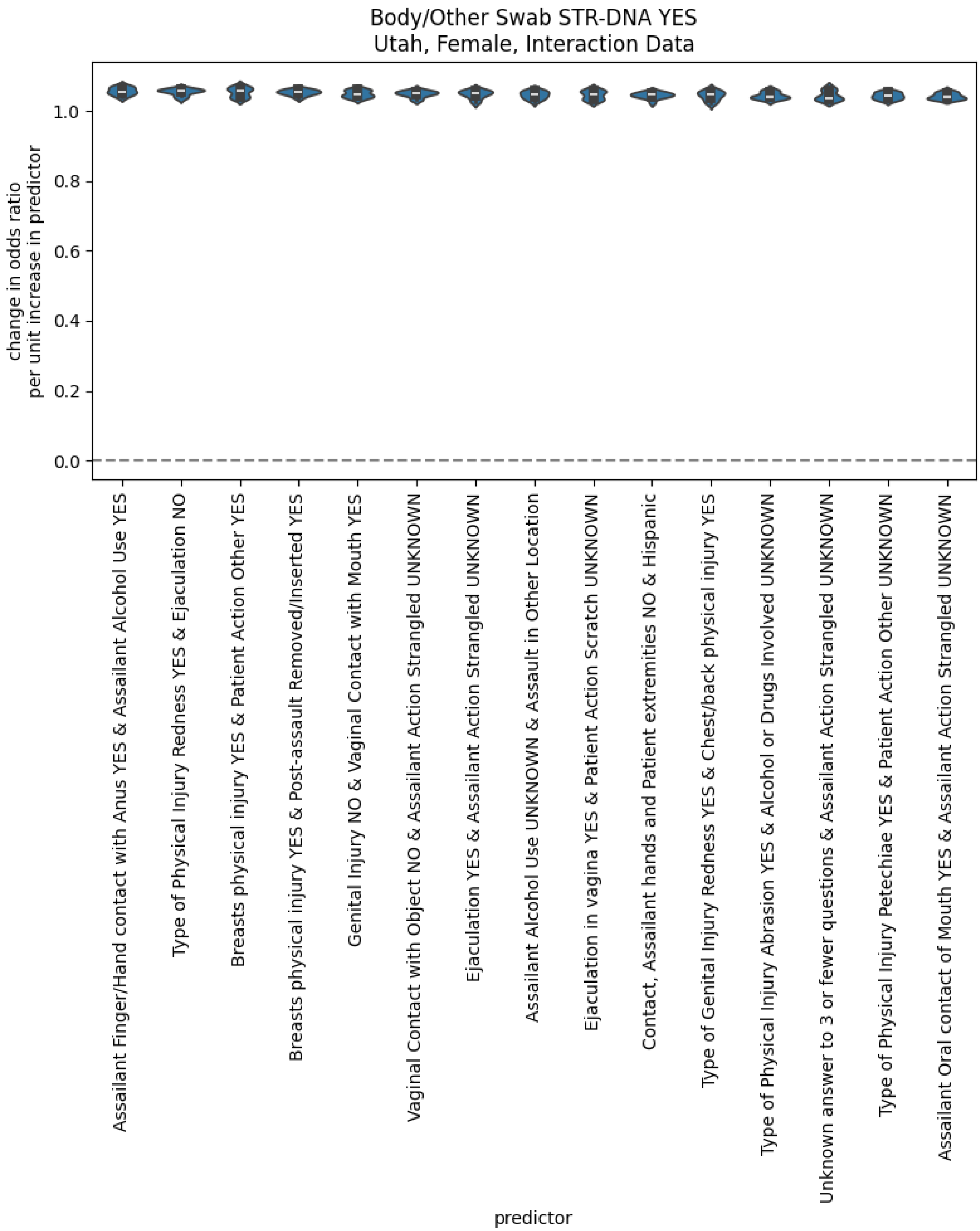


Figure 67. Body Swab, Females, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Body Swabs, not Breasts or Neck, from Females

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the neck swabs of females include vaginal penetration by penis; assailant alcohol use; lack of cervical injury; acquaintance assailant; ejaculation occurred; oral contact by assailant of breasts, mouth, and other body parts; and strangulation.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Males

Figure 68. Body Swab, Males, Normalized, Non-Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

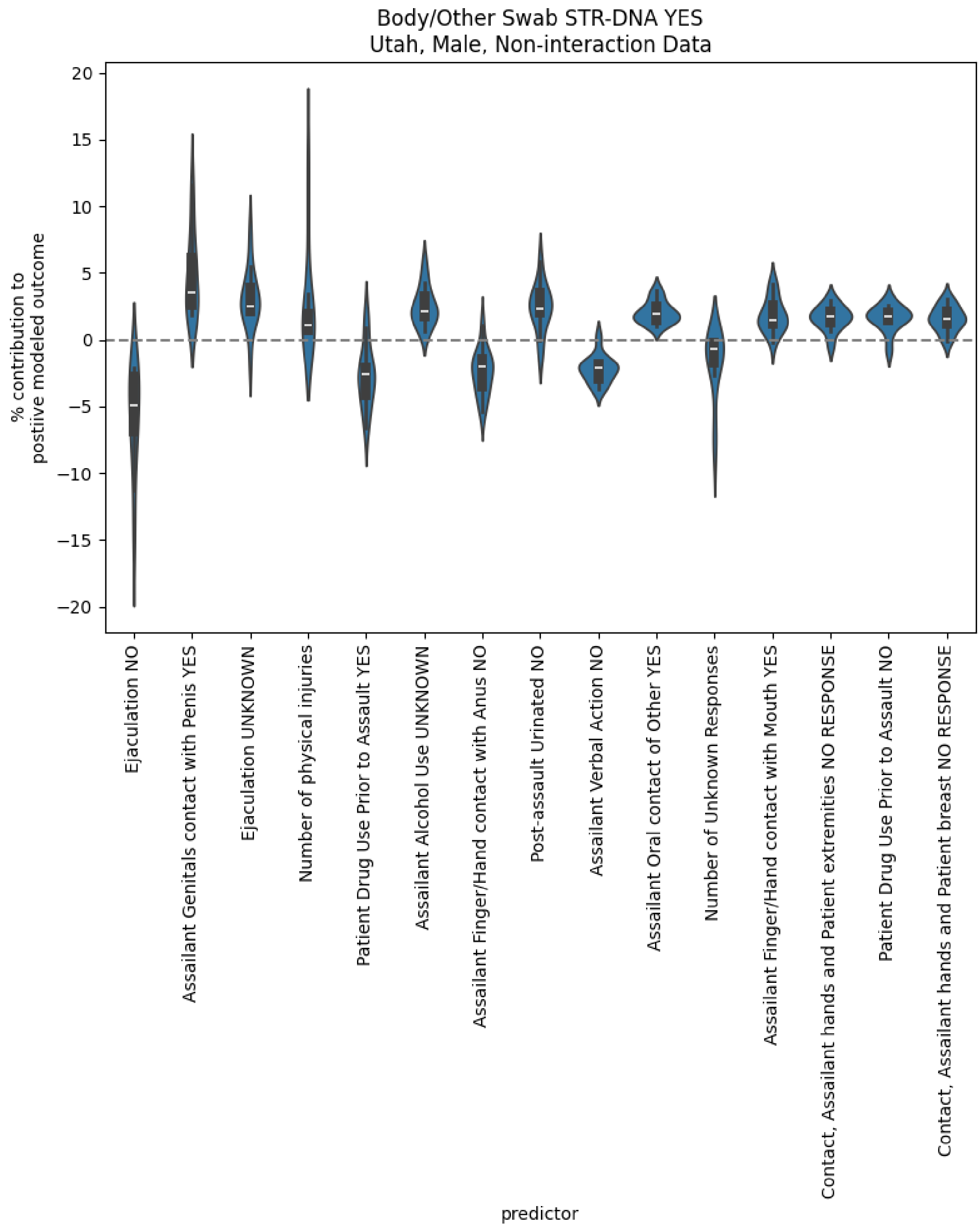


Figure 69. Body Swab, Males, Not Normalized, Non-Interaction Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

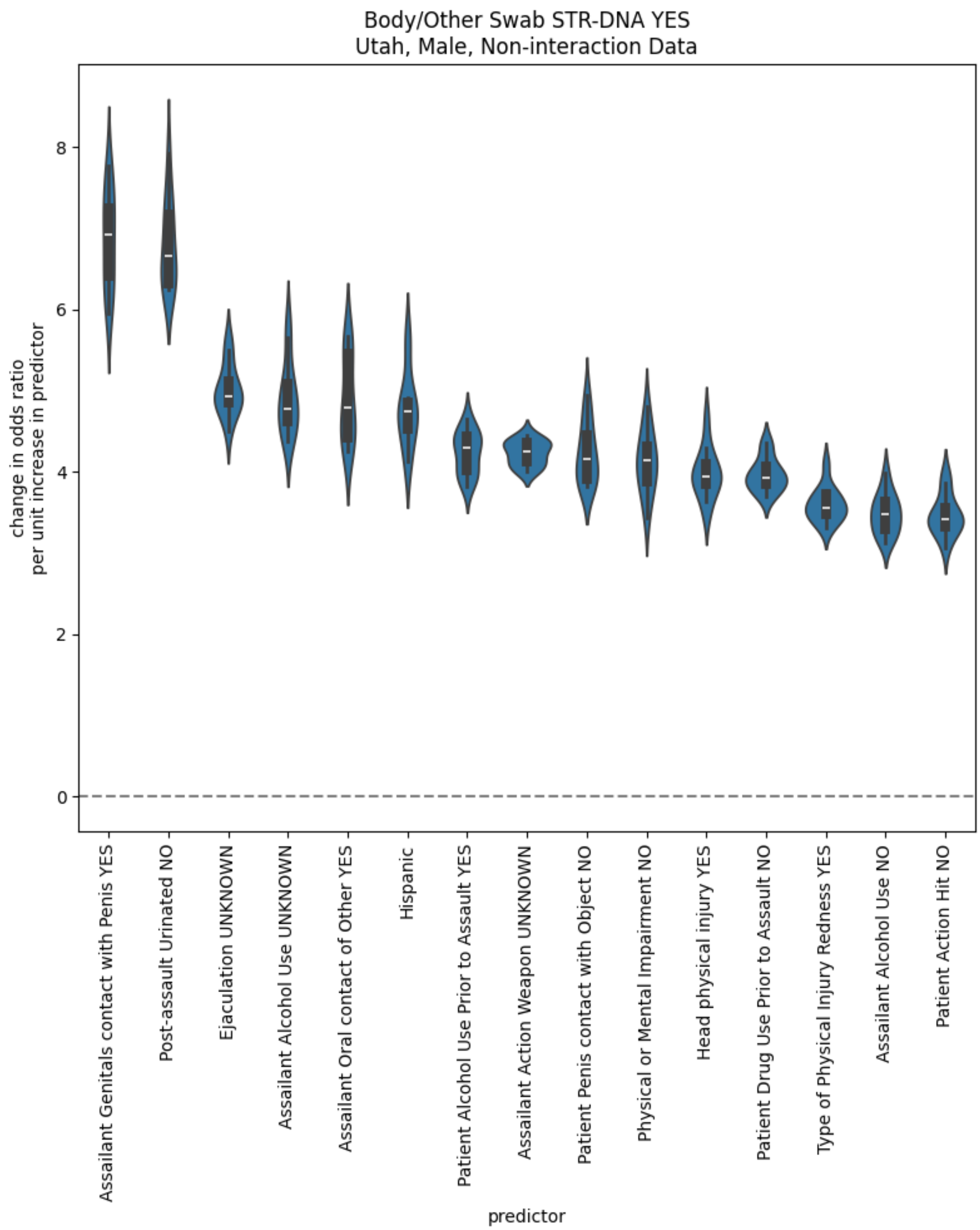


Figure 70. Body Swab, Males, Normalized, Interaction Percent Contribution to the Model Decision-Making of Development of Full/Partial STR DNA Profile of Foreign Contributor(s)

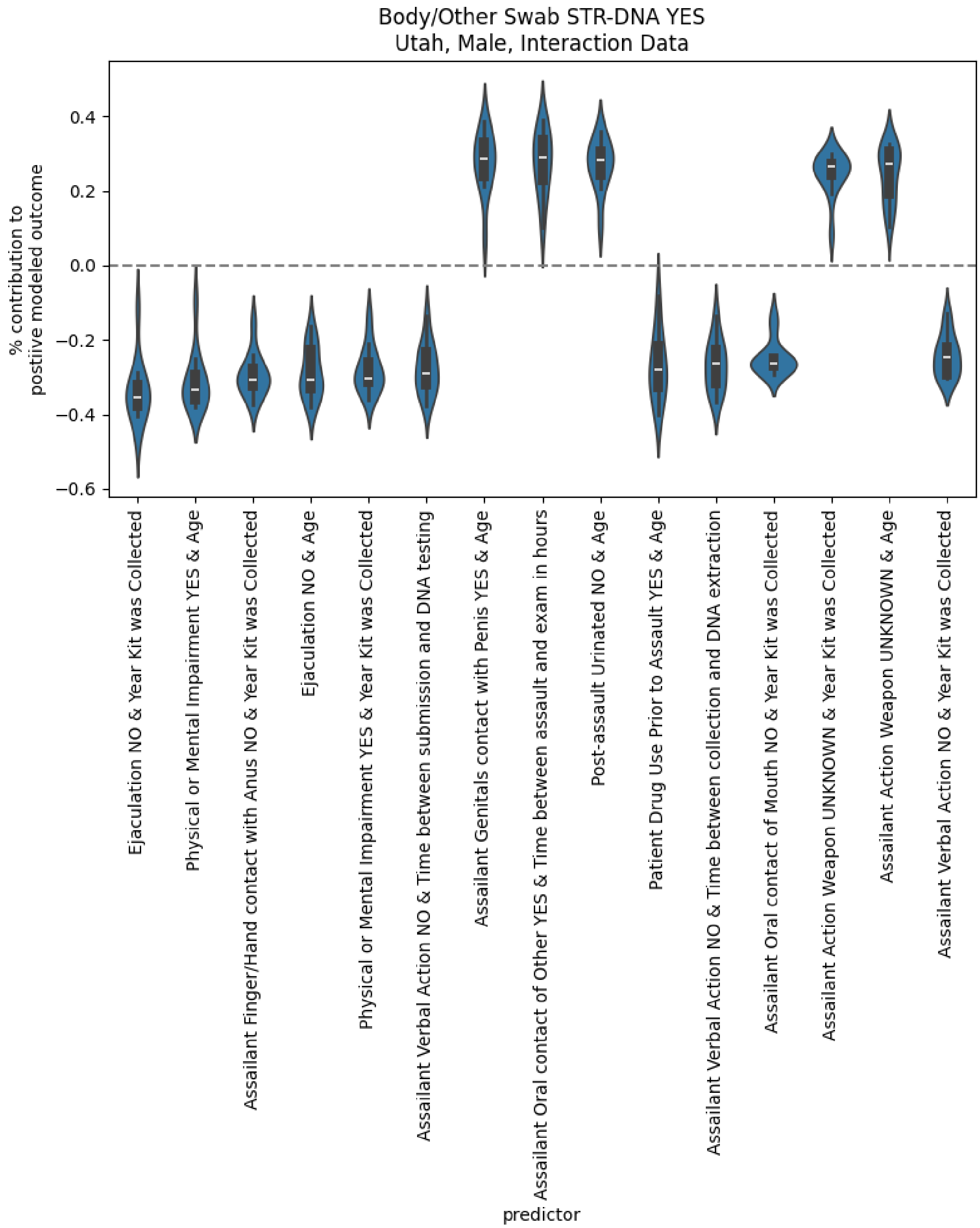
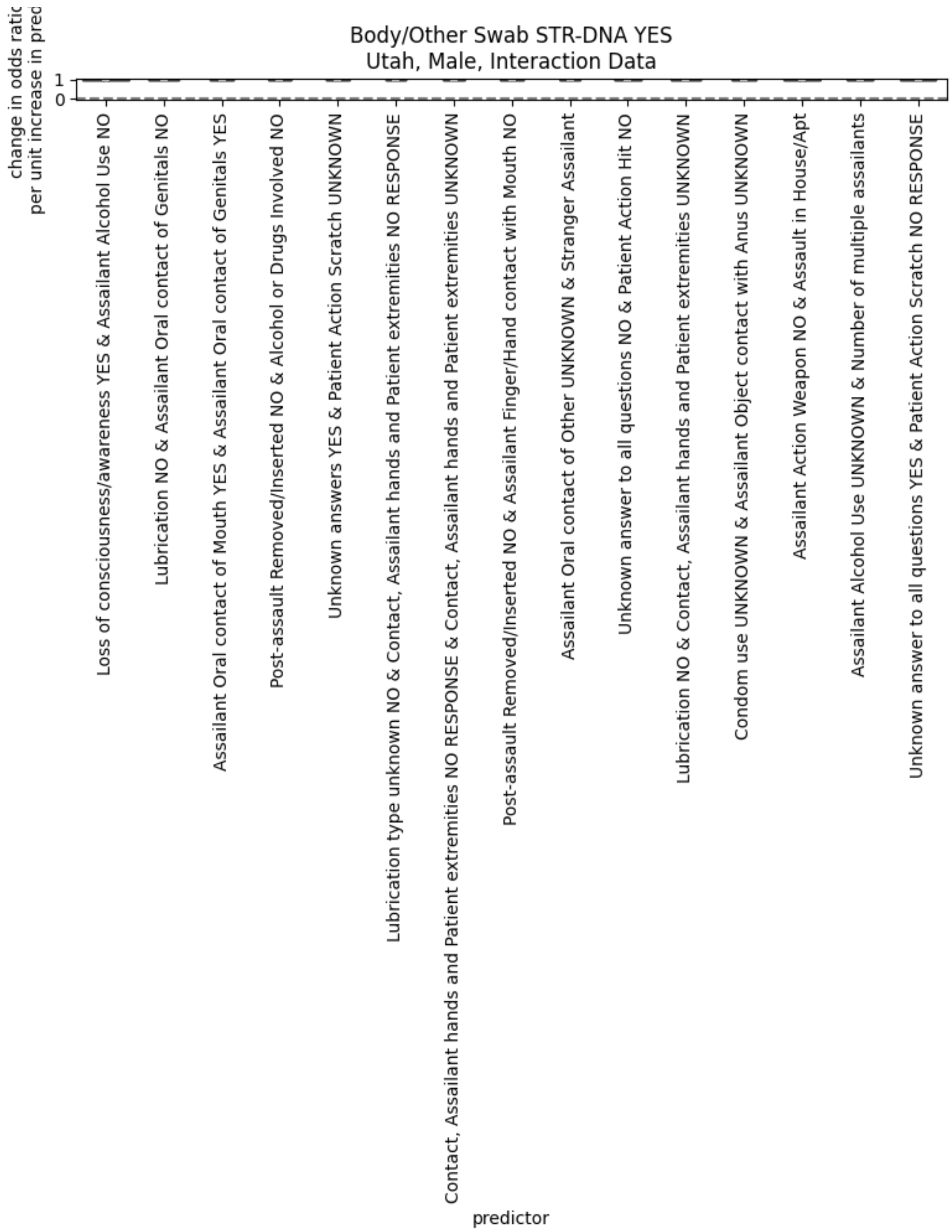


Figure 71. Body Swab, Males, Not Normalized with Interactions Change in Odds Ratio of Predictors for Development of Full/Partial STR DNA Profile of Foreign Contributor(s)



Summary of Key Findings from Body Swabs, not Breasts or Neck, from Males

In summarizing the models, variables significant in predicting the development of full or partial STR DNA profile of foreign contributors from the neck swabs of males include assailant genital contact with victim's penis, lack of post-assault urination, unknown ejaculation, assailant oral contact of body parts, and Hispanic race.

The interaction models indicate that multiple variables have relationships that can improve the accuracy of the model predictions. A benefit of utilizing machine learning logistic regression is that these interactions can inform swab selection.

Research questions under Part 2 of the study:

Research question #5: What is the reliability and validity of the Sexual Assault Kit evidence Machine Learning Model (SAK-ML) software program in predicting STR DNA profiles entered into CODIS using retrospective data?

> We assessed the reliability of the Sexual Assault Kit evidence Machine Learning Software (SAK-ML) using two methods: 1) by measuring the accuracy of the models, and 2) by measuring the "percent better than guessing the distribution mode." This second measure is intended to explain the extent to which the models were able to overcome the bias towards positive samples that was present in the data. For example, in some swabs, almost 90% of the available samples yielded positive STR-DNA profiles. Accordingly, a model which always output "YES" would have an accuracy of 90%, without capturing anything of the relationship between prediction features (patient age, race, action, etc.) and the probability of developing a positive STR-DNA profile. "Percent better than guessing the distribution mode" is therefore a measure of the extent to which information from the SAMFE actually predicts the development of a positive STR-DNA profile. As an example, a model with 100% accuracy on a dataset with 90% positive samples would be ~11% better than the "zero-context" model given by always guessing "YES."

Research question #6: Which method of selecting swabs from SAKs (forensic analysts determine which swabs to analyze and number of swabs, OR use of SAK-ML Model) yields a higher percentage of STR DNA profiles entered into CODIS?

We were unable to answer this question. To develop a machine learning model with improved accuracy in predicting which swabs to analyze to develop uploaded CODIS STR DNA profiles, all of the swabs from the SAKs would need to be tested. Additionally, thousands of SAKs would need to be included in a study of testing all SAK swabs.

The bulk of our data was from UBFS which tests selected swabs based upon the likelihood of developing meaningful DNA information. Due to this, the dataset was biased as the majority of data was from swab samples that were more likely to develop meaningful DNA information. An unbiased dataset would need all swabs tested to train or develop an accurate machine learning model. To develop a reliable model to predict the development of STR DNA profiles of foreign contributors per swab, a large dataset of SAKs for which all swabs were tested, regardless of the information in the SAMFE or the expertise of the forensic analysts, would be needed. Furthermore, definitive statements about the effectiveness of one approach over another are hampered by statistical power; for some swabs, information was available for only a few dozen patients. This limited sample size precluded separating the data into a "train/test" split that is common for validation in machine learning contexts (hence the decision to use logistic regression instead of data hungry and black-box models such as random forests). However, the capacity of most models to have a positive "percent better than guessing distribution mode" suggests that the models were able to improve upon the selection process that caused the data to be biased towards positive samples. More validation would be needed in a cohort of SAKs for whom testing was completed on every swab, but the present result suggests that data-driven models, coupled with a human-in-the-loop decision process about swab testing, may improve the efficiency of testing processes.

Research question #7: What is the impact of using SAK-ML Model on the following outcomes: development of STR DNA profiles entered into CODIS, crime lab efficiency, and crime lab cost savings?

We did not explore this question as we did not launch a machine learning model in practice.

<u>Models Skills Comparisons for Utah Data on Females (Figures 72 – 79)</u>

Figures 72 – 79 represent evaluation of the accuracies of the models and "percent better than guessing" of model performance on Utah data of female victims. As explained previously, we trained two sets of models using different data massaging techniques. The first set used data normalized so that all values were between 0 and 1, which was useful for a certain type of parameter/coefficient explainability wherein the coefficients were normalized so their absolute value summed to 100 (denoted "sum to 100" in the following figures). The second set used un-normalized data, which, in the context of logistic regression, was useful for interpretation of the exponentiated coefficients as a change in odds ratio per unit increase of each variable (denoted "exponentiated" in the following figures). In each of these instances, we also trained models using two sets of data: data that included the original features, and a second dataset that included the original features *as well as* the pair-wise interactions between features. Where the "individual terms" data (as indicated on the following figures) included roughly 200 features, the "interactions" data included tens of thousands of features. Enriching data in this way frequently allows for improved model performance, and this proved to be true in this context, as evidence by the following comparison plots between models trained on the 200 features ("individual terms") and models trained on the >40,000 features ("interactions").

The number of coefficients in a logistic regression model is equal to the number of features, and thus the “interactions” models were significantly more complex than the “individual terms” models. This is similar to the difference in complexity between logistic regression models on individual terms and random forests. However, the types of interactions and relationships between features learned in a “black box” random forest model are significantly harder to specify than the exhaustive list of pairwise interactions included in the “interactions” logistic regression models considered here. Given that the datasets for individual swabs frequently had too few samples to reasonably apply black-box machine learning models, the alternative of including a massive number of clearly specified interaction columns seemed a reasonable “white box” method for quantifying interacting relationships between variables. The comparisons between model skills using these various methods of data massaging (including interactions or not, normalizing the data or not) are plotted below.

As a final note, given the size of the “interaction” data, we opted to use a gradient descent-based optimization method called “ADAM” for solving for optimal model parameters. This is an optimization technique that is frequently used in training large models such as neural networks, since it scales well with the number of parameters. However, the process involves random initializations of the model parameters or coefficients, and, using repeated random samples of the data, updating the parameters to increase model fit. Because of stochasticity in this process, different runs of the training process can yield different model skills and coefficients. To quantify this variability, we ran 12 models on each of the 4 types of massaged datasets, and on each target feature (swabs and overall CODIS profile outcomes). The distributions of model accuracies, as well as the percent better than guessing the distribution (per the above discussion), are plotted below.

Figure 72. Accuracy Percent of Models on Female, Utah, Un-normalized Data

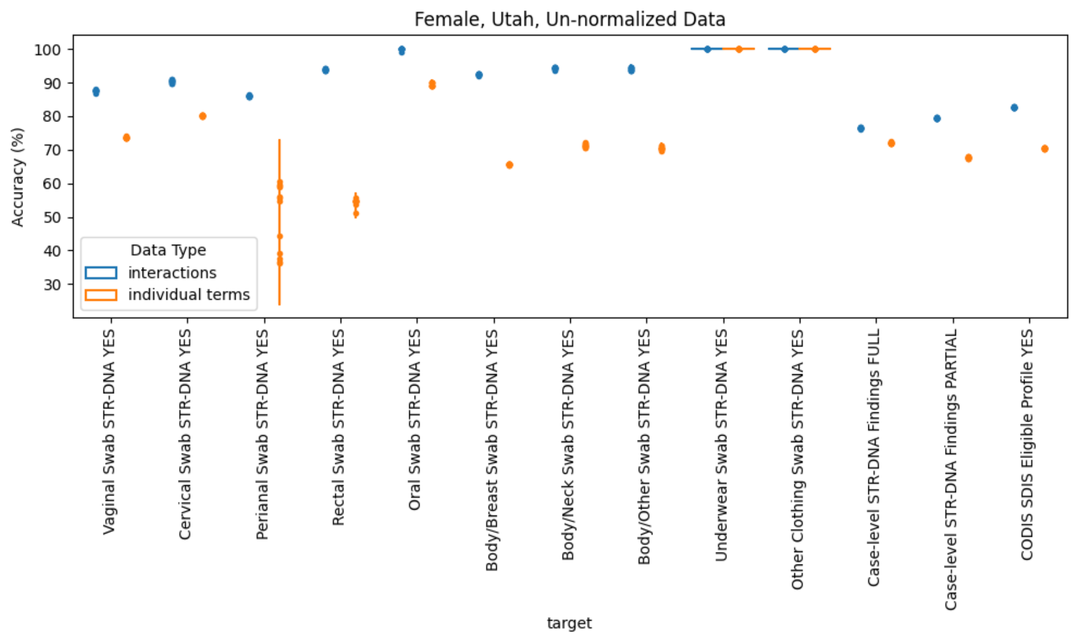


Figure 73. Accuracy Percent of Models on Female, Utah, Normalized Data

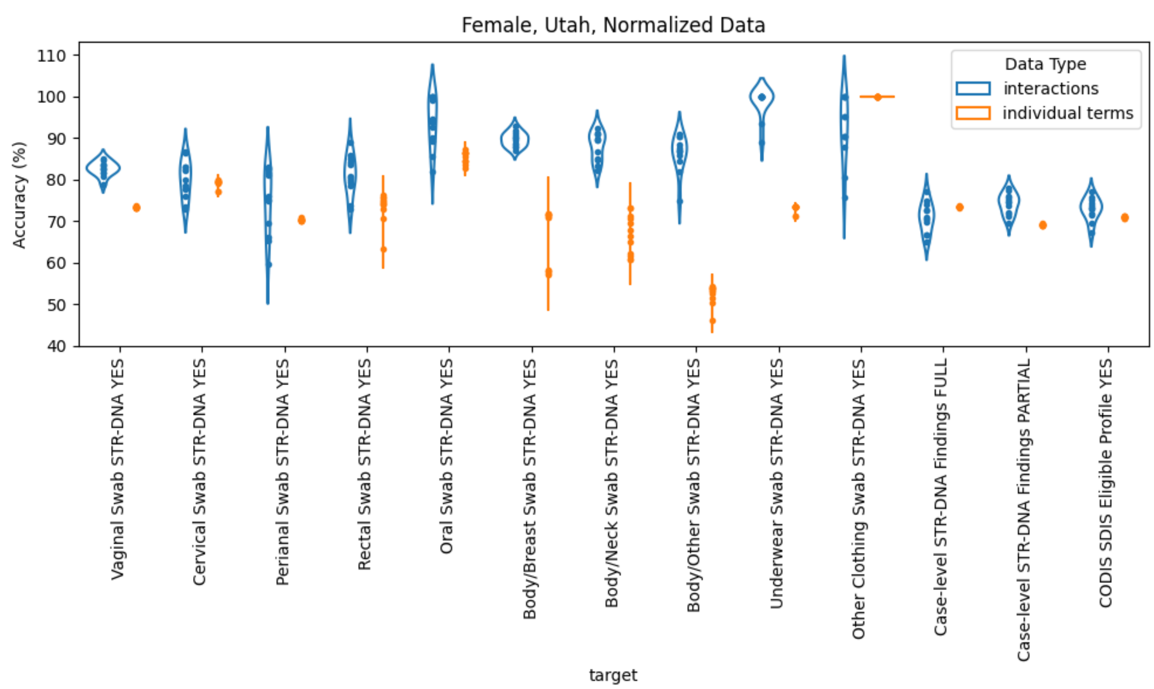


Figure 74. Percent Better than Guessing of Models on Female, Utah, Un-normalized Data

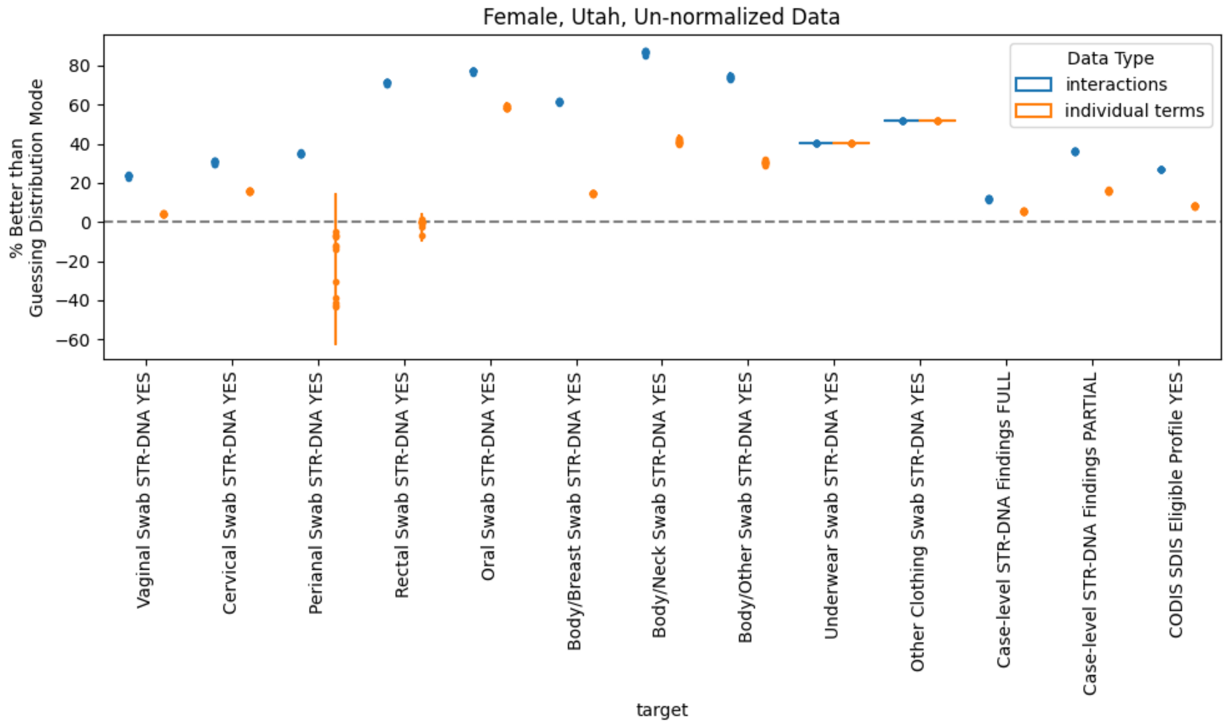


Figure 75. Percent Better than Guessing of Models on Female, Utah, Normalized Data

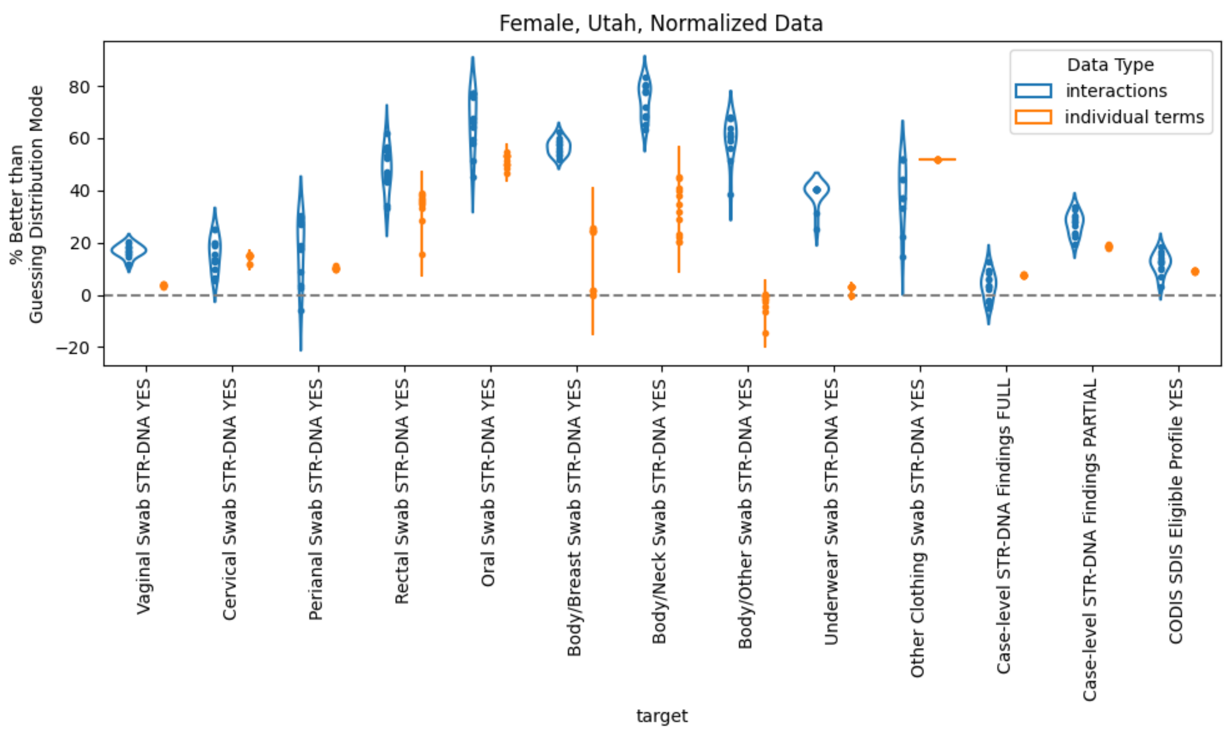


Figure 76. Accuracy of Models on Female, Utah, Interaction-Augmented Data

Figure 77. Accuracy of Models on Female, Utah, Non-Interaction Data

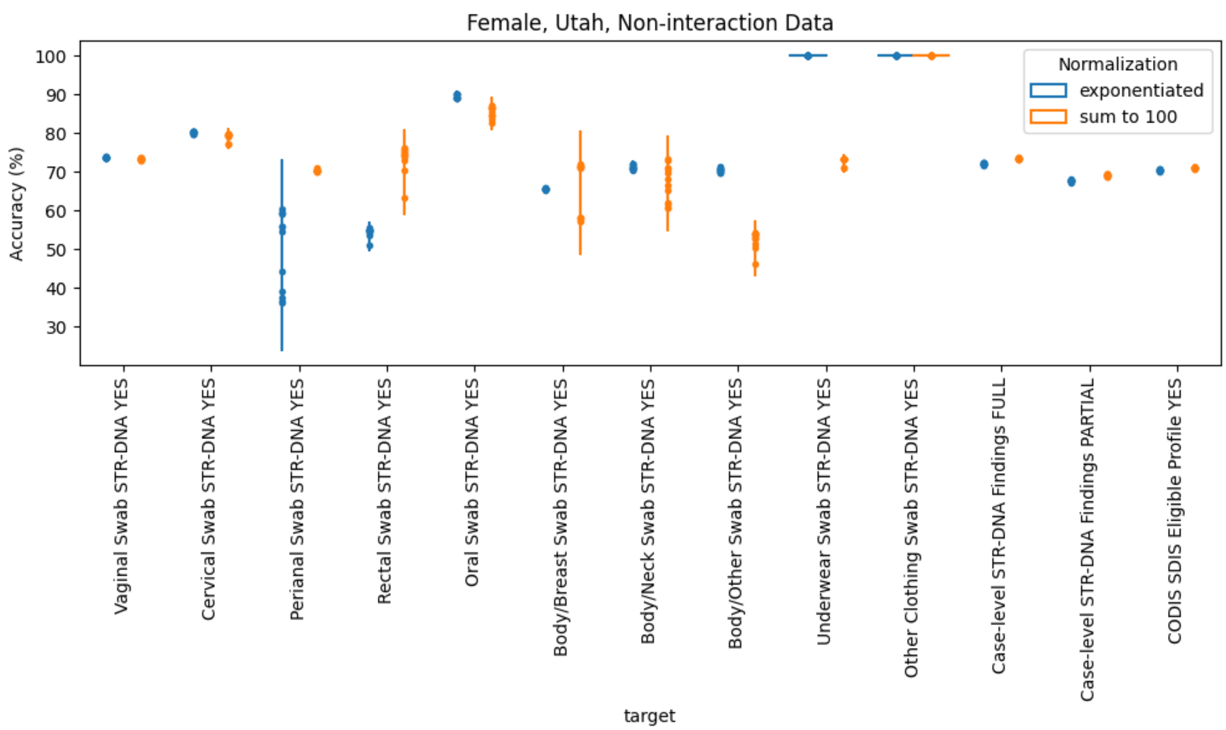


Figure 78. Percent Better than Guessing of Models on Female, Utah, Interaction-Augmented Data

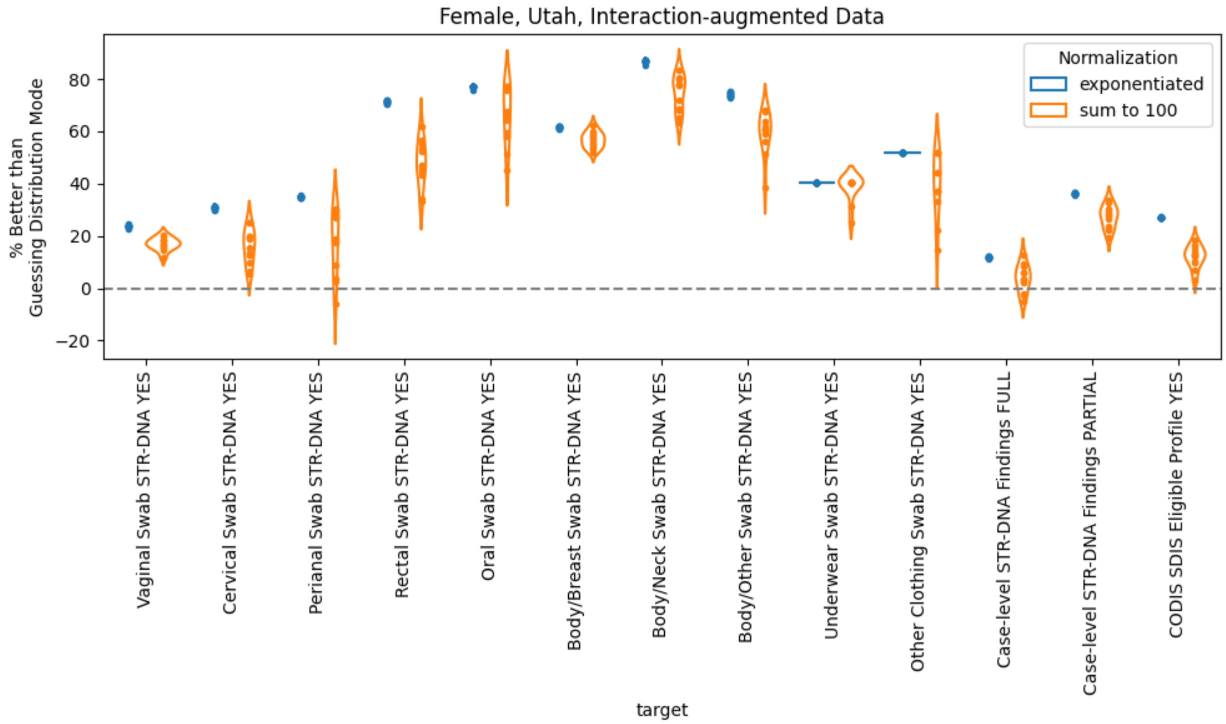


Figure 79. Percent Better than Guessing of Models on Female, Utah, Non-Interaction Data

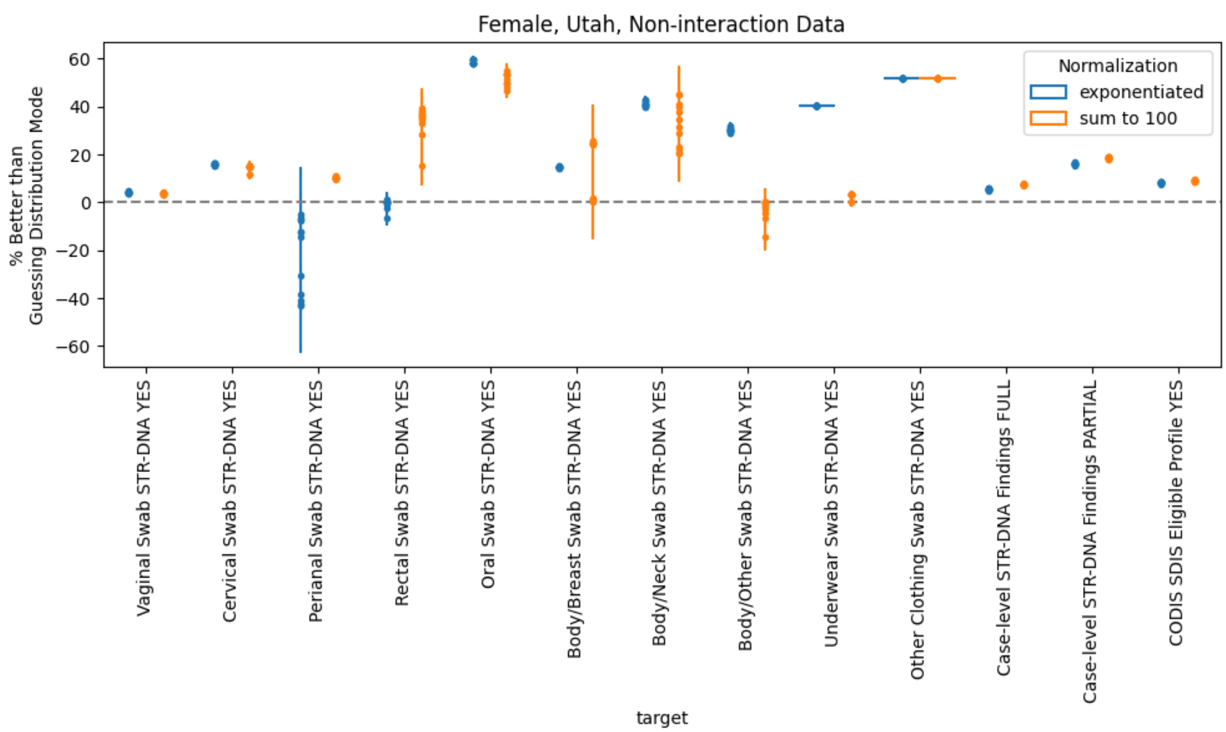


Figure 80. Accuracy Percent of Models on Males, Utah, Un-normalized Data

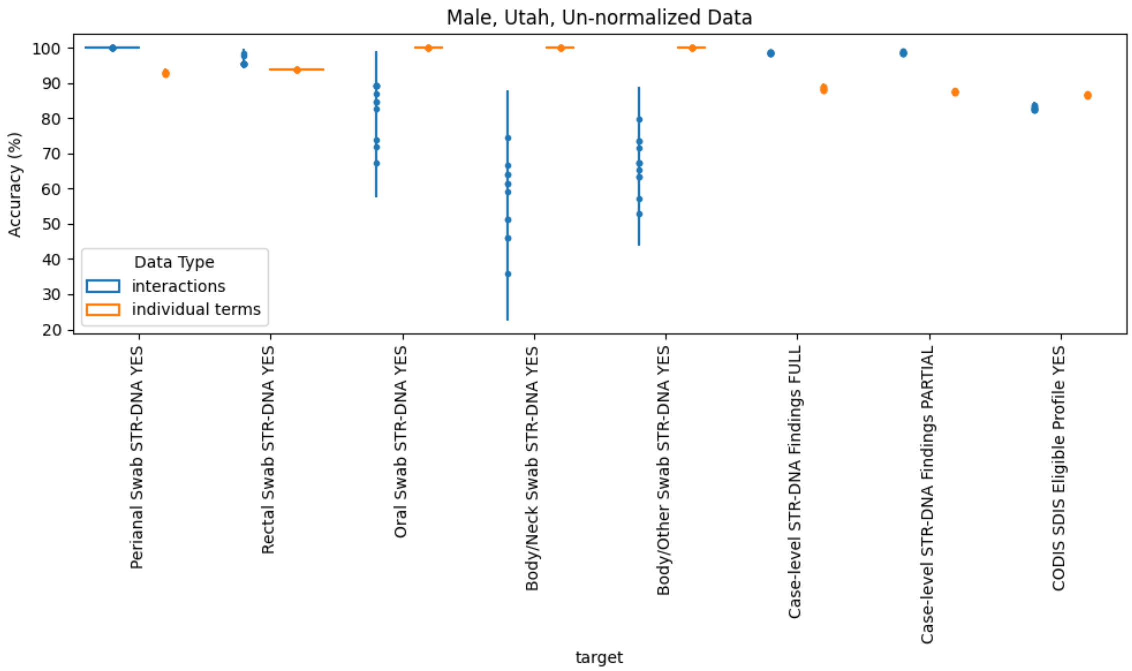


Figure 81. Accuracy Percent of Models on Males, Utah, Normalized Data

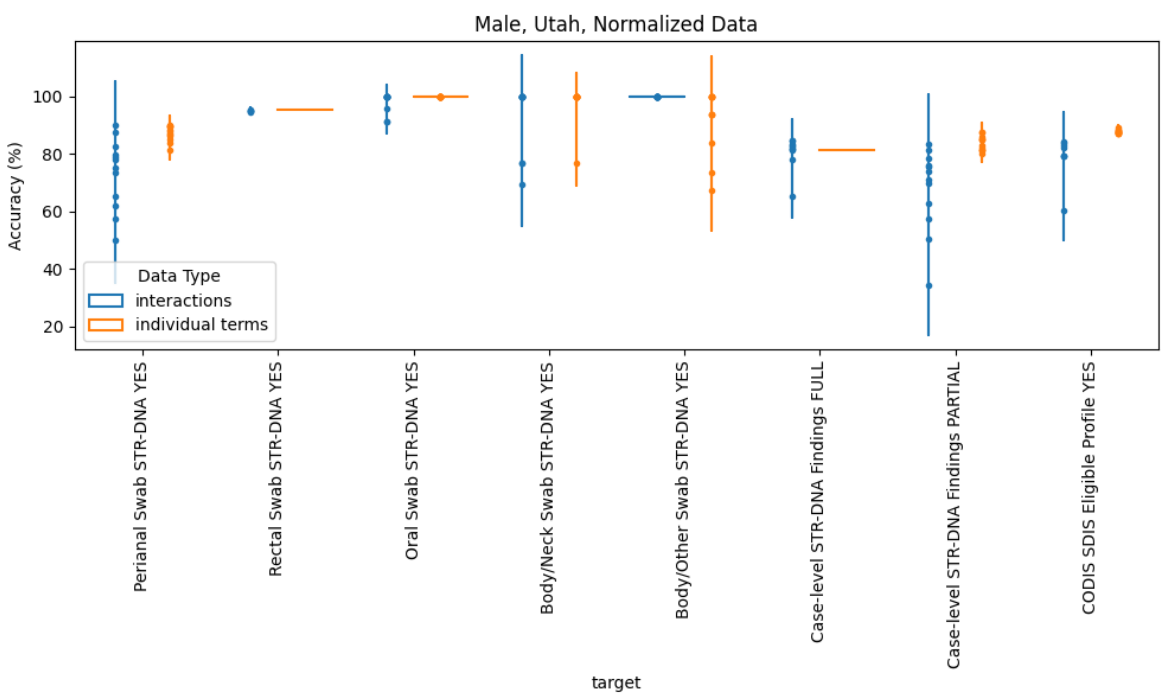


Figure 82. Percent Better than Guessing of Models on Males, Utah, Un-normalized Data

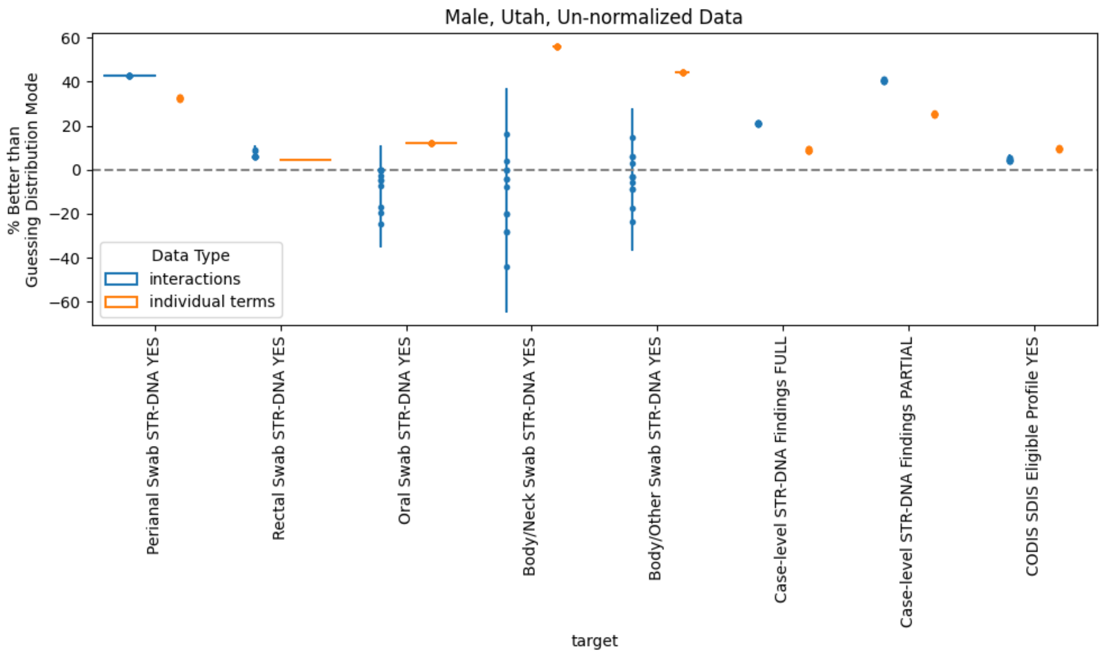


Figure 83. Percent Better than Guessing of Models on Males, Utah, Normalized Data

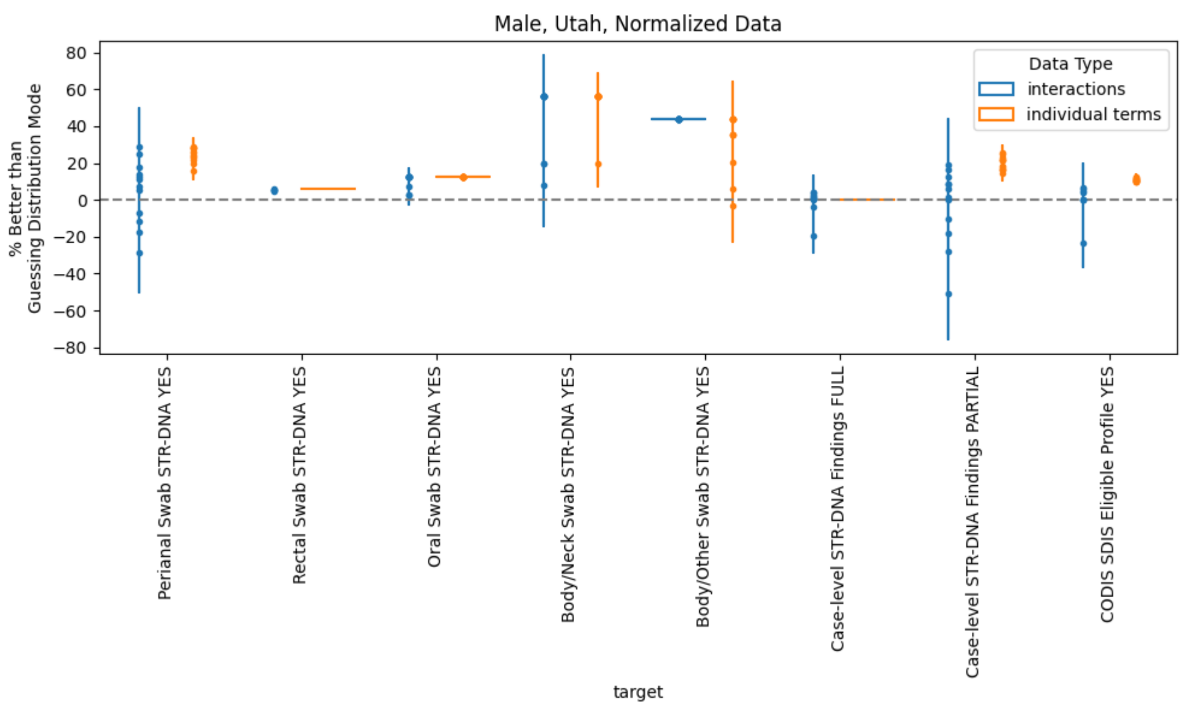


Figure 84. Accuracy of Models on Males, Utah, Interaction-Augmented Data

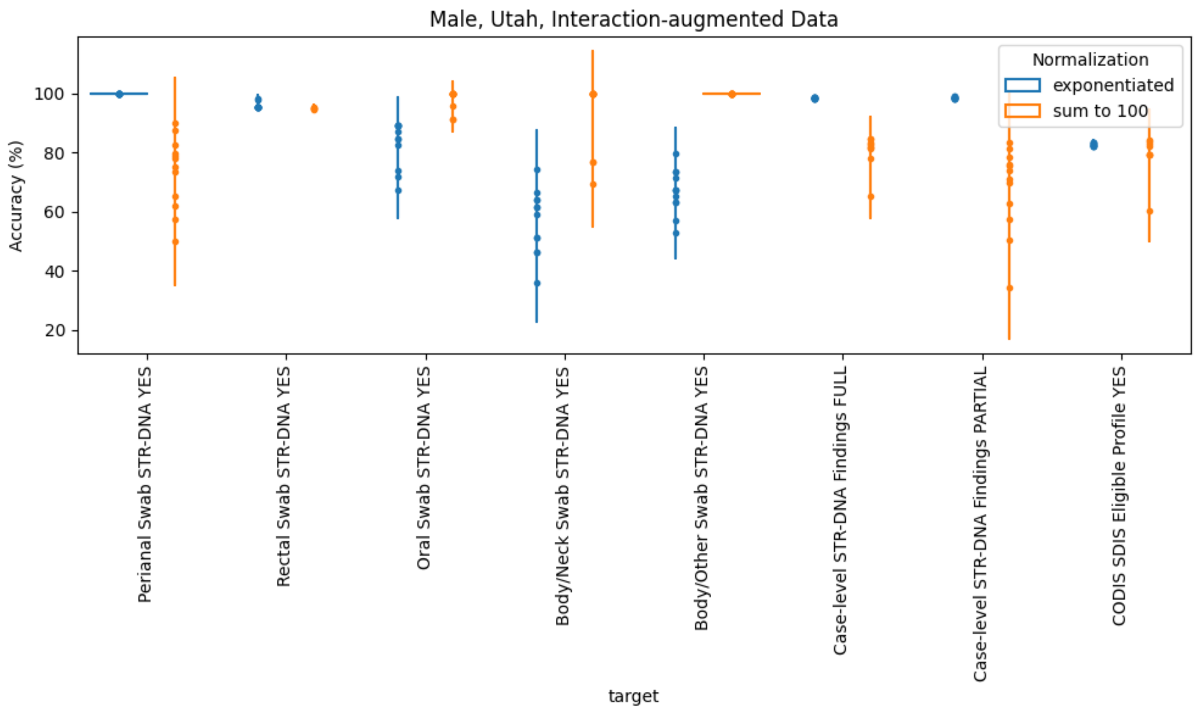


Figure 85. Accuracy of Models on Males, Utah, Non-Interaction Data

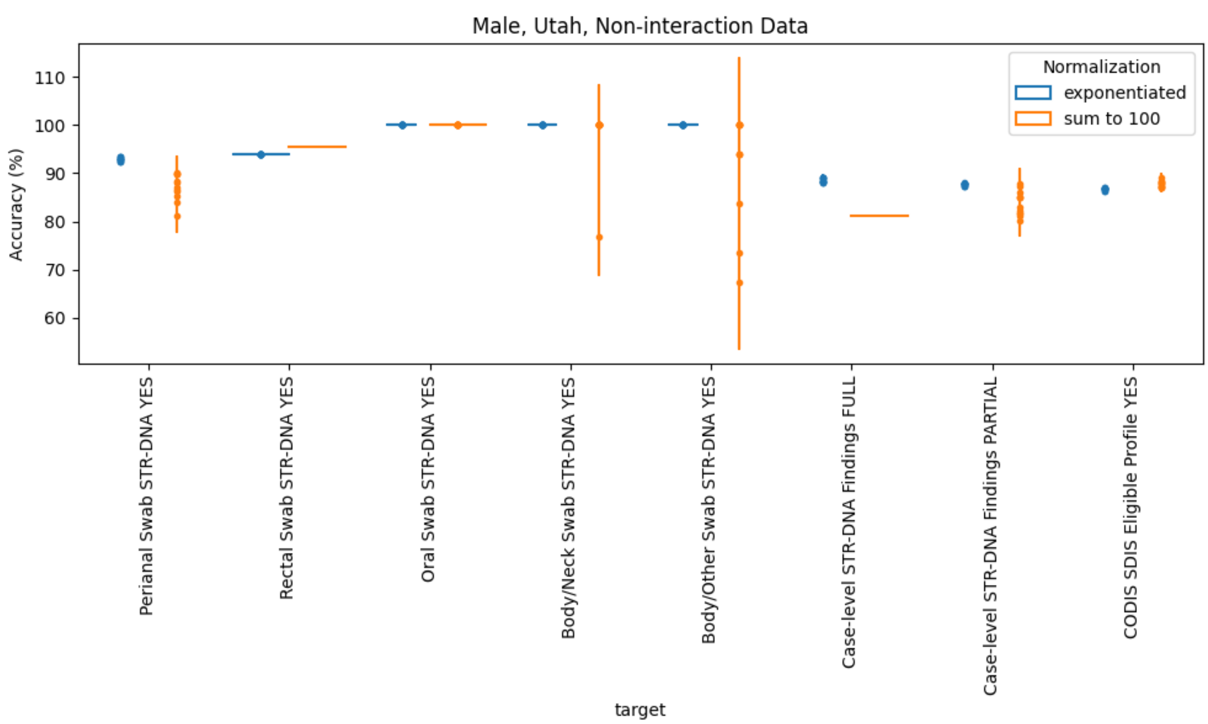


Figure 86. Percent Better than Guessing of Models on Males, Utah, Interaction-Augmented Data

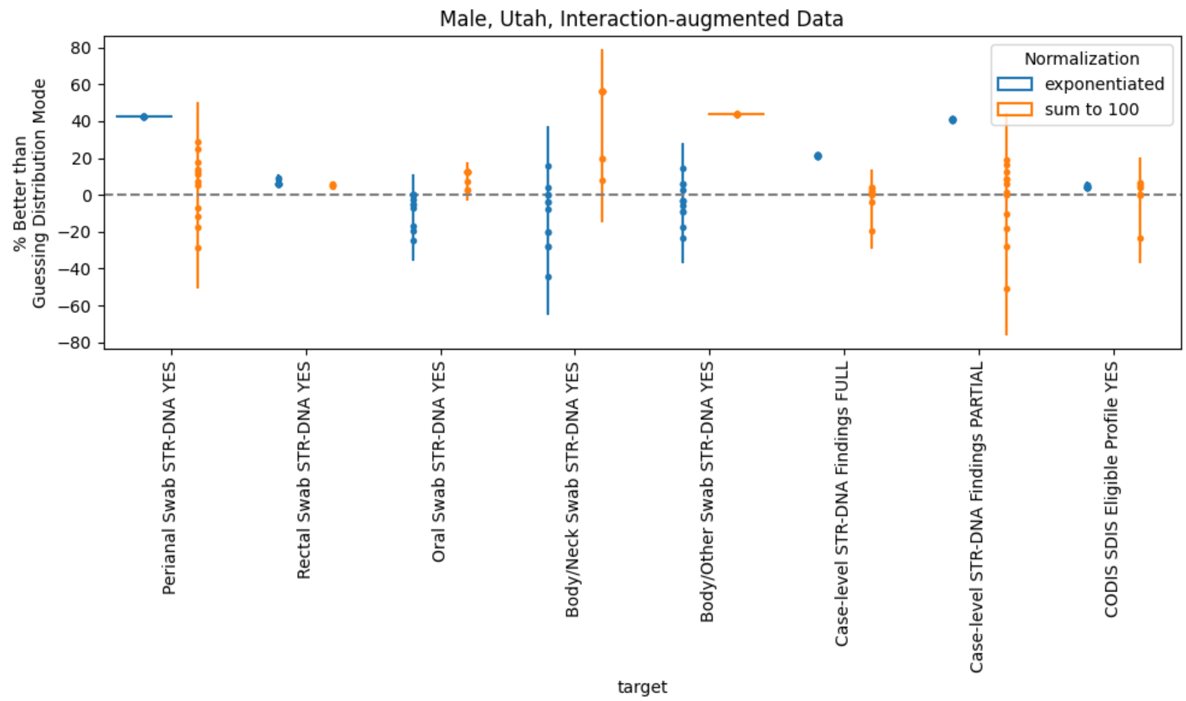


Figure 87. Percent Better than Guessing of Models on Males, Utah, Non-Interaction Data

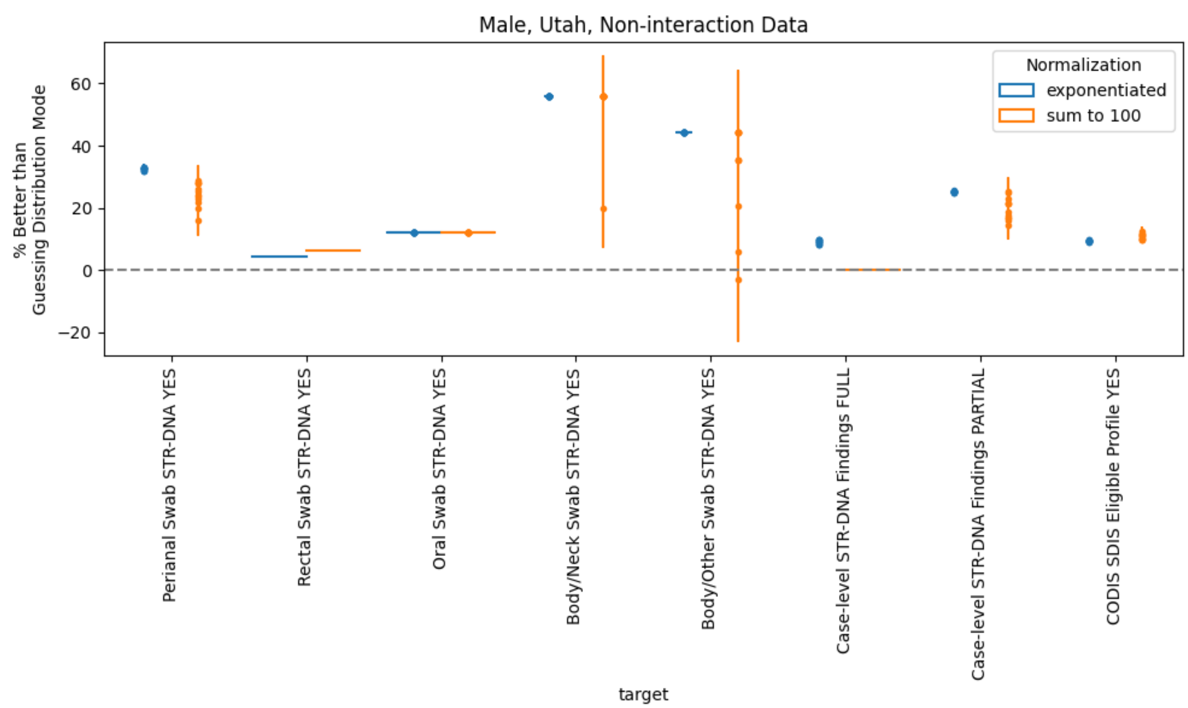


Figure 88. Accuracy Percent of Models on Females, Orange County, Un-normalized Data

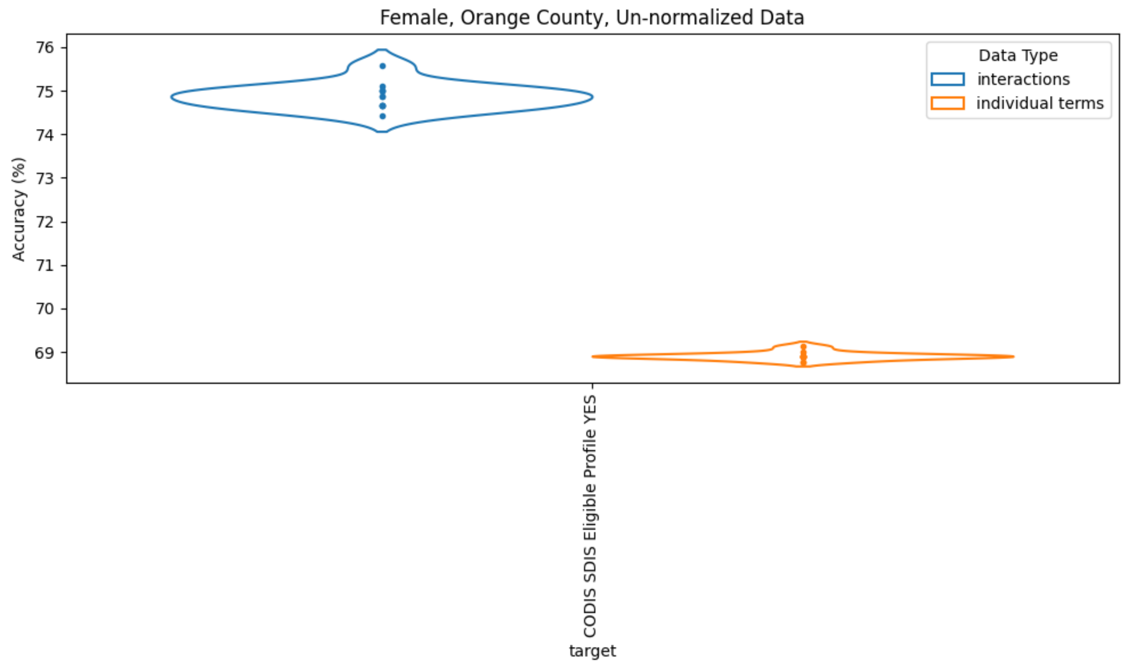


Figure 89. Accuracy Percent of Models on Female, Orange County, Normalized Data

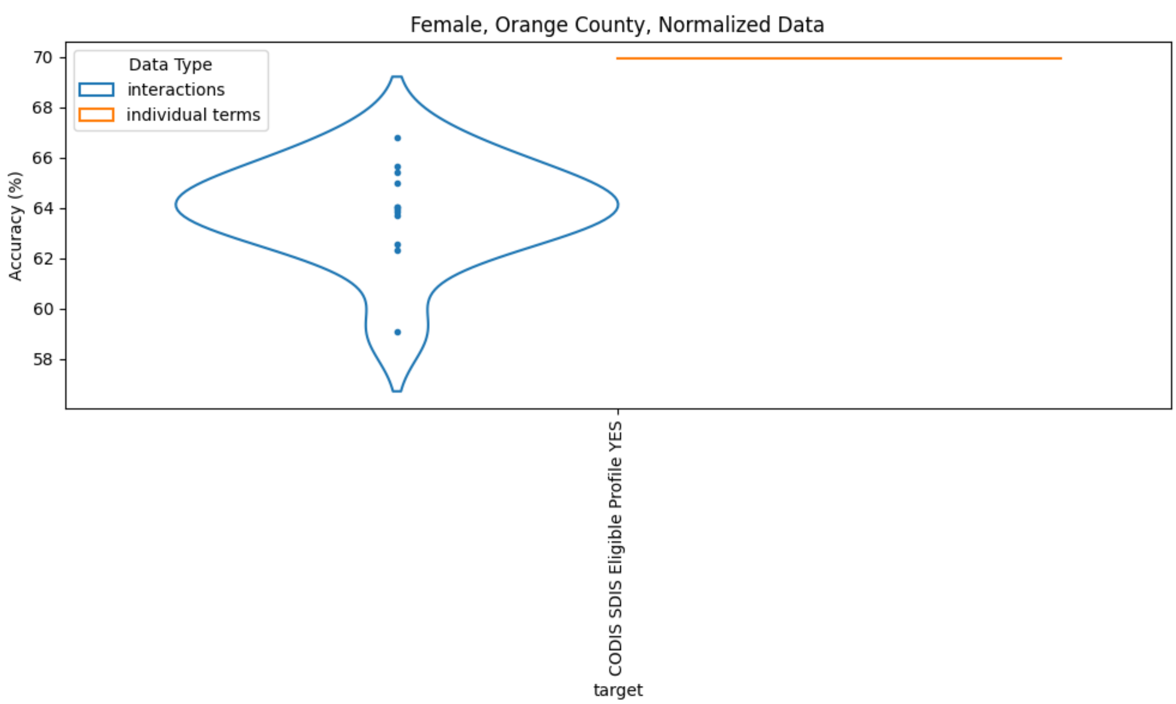


Figure 90. Percent Better than Guessing of Models on Female, Orange County, Un-normalized Data

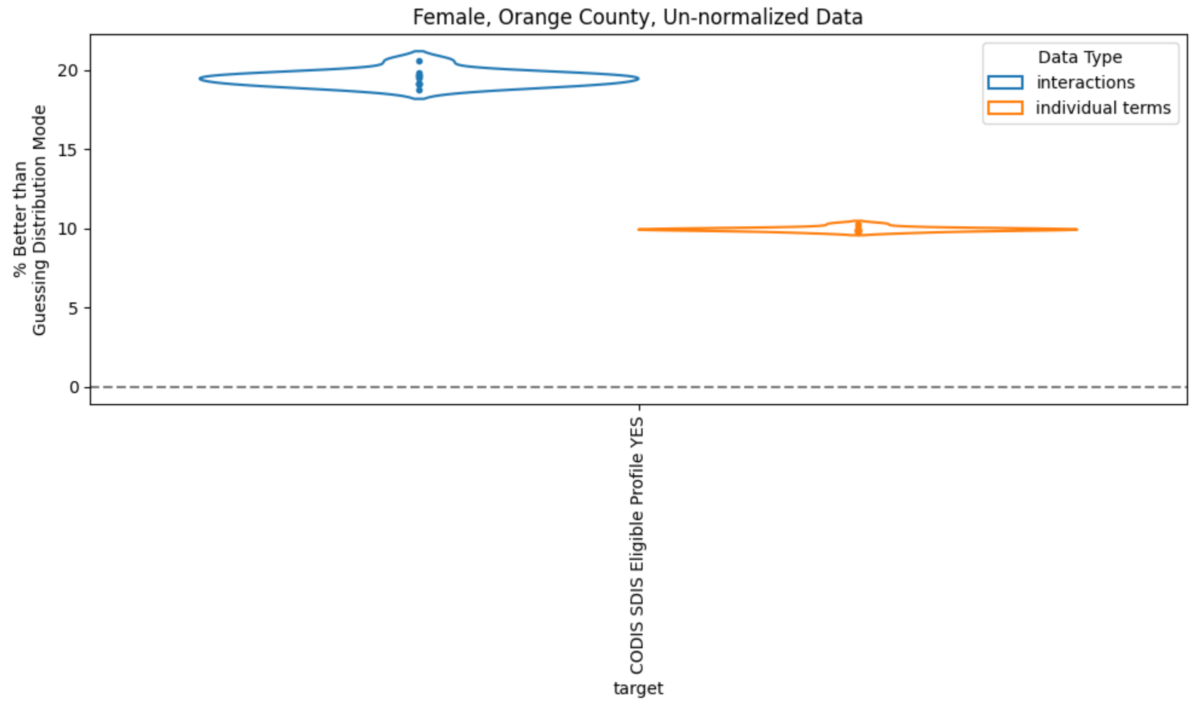


Figure 91. Percent Better than Guessing of Models on Females, Orange County, Normalized Data

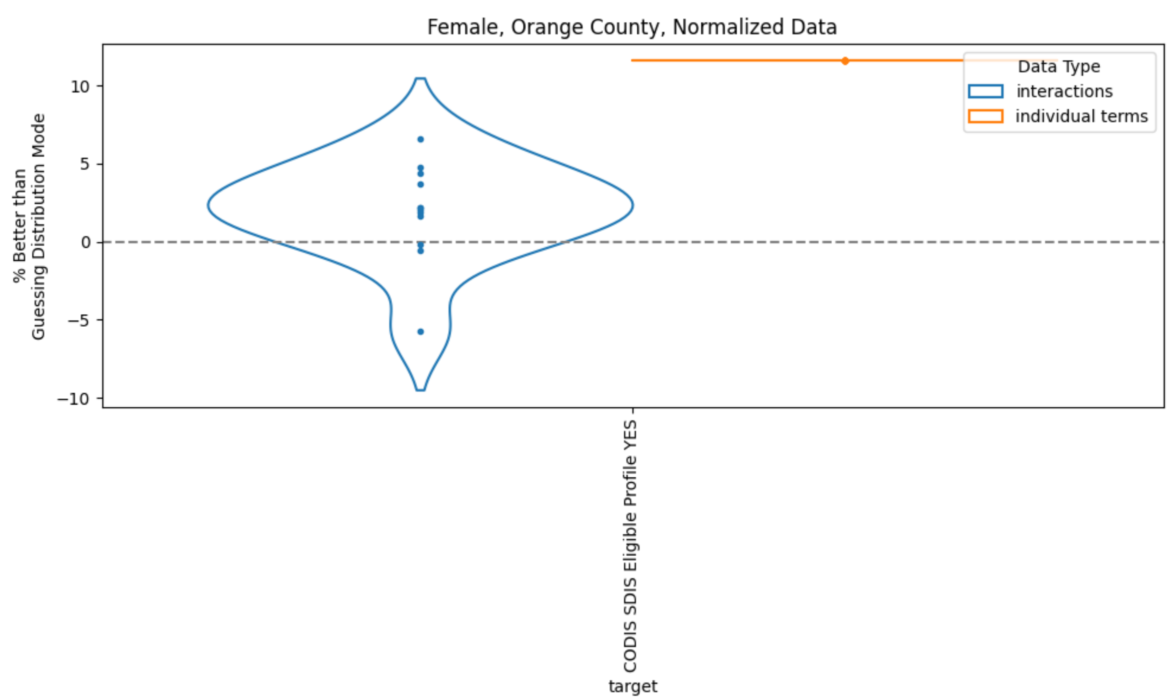


Figure 92. Accuracy of Models on Female, Orange County, Interaction-Augmented Data

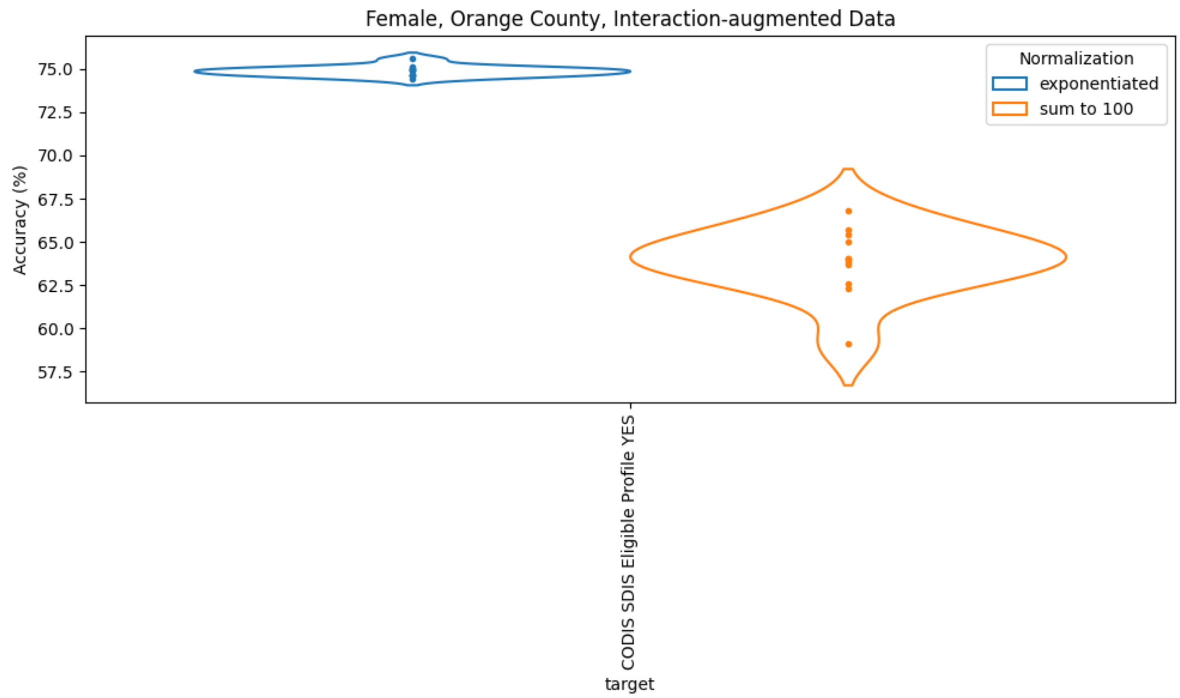


Figure 93. Accuracy of Models on Females, Orange County, Non-Interaction Data

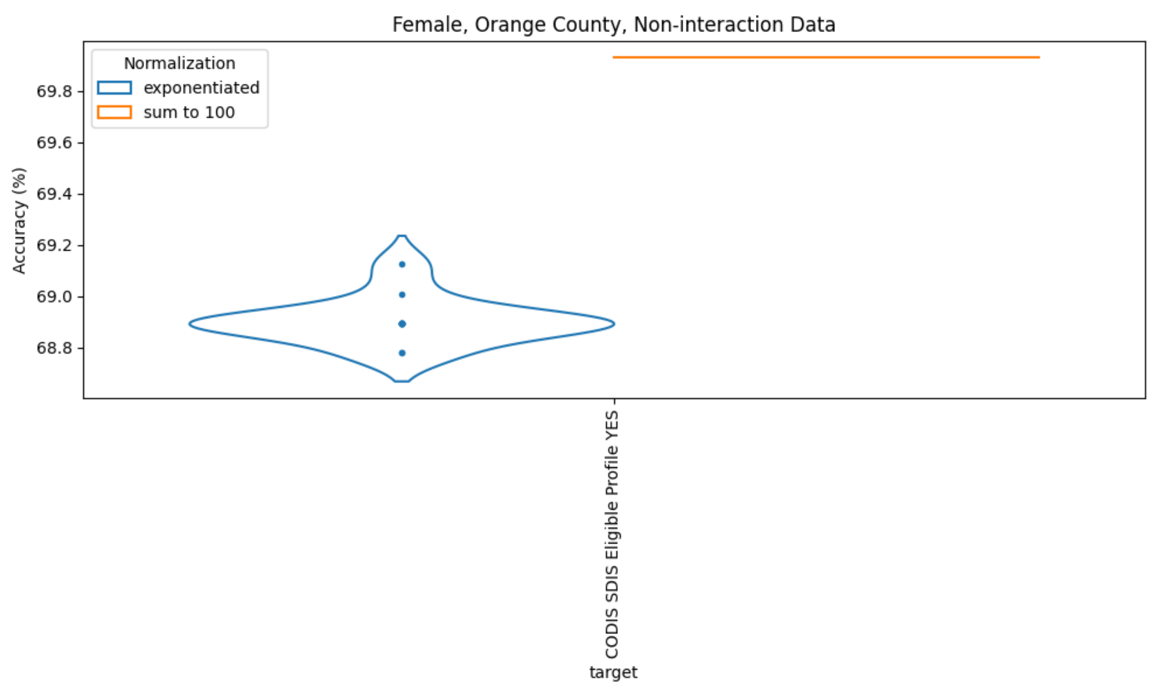


Figure 94. Percent Better than Guessing of Models on Females, Orange County, Interaction-Augmented Data

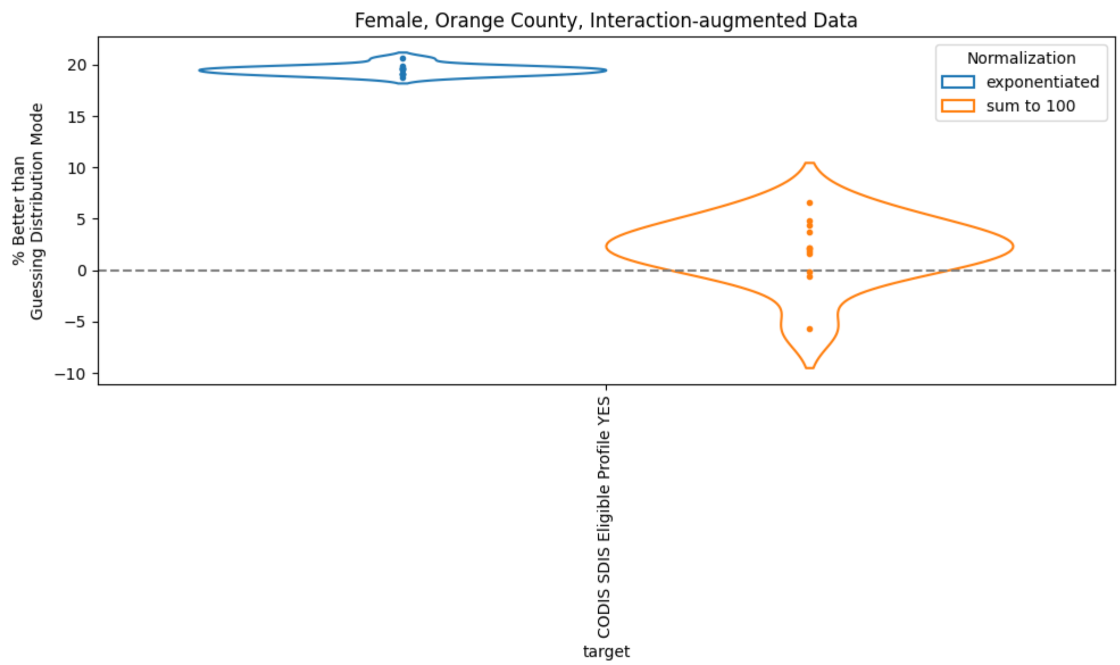


Figure 95. Percent Better than Guessing of Models on Females, Orange County, Non-Interaction Data

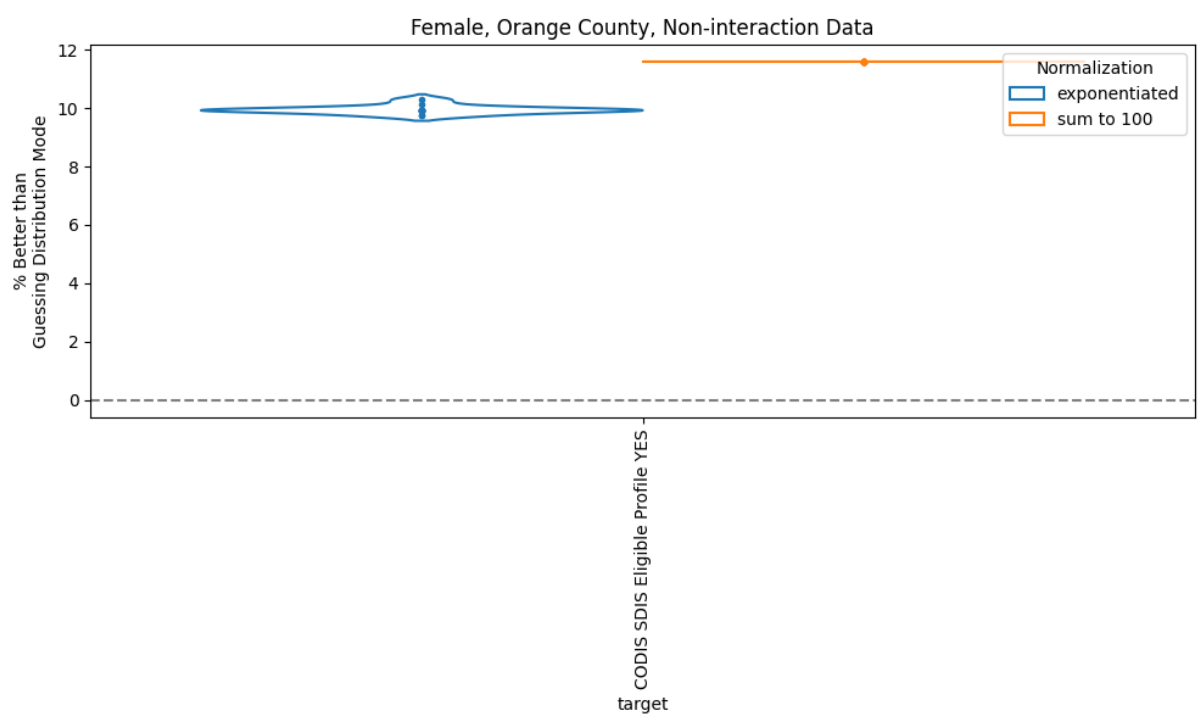


Figure 96. Accuracy Percent of Models on Females, Idaho, Un-normalized Data

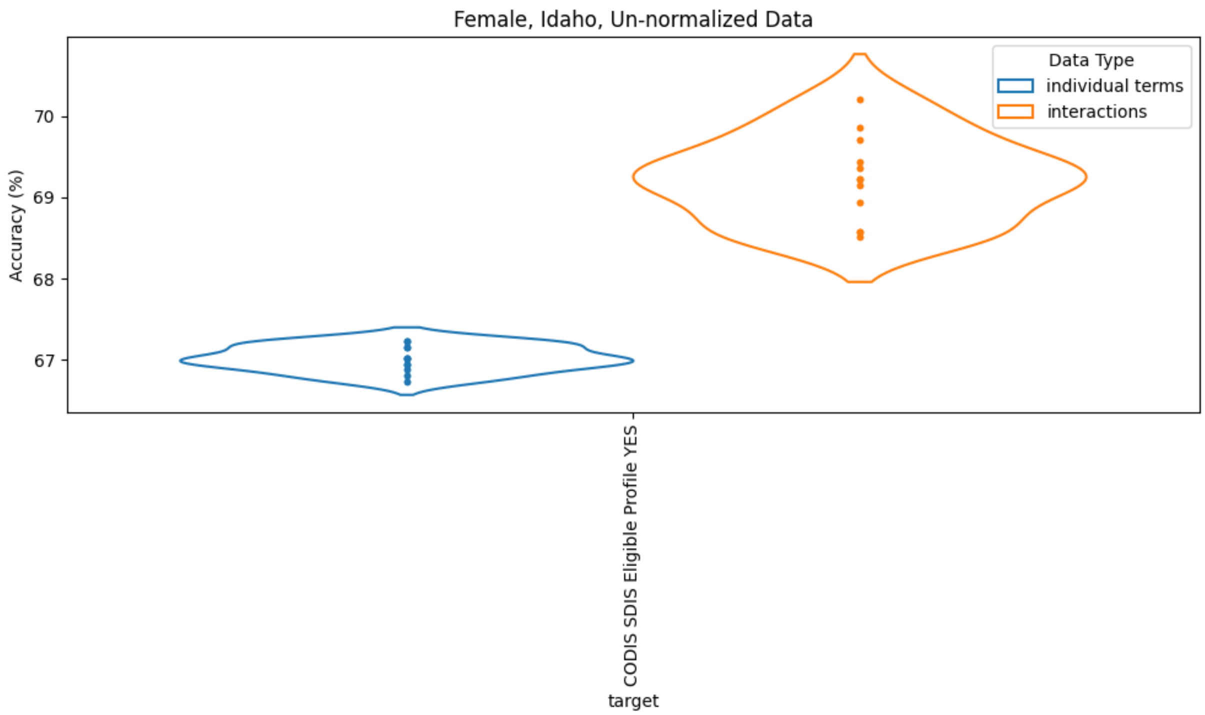


Figure 97. Accuracy Percent of Models on Females, Idaho, Normalized Data

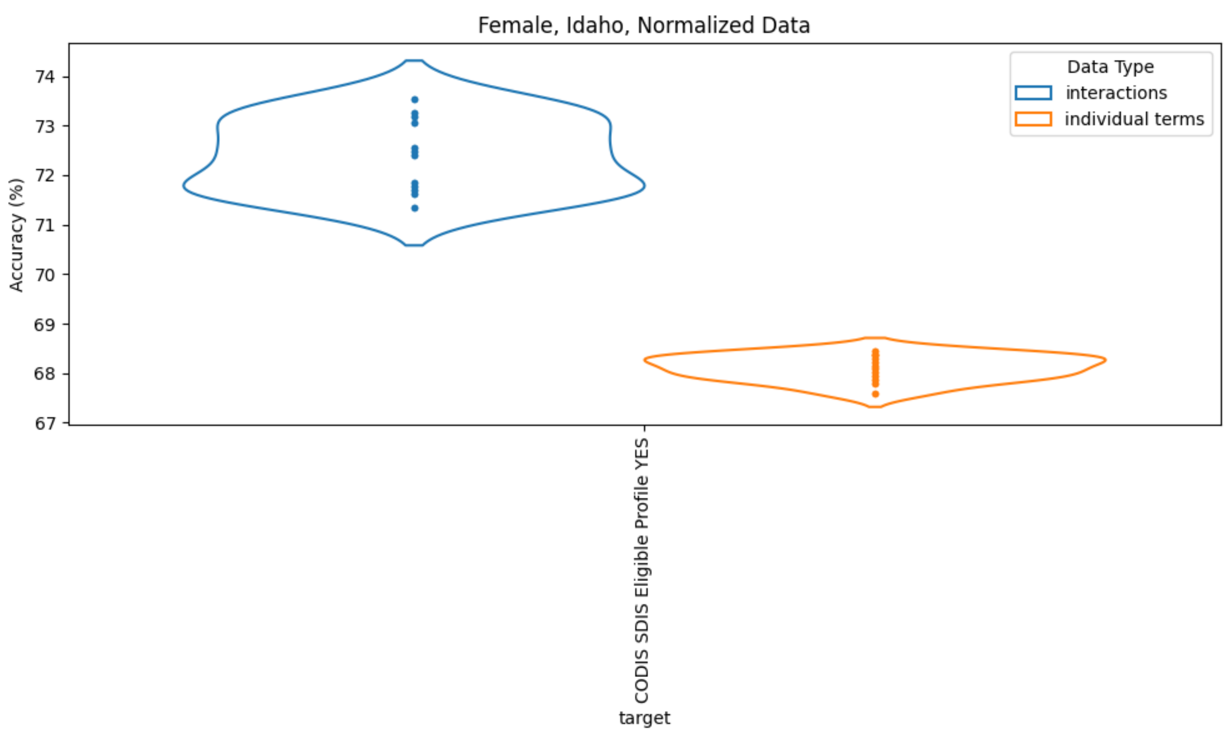


Figure 98. Percent Better than Guessing of Models on Females, Idaho, Un-normalized Data

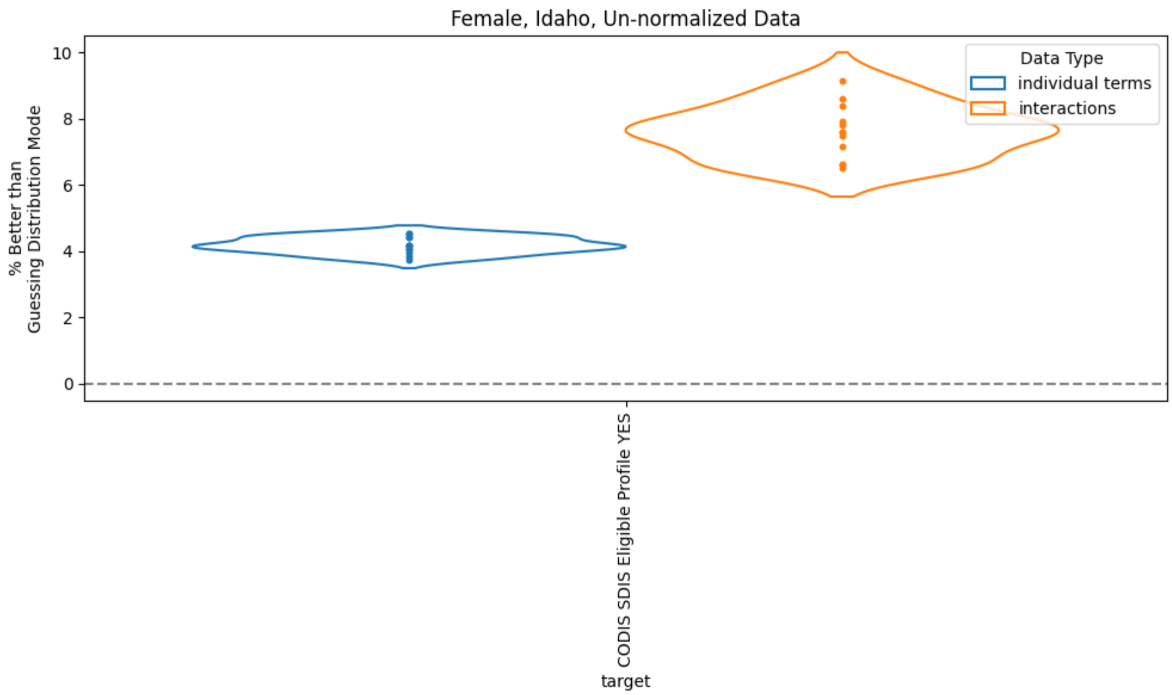


Figure 99. Percent Better than Guessing of Models on Females, Idaho, Normalized Data

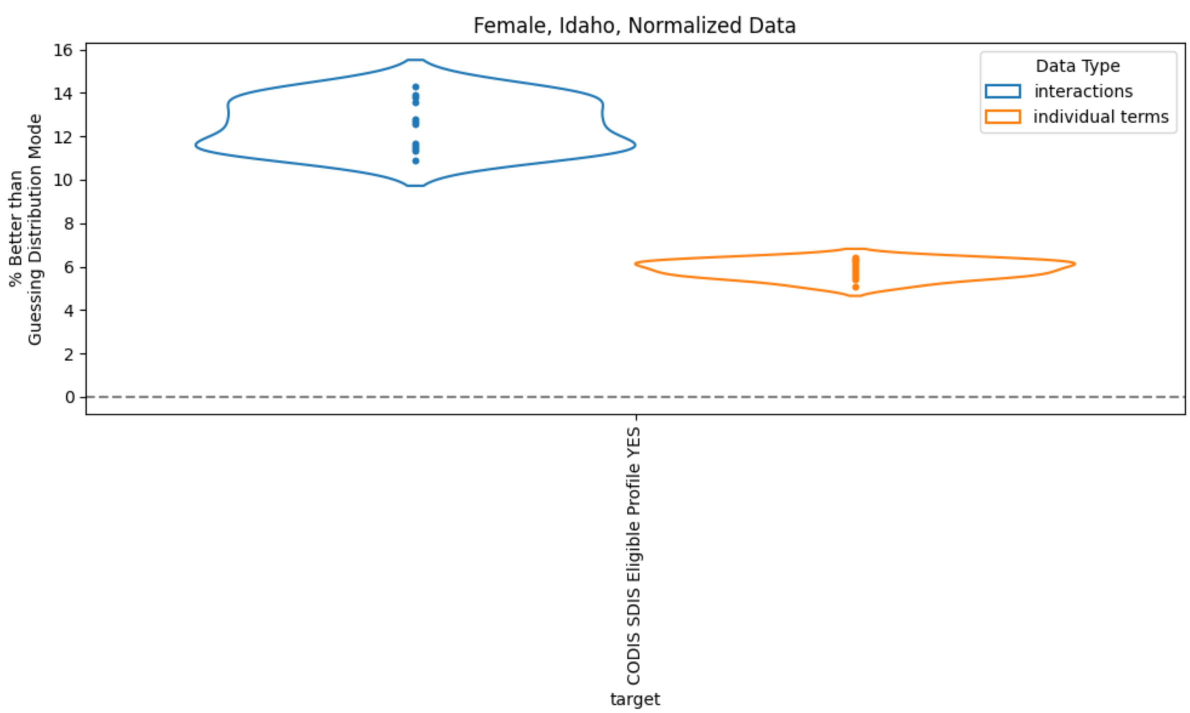


Figure 100. Accuracy of Models on Females, Idaho, Interaction-Augmented Data

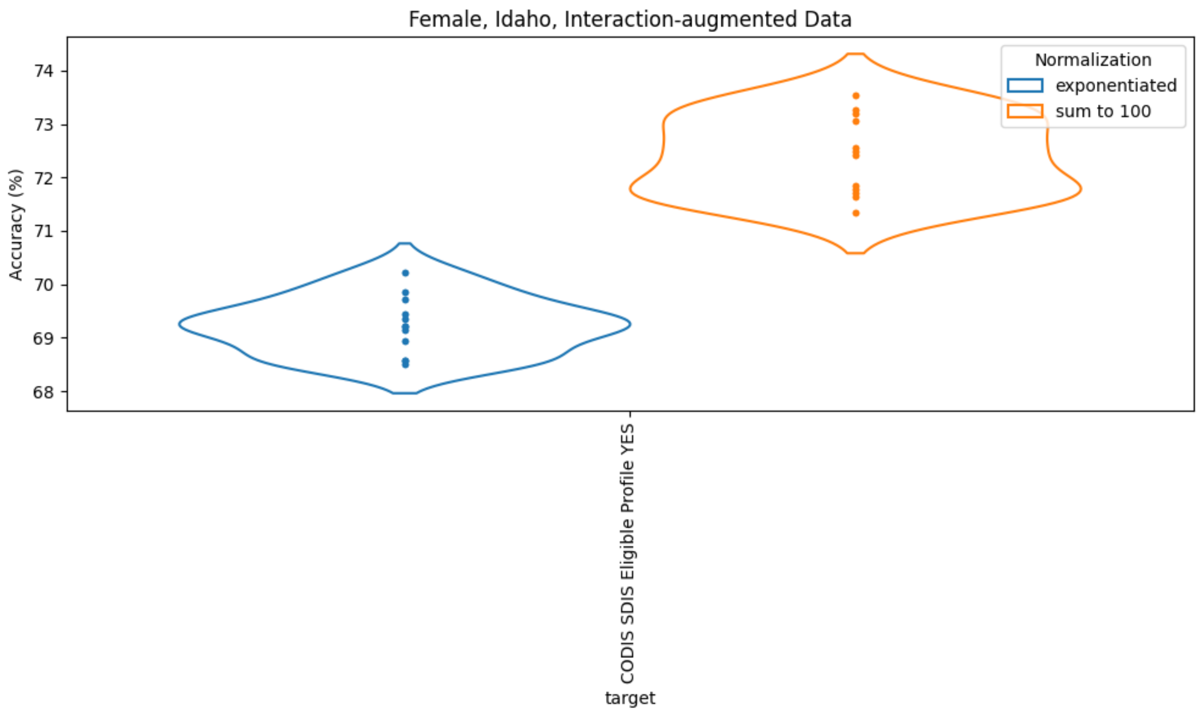


Figure 101. Accuracy of Models on Female, Utah, Non-Interaction Data

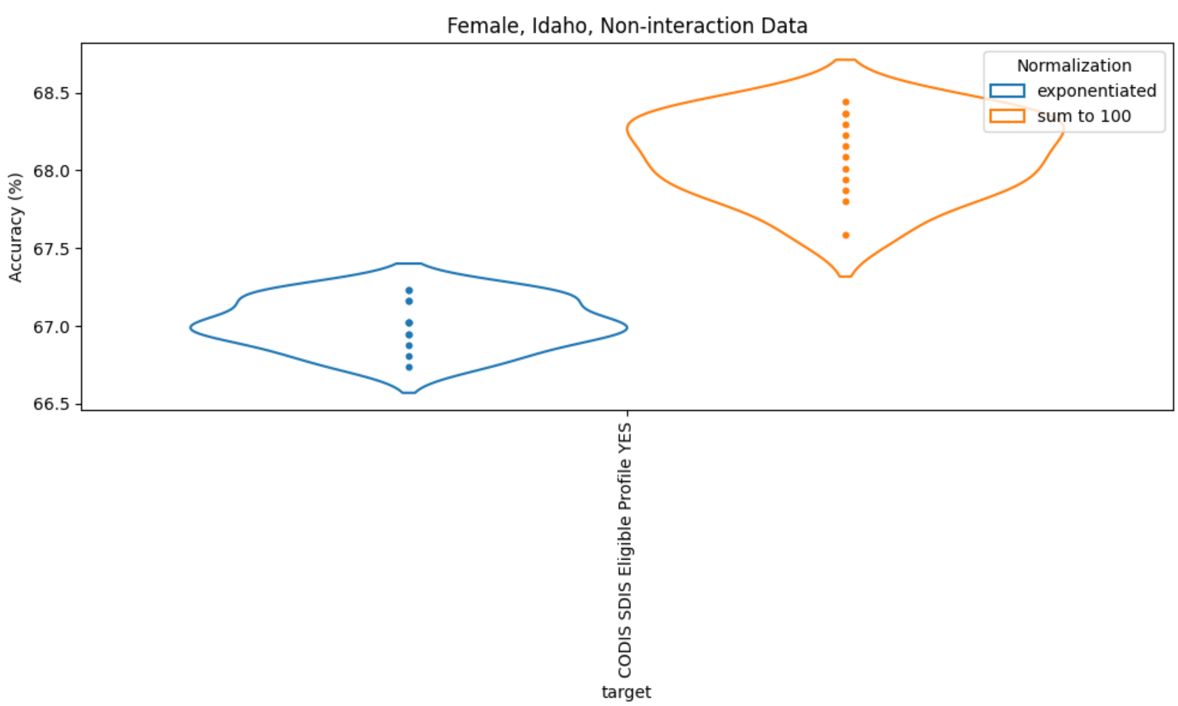


Figure 102. Percent Better than Guessing of Models on Females, Idaho, Interaction-Augmented Data

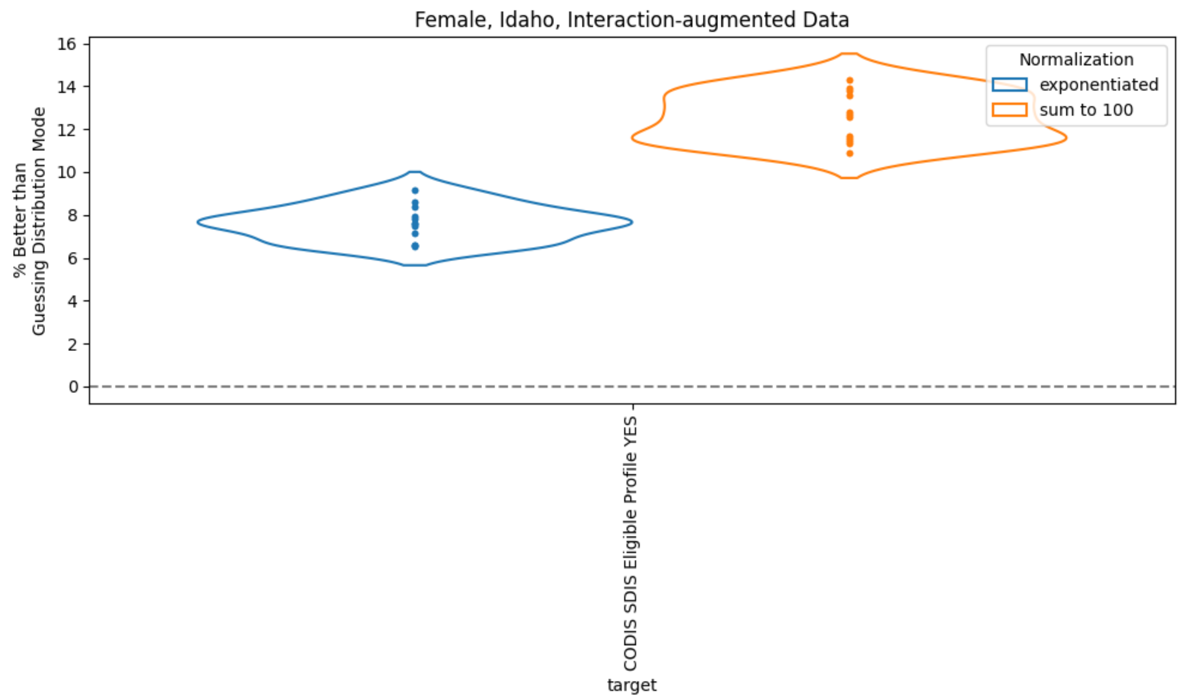
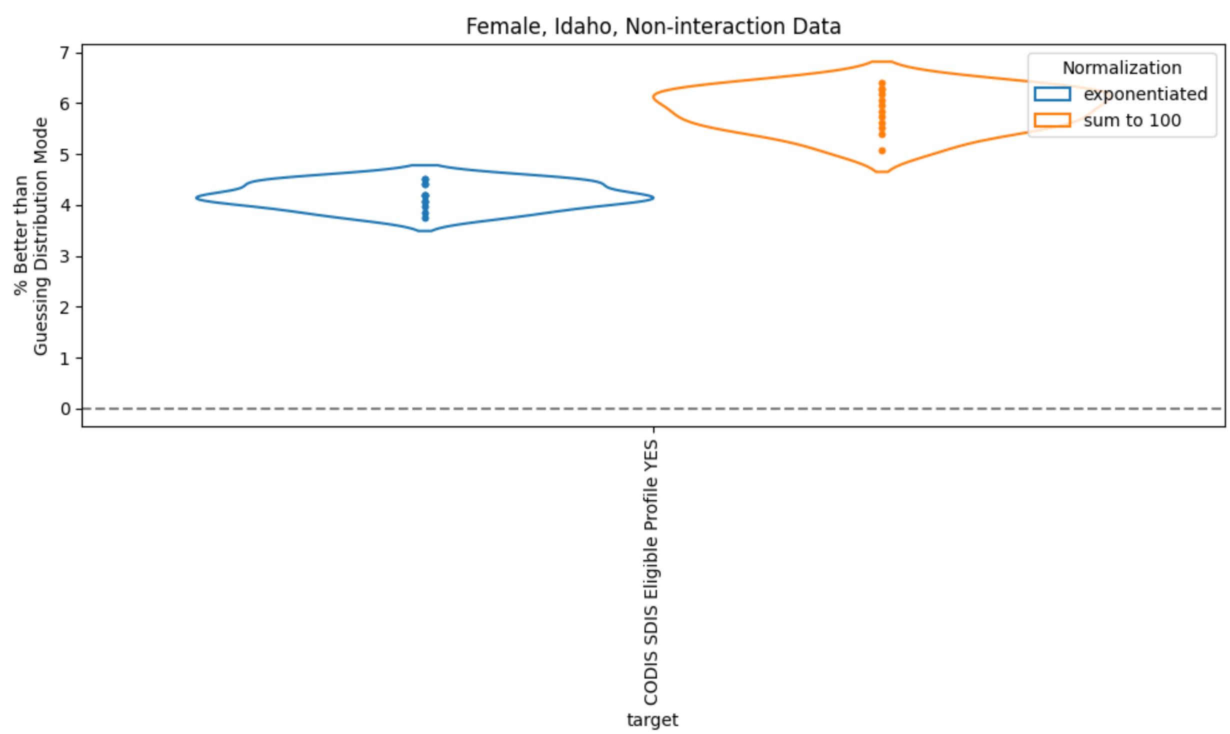


Figure 103. Percent Better than Guessing of Models on Females, Idaho, Non-Interaction Data



Applicability to Criminal Justice

The findings from this study have significant implications for practice and policy recommendations for SAK evidence collection and analysis and, therefore, implications for criminal justice in the investigation and prosecution of sexual assault cases. The dataset created for this study with data from SAMFE forms and crime laboratory databases is currently the largest dataset of its kind in the U.S. with information on 11,715 patients seen for SAMFEs and 9,599 SAKs. The findings from this report and future publications, presentations, and other dissemination methods will hopefully aid in developing evidence-based, multidisciplinary practice recommendations.

The percentage of SAKs that developed uploaded CODIS SDIS profiles in this study was dependent upon the site or crime lab: 33.3% (ISPFS), 34.2% (UBFS), and 46.3% (OCCL). In a review of the few studies exploring the percentage of uploaded CODIS profiles, the range was found to be 25.4% to 57%. The prior studies do not indicate if CODIS was SDIS or NDIS. A multitude of factors may account for this substantial range found in the literature and within this study. Firstly, the development of uploaded CODIS profiles is somewhat dependent upon the expertise and experience of the SANEs or examiners within a jurisdiction and their evidence collection decisions. Further evaluation of evidence collection practices would be useful. Secondly, the practices, policies, expertise, equipment, and interpretation methods within a crime lab would influence the development of uploaded CODIS profiles. Thirdly, the FBI has guidelines for determining CODIS eligibility of developed profiles. Interpretation of these guidelines may vary between crime labs with some labs taking a more conservation approach in CODIS profile upload decisions.

The end-point product of this study was to develop a machine learning model to guide decision-making in the selection of SAK evidence/swabs for analysis. As the project developed, we faced a substantial obstacle in the development of an unbiased machine learning model for SAK evidence. As noted previously, to develop a highly accurate machine learning model, testing *ALL* swabs in thousands of SAKs would be necessary. Unfortunately, this is not a reasonable option due to time, resources, and financial constraints within publicly funded crime laboratories. Yet, our findings do indicate that utilizing logistic regression machine learning models augmented with human interaction could be useful.

If a valid, reliable, and accurate machine learning model was developed, another obstacle exists for widespread utilization in the U.S. – lack of a standardized, national SAK and SAMFE paperwork. As noted in our data collection, each of the three sites collects different information as part of the SAMFE forms resulting in different variables to include in the models. This implies that different jurisdictions and crime labs would require unique machine learning models to guide selection for SAK evidence analysis. If a standardized, national SAK with forensic electronic medical record data was implemented nationally, then the development of a machine learning model to aid in selection of SAK evidence could be very beneficial.

A further challenge in the use of machine learning models for selection of SAK evidence is the time involved to enter the required data, primarily in areas without electronic SAMFE forms. Many U.S. sites continue to use paper SAMFE forms necessitating hand entry of key data points for a machine learning model. For U.S. sites with electronic SAMFE documentation, machine learning models could be implemented if a software bridge was created to extract data from the SAMFE into the machine learning model.

We hope this study highlights the benefits of data collection and analysis from SAMFE forms and SAK testing outcomes. By aggregating de-identified data across disciplines, we aim to develop greater collaboration within communities and improve criminal justice outcomes for survivors.

Products

A list of previous and pending scholarly products and dissemination activities resulting from this funding is provided.

Scholarly Products:

- Valentine, J.L., Miles, L.M., Brown, B., Alder, C., Johnson, L., Criddle, A., Asay, N., & Grimsman, D. (2024) Development of Combined DNA Index System (CODIS) Profiles from sexual assault kits of female victims and associated victim and assault features. (Manuscript in process).
- Valentine, J.L., Miles, L.M., Brown, B., Alder, C., Johnson, L., Criddle, A., Asay, N., & Grimsman, D. (2024) Development of Combined DNA Index System (CODIS) Profiles from sexual assault kits of male victims and associated victim and assault features. (Manuscript in process).
- Valentine, J.L, & Miles, L.M. (2024). Retrospective review of deoxyribonucleic acid analysis findings from sexual assault kits: Implications for forensic nursing practice. (Manuscript in process).
- Allen, C.I., Payne, S., & Valentine, J.L. (2023). Ethical data sharing in forensic research. Forensic Science International: Synergy, 6. https://doi.org/10.1016/j.fsisyn.2023.100322
- Coding of all models referenced in this Technical Summary to be uploaded on Zenodo.

- Archived data road map with link to model codes on the National Archive of Criminal Justice Data website.

Dissemination Activities

International/National Conferences:

- Valentine, J.L. & Miles, L.W. (2023). *Sexual assault of victims born with male genitalia*. American Society of Criminology, 78th Conference, Philadelphia, PA.
- Valentine, J.L., Miles, L.W., & Payne, S. (2023). *Sexual assault kits and development of uploaded CODIS STR DNA profiles*. American Society of Criminology, 78th Conference, Philadelphia, PA.
- Valentine, J.L., Miles, L.W., & Andrelczyk, J. (2023). *Does age matter? Descriptive data and sexual assault kit DNA analysis findings of elderly sexual assault victims.* International Association of Forensic Nurses Conference 2023, Phoenix, AZ.
- Valentine, J.L, Allen, C., Momberger, J., Pugh, S., Payne, S., & Miles, L. (2023). *DNA analysis findings from male sexual assault victims: Multidisciplinary practice implications*. National Institute of Justice Research and Development Symposium, Orlando, FL.
- Valentine, J.L., & Miles, L. (2023). *Does age matter? Descriptive data and sexual assault kit DNA analysis findings of elderly sexual assault victims*. American Academy of Forensic Sciences Annual Conference 2023, Orlando, FL.
- Valentine, J.L., Payne, S., Miles, L., Alder, C., Black, E., & Johnson, L., (2022). *Sexual assault victim and assault characteristics and development of Combined DNA Index System (CODIS)-eligible short tandem repeat (STR) DNA profiles*. American Academy of Forensic Sciences Annual Conference 2022, Seattle, WA.

- Valentine, J.L., & Miles, L. (2021). *DNA analysis findings from >4,000 sexual assault kits: Impact on interdisciplinary practices and policies*. American Society of Criminology 2021Conference: Science and Evidence-Based Policy in a Fractured Era, Chicago, IL.
- Valentine, J.L., Miles, L., & Payne, S. (2022, January). *Retrospective Study on DNA Analysis Findings from Sexual Assault Kits: Implications on Practice and Policy*. National Institute of Justice, Forensic Technology Center of Excellence, virtual.
- Valentine, J.L. (2022, January). *Round table discussion with subject matter experts*, panelist. National Institute of Justice, Forensic Technology Center of Excellence, virtual.
- Black, E., Payne, S., & Valentine, J.L. (2022, February). *The dirty truth: Does bathing after sexual assault prevent the development of Combined DNA Index System (CODIS)-eligible DNA profiles*? American Academy of Forensic Sciences 2022 Conference, Seattle, WA.
- Valentine, J.L., Payne, S., & Miles, L. (2022, September). Development of Combined DNA Index System (CODIS) eligible profiles from sexual assault kits of female victims and associated victims' and assault features, Northwest Association of Forensic Scientists, virtual presentation.
- Valentine, J.L., Payne, S., & Miles, L. (2021, November). *Assessment of Sexual Assault Kit (SAK) Evidence Selection Leading to Development of SAK Evidence Machine-Learning Model (SAK-ML Model) Research Update*, National Institute of Justice, Combined DNA Index System National Conference, virtual conference.

References

American College Health Association. (2022). National College Health Assessment: Undergraduate Student Reference Group, Executive Summary. Retrieved from https://www.acha.org/documents/ncha/NCHA-III_SPRING_2022_UNDERGRAD_REFERENCE_GROUP_EXECUTIVE_SUMMARY.pdf

Campbell, R., Javorka, M., Sharma, D.B., Gregory, K., Opsommer, M., Shelling, K., Lu, L. (2020). A state census of unsubmitted sexual assault kits: Comparing forensic DNA testing outcomes by geographic and population density characteristics. *Journal of Forensic Sciences*, *65*(6), 1820-1827.

Davis, R.C., Jurek, A., Wells, W., & Shadwick, J. (2021). Investigative outcomes of CODIS matches in previously untested sexual assault kits. *Criminal Justice Policy Review*, *32*(8), 841-864.

FBI. (n.d.). Frequently asked questions on CODIS and NDIS. Retrieved from https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet

Kerka, J.E., Heckman, D.J., Albert, J.H., Sprague, J.E., & Maddox, L.O. Statistical modeling of the case information from the Ohio attorney general's sexual assault kit testing initiative. *Journal of Forensic Sciences*, *63*(4), 1122-1133.

Nelson, M.S. (2013). Analysis of untested sexual assault kits in New Orleans. U.S. Department of Justice, Office of Justice Programs. NCJ Number 242313. Retrieved from https://www.ojp.gov/pdffiles1/nij/242312.pdf

Peterson, J., Johnson, D., Herz, D., Graziano, L., & Oehler, T. (2012). Sexual assault kit backlog study. Washington D.C.: The National Institute of Justice.

U.S. Census Bureau. (2022). Quick Facts. Retrieved from https://www.census.gov/quickfacts/fact/table/UT,orangecountycalifornia,ID/PST045222

Wang, Z., MacMillan, K., Powell, M., & Wein, L. (2020). A cost-effective analysis of the number of samples to collect and test from a sexual assault. PNAS, 117(24), 13421-13427. https://doi.org/10.1073/pnas.2001103117

References for Model Development and Coding Decisions

Jović, K. Brkić & N. Bogunović, (2015). A review of feature selection methods with applications, *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), Art. 9.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, *50*(6), 1-45. https://doi.org/10.1145/3136625

Marcinkeviˇcs, R., & Vogt, J. E. (2023) Interpretable and explainable machine learning: A methods-centric overview with concrete examples. WIREs Data Mining and Knowledge Discovery, e1493. doi: 10.1002/widm.1493

Appendix A

Descriptive Data				
		2010-2022	2015-2020	2013-2020
#	Variable	Utah *N* = 8981 patients/SAKs *N* = 6865 submitted SAKs	Orange County *N* = 1207 SAKs	Idaho *N* = 1527 SAKs
2	**Site** Site A Site B Site C Site D Site E	 *n*=5343 *n*=494 *n*=214 *n*=1378 *n*=1534		
3	**Exam by SANE** 0= No 1= Yes	 7% 93%	 0% 100%	
4	**Year SAK collected** 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022	 540 (6.0%) 548 (6.1%) 566 (6.3%) 520 (5.8%) 521 (5.8%) 630 (7.0%) 709 (7.9%) 860 (9.6%) 849 (9.5%) 900 (10%) 786 (8.8%) 848 (9.4%) 703 (7.8%)		
5	**Kit Brought to Crime Lab** No Yes	 1869 (20.8%) 7119 (79.3%)	 0% 100%	
6	**SAK Submission Time from collection** Not submitted Submitted within 1 month Submitted 1 month – 1 year Submitted after 1 year	 1869 (20.8%) 5394 (60.1%) 970 (10.8%) 746 (8.3%)		
7	**Age** Mean Median Mode Std. Deviation	 27.71 24.0 18 11.4	 24.44 22.0 21 12.003	 25.29 21 16 11.969

	Range	14-95		14-94
	Missing	13	0	92
	Percentiles			
	25	19.0	17.0	17.0
	50	24.0	22.0	21.0
	75	34.0	30.0	31.0
8	**Gender**			
	Female	8468 (94.3%)	1159 (96%)	1524 (97%)
	Male	430 (4.8%)	48 (4%)	48 (3%)
	Transgender/Intersex	83 (0.9%)		
9	**Race**			n=431/1572 Valid %
	White	6719 (74.8%)	554 (45.9%)	333 (77.3%)
	Hispanic	1098(12.2%)	51 (4.2%)	66 (15.3%)
	Black	316 (3.5%)	484 (40.1%)	4 (0.9%)
	Native American	265 (3.0%)	79 (6.5%)	13 (3.0%)
	Other	270 (3.0%)	79 (0.1%)	XXX
	Asian/Pacific Islander	203 (2.3%)	29 (2.4%)	4 (0.9%)
	Unknown	42 (0.5%)	9 (0.7%)	11 (2.6%)
	Missing	68	0	1141
	Patient with Physical or Mental Impairment			
	No	8004 (89.1%)		
	Yes	884 (9.8%)		
	Unknown	49 (0.5%)		
	Missing	44		
10	**Time (Hours) from between assault and exam**			
	Mean	28.8	25.898	
	Median	16.0	16.0	
	Mode	4.0	6.0	
	Std. Deviation	38.13	28.72	
	Skewness	6.1	2.3	
	Std. Error- skewness	0.03	0.066	
	Range	1025	245	
	Min	1	1	
	Max	1025.0	245	
	Percentiles			
	25	6.5	7.0	
	50	16.0	16.0	
	75	37.0	33.0	
			0	

	Missing	160		
11	**Consensual Sexual Contact Within 120 Hours of assault**			
	No	6185 (68.9%)	841 (69.7%)	1006 (64%)
	Yes	2566 (28.6%)	342 (28.3%)	294 (18.7%)
	Unknown	86 (1.0%)	24 (2.0%)	93 (5.9%)
	Missing	159	0	179
12	**Suspect Relationship**			*Valid % (n=509)*
	Stranger	1626 (18.1%)	121 (10%)	62 (12.2%)
	Acquaintance	5182 (57.7%)	738 (61.1%)	331 (65%)
	Spouse/Partner	620 (6.9%)	64 (5.3%)	23 (4.5%)
	Other	572 (6.4%)	90 (7.5%)	45 (8.8%)
	Ex-partner	519 (5.8%)	124 (10.3%)	28 (5.5%)
	Unknown by patient	438 (4.9%)	70 (5.8%)	20 (3.9%)
	Missing	24	0	1063
13	**Location of Assault**			*Valid % (n=461)*
	House/Apt.	5602 (62.4%)	638 (52.9%)	330 (71.6%)
	Other	1221 (13.6%)	160 (13.3%)	24 (5.2%)
	Car	844 (9.4%)	150 (12.4%)	47 (3%)
	Outside	810 (9.0%)	62 (5.1%)	31 (6.7%)
	Unknown by patient	351 (3.9%)	63 (5.2%)	9 (2%)
	Hotel/Motel/Inn	125 (1.4%)	134 (11.1%)	20 (4.3%)
	Missing	28	0	1111
14	**Multiple Suspects**			*Valid % (n=837)*
	No	7685 (85.6%)	1043 (86.4%)	721 (86.1%)
	Yes	862 (9.6%)	98 (8.1%)	75 (9%)
	Unknown by patient	411 (4.6%)	66 (5.5%)	41 (4.9%)
	Missing	23	0	735
15	**Multiple Suspects Number**			
	Mean	2.49	2.58	
	Median	2	2.0	
	Mode	2	2	
	Std. Deviation	1.134	1.437	
	Min	2	2	
	Max	17	15	
	Percentiles			
	25	2	2.0	
	50	2	2.0	
	75	3	3.0	

16	**Patient Action scratch suspect (n=4919)** No Yes Unknown	3031 (61.6%) 464 (9.4%) 1424 (28.9%)		
17	**Patient Action bit suspect (n=4920)** No Yes Unknown	3509 (71.3%) 231 (4.7%) 1180 (24.0%)		
18	**Patient Action hit suspect (n=4917)** No Yes Unknown	3160 (64.3%) 570 (11.6%) 1187 (24.1%)		
19	**Patient Action kick suspect (n=4919)** No Yes Unknown	3256 (66.2%) 456 (9.3%) 1207 (24.5%)		
20	**Patient Action other action against suspect, primarily shoved/pushed (n=4842)** No Yes Unknown	2843 (58.7%) 846 (17.5%) 1153 (23.8%)		
21	**Suspect Action verbal threat or coercion (n=6215)** No Yes Unknown by patient	2585 (41.6%) 2368 (38.1%) 1262 (20.3%)		
22	**Suspect Action grabbed or held patient** No Yes Unknown by patient Missing	1565 (17.4%) 5415 (60.3%) 1955 (21.8%) 46		*Valid % (n=133)* 4 (26.1%) 112 (84.2%) 17 (12.8%) 1503
23	**Suspect Action hit patient** No Yes Unknown by patient	5533 (61.6%) 1451 (16.2%) 1949 (21.7%)		*Valid % (n=69)* 18 (26.1%) 33 (47.8%) 18 (26.1%)

	Missing	48		1503
24	**Suspect Action strangled patient** No Yes Unknown by patient Missing	 5524 (61.5%) 1491 (16.6%) 1918 (21.4%) 48		*Valid % (n=453)* 339 (74.8%) 73 (16.1%) 41 (9.1%) 1119
25	**Suspect Action used weapon** No Yes Unknown by patient Missing	 6032 (67.2%) 916 (10.2%) 1986 (22.1%) 47		*Valid % (n=49)* 18 (36.7%) 14 (28.6%) 17 (34.7%) 1523
26	**Suspect Action used restraints** No Yes Unknown by patient Missing	 6605 (73.5%) 456 (5.1%) 1874 (20.9%) 46		Valid % (n=53) 22 (41.5%) 13 (24.5%) 18 (34%) 1519
27	**Suspect Action burned patient** No Yes Unknown by patient Missing	 7189 (80%) 126 (1.4%) 1615 (18%) 51		
28	**Suspected Drug Facilitated** No Yes Unknown by patient Missing	 6979 (77.7%) 1481 (16.5%) 463 (5.2%) 58		*Valid % (n=297)* 240 (80.8%) 52 (17.5%) 5 (1.7%) 1275
29	**Patient Drug Use before assault** 0= No 1= Yes 2= Unknown by patient Missing	 7289 (81.2%) 1481 (16.5%) 463 (5.2%) 70	 704 (59.2%) 358 (30.1%) 128 (10.8%) 17	Valid % (n=237) 201 (84.8%) 34 (14.3%) 2 (0.8%) 1335
30	**Patient Alcohol Use before assault** 0= No 1= Yes	 5189 (57.9%) 3606 (40.2%)	 482 (39.9%) 692 (57.3%)	Valid % (n=275) 140 (50.9%) 133 (48.4%)

	2= Unknown by patient Missing	101 (1.1%) 76	18 (1.5%) 15	2 (0.7%) 1297
31	**Suspect Drug Use in assault** 0= No 1= Yes 2= Unknown Missing	 3905 (43.5%) 1620 (18%) 3390 (37.7%) 66		
32	**Suspect Alcohol Use in assault** 0= No 1= Yes 2= Unknown Missing	 2656 (29.6%) 2981 (33.2%) 3280 (36.5%) 64		
33	**Patient or Suspect Drug or Alcohol Use** 0= No 1= Yes 2= Unknown Missing	 1808 (20.1%) 5134 (57.2%) 1975 (22%) 64		
34	**Loss of Consciousness or Awareness** 0= No 1= Yes 2= Unknown Missing	 4526 (50.4%) 4295 (47.8%) 99 (1.1%) 61	 633 (52.4%) 556 (46.1%) 0 (0%) 18	*Valid % (n=312)* 163 (52.2%) 148 (47.4%) 1 (0.3%) 1260
35	**Patient reported one or more unknown answer to questions regarding penetrative acts** No Yes Unknown Missing	 4528 (50.4%) 4390 (48.9%) 19 (0.2%) 44		
36	**Patient reported four or more unknown answer to questions regarding penetrative acts** No Yes Unknown Missing	 5941 (66.2%) 2977 (33.1%) 18 (0.2%) 45		
37	**Patient reported unknown for all answers to questions regarding penetrative acts**			

	No Yes Unknown Missing	7331 (81.6%) 1583 (17.6%) 23 (0.3%) 44		
38	**Number of unknown responses regarding patients' answers to questions regarding penetrative acts** Mean Median Mode Std. Deviation Minimum Maximum Percentiles 25% 50% 75% Missing	 4.43 0 0 6.098 0 18 0 0 11 81		
39	**Patient reported as asleep and awakened to being sexually assaulted** No Yes Unknown by patient Missing	 7741 (86.2%) 1096 (12.2%) 115 (1.3%) 29		Valid % (n=312) 206 (80.8%) 48 (18.8%) 1 (0.4%) 1317
40	**Assaultive Act** **Contact with Pt's Vagina by Assailant Penis/Genitals** No Yes Unknown by patient NA; Male patient Attempted Missing	 974 (10.8%) 5367 (59.8%) 2168 (24.1%) 430 (4.8%) XXXX 39	 120 (9.9%) 646 (53.5%) 376 (31.2%) 25 (2.1%) 40	 136 (8.7%) 915 (58.2%) 342 (21.8%) 47 (3%) XXXX 132
41	**Assaultive Act** **Contact with Pt's Vagina by Assailant Finger/Hand** No Yes Unknown NA; Male patient Attempted Missing	 1360 (15.1%) 4532 (50.5%) 2619 (29.2%) 430 (4.8%) XXXX 79	 269 (22.3%) 384 (31.8%) 493 (40.8%) 14 (1.2%) 47	Valid % (n=551) 89 (16.2%) 314 (57%) 126 (22.9%) 22 (4 %) XXX 1021

42	**Assaultive Act** **Contact with Pt's Vagina by Assailant Mouth/Tongue** No Yes Unknown by patient NA; Male patient Missing	 4089 (45.5%) 1780 (19.8%) 2638 (29.4%) 430 (4.8%) 42	 466 (38.6%) 189 (15.7%) 513 (42.5%) 8 (0.7%) 31	 761 (48.4%) 166 (10.6%) 389 (24.7%) 41 (2.6%) 215
43	**Assaultive Act** **Contact with Pt's Vagina by object** No Yes Unknown by patient NA; Male patient Missing	 4751 (52.9%) 327 (3.6%) 2624 (29.2%) 430 (4.8%) 52	 614 (50.9%) 18 (1.5%) 518 (42.9%) 1 (0.1%) 56	Valid % (n=373) 252 (67.6%) 10 (2.7%) 94 (25.2%) 17 (2.6%) 1199
44	**Assaultive Act** **Contact with Pt's Anus by Assailant Penis/Genitals** No Yes Unknown Attempted Missing	 4751 (52.9%) 1603 (17.8%) 2780 (31%) XXXX 40	 517 (42.8%) 121 (10%) 500 (41.4%) 39 (3.2%) 62	 772 (49.1%) 230 (14.6%) 404 (25.6%) XXXX 166
45	**Assaultive Act** **Contact with Pt's Anus by Assailant Finger/Hand** No Yes Unknown Attempted Missing	 4854 (54%) 1304 (14.5%) 2780 (31%) XXXX 43	 557 (46.1%) 86 (7.1%) 520 (43.1%) 14 (1.2%) 30	Valid % (n=358) 214 (59.8%) 33 (9.2%) 111 (31%) XXXX 1214
46	**Assaultive Act** **Contact with Pt's Anus by Assailant Mouth/Tongue** No Yes Unknown Missing	 5785 (64.4%) 391 (4.4%) 2756 (30.7%) 49	 603 (50%) 43 (3.6%) 530 (43.9%) 31	*Valid % (n=365)* 250 (68.5%) 13 (3.6%) 102 (27.9%) 1207
47	**Assaultive Act** **Contact with Pt's Anus by Object** No Yes	 6055 (67.4%) 217 (2.4%)	 634 (52.5%) 12 (1%)	

	Unknown Attempted Missing	2653 (29.5%) 56	517 (2.8%) 2 (0.2%) 42	
48	**Assaultive Act – Male only (n=430)** **Contact with Pt's Penis by Assailant Genitals** No Yes Unknown Missing	 138 (32.1%) 135 (29.1%) 164 (38.1%) 3		
49	**Assaultive Act – Male only (n=430)** **Contact with Pt's Penis by Assailant Finger/Hand** No Yes Unknown Missing	 90 (20.9%) 189 (44%) 147 (34.2) 4		
50	**Assaultive Act – Male only (n=430)** **Contact with Pt's Penis by Object** No Yes Unknown Missing	 243 (58.8%) 12 (2.8%) 160 (37.2%) 5		
51	**Assaultive Act** **Contact with Pt's Mouth by Assailant Penis/Genitals** No Yes Unknown Attempted Missing	 5020 (55.9%) 2004 (22.3%) 1914 (21.3%) XXXX 43	 434 (36%) 213 (17.6%) 486 (40.3%) 39 (3.2%) 35	 748 (47.6%) 262 (16.7%) 379 (24.1%) XXXX 183
52	**Assaultive Act** **Contact with Pt's Mouth by Assailant Finger/Hand** No Yes Unknown Missing	 5523 (61.8%) 1307 (14.6%) 2108 (23.5%) 43		
53	**Assaultive Act** **Contact with Pt's Mouth by Assailant Mouth/Tongue**			Valid % (n=368)

	No Yes Unknown Missing	3038 (33.8%) 4046 (45.1%) 1851 (20.6%) 46		179 (48.6%) 82 (22.3%) 107 (29.1%) 1204
54	**Assaultive Act** **Contact with Pt's Mouth by Object** No Yes Unknown Missing	 6782 (75.5%) 130 (1.4%) 2011 (22.4%) 58		
55	**Assaultive Act** **Suspect Mouth Contact with Patient's Genitals** No Yes Unknown Attempted Missing	 4816 (53.6%) 1937 (21.6%) 2186 (24.3%) XXXX 42	 466 (38.6%) 189 (15.7%) 513 (42.5%) 8 (0.7%) 31	 804 (51.1%) 171 (10.9%) 415 (26.4%) XXXX 182
56	**Assaultive Act** **Suspect Mouth Contact with Patient's Breasts** No Yes Unknown Missing	 3675 (40.9%) 3085 (34.4%) 2172 (24.2%) 49		*Valid % (n=406)* 168 (41.4%) 130 (32%) 108 (26.6%) 1166
57	**Assaultive Act** **Suspect Mouth Contact with Patient's Mouth** No Yes Unknown Missing	 2710 (30.2%) 4360 (48.5%) 1868 (20.8%) 43		*Valid % (n=373)* 152 (40.8%) 113 (30.3%) 108 (29%) 1199
58	**Assaultive Act** **Suspect Mouth Contact with other parts of patient's body** No Yes Unknown Missing	 4271 (47.6%) 2471 (27.5%) 2142 (23.9%) 97		*Valid % (n=410)* 118 (45.9%) 114 (7.3%) 108 (26.3%) 1162
59	**Assaultive Act**			

	Assailant's hands touch patient's breasts No Yes Unknown Missing	 1349 (15%) 2892 (32.2%) 1821 (20.3%) 2919		
60	**Assaultive Act** **Assailant's hands touch patient's extremities** No Yes Unknown Missing	 1223 (13.6%) 3104 (34.6%) 1735 (19.3%) 2919		
61	**Assaultive Act** **Assailant's hands touch patient's other body parts** No Yes Unknown Missing	 1944 (21.6%) 2185 (24.3%) 1843 (20.5%) 3009		
62	**Number of assaultive/penetrative acts** Fondling (no penetration) 1 penetrative act 2 penetrative acts 3 penetrative acts 4 penetrative acts Unknown Missing	 239 (2.7%) 2822 (31.4%) 2304 (25.7%) 1182 (13.2%) 419 (4.7%) 1473 (16.4%) 542		 38 (2.4%) 648 (41.2%) 310 (19.7%) 88 (5.6%) 23 (1.5%) 276 (17.6%) 189
63	**Ejaculation Occurred** No Yes Unknown Missing	 1280 (14.3%) 2974 (33.1%) 4666 (52%) 61	 165 (13.7%) 304 (25.2%) 698 (57.8%) 40	 226 (14.4%) 391 (24.9%) 809 (51.5%) 147
64	**Ejaculation Site** Vagina Internal anus/rectum Internal oral cavity External genitalia External body site not genitalia External site, (i.e. bedding/clothing) not on patient	 1175 (13.1%) 169 (1.9%) 211 (2.3%) 35 (0.4%) 569 (6.3%) 339 (3.8%)	 180 (14.9%) 18 (1.5%) 27 (2.2%) 3 (0.2%) 61 (5%) 25 (2.1%)	*Valid % (n=138)* 95 (68.8%) 14 (10.1%) 7 (5.1%) 1 (0.7%) 18 (13%) 1 (0.7%)

	External site, NA (i.e. furniture, car seat, condom)	1351 (15%)	28 (2.3%)	2 (1.4%)
65	**Condom Use**			
	No	5784 (64.4%)	570 (47.2%)	794 (50.5%)
	Yes	641 (7.1%)	86 (7.1%)	81 (5.2%)
	Unknown	2487 (27.7%)	514 (42.6%)	550 (35%)
	Not Applicable	15 (0.2%)		
	Missing	54	81	147
66	**Lubrication**			
	No	5615 (62.5%)		
	Yes	794 (8.8%)		
	Unknown	2516 (28%)		
	Missing	52		
67	**Lubrication Type**			
	Assailant Saliva	344 (3.8%)		
	Commercial oil/lubricant	161 (1.8%)		
	Lotion/soaps	77 (0.9%)		
	Other/unknown product	18 (0.2%)		
68	**Patient Urinated Post-assault**			*Valid % (n=450)*
	No	1063 (11.8%)	184 (15.2%)	48 (10.1%)
	Yes	7685 (85.6%)	1023 (84.8%)	398 (88.4%)
	Unknown by patient	182 (2%)	0	4 (0.9%)
	Missing	51	0	1122
28	**Patient Defecated Post-assault**			*Valid % (n=419)*
	No	5189 (57.8%)	746 (61.8%)	231 (55.1%)
	Yes	3452 (38.4%)	461 (38.2%)	180 (43%)
	Unknown by patient	285 (3.2%)		8 (1.9%)
	Missing	55	0	1153
29	**Patient Vomited Post-assault**			Valid % (n=392)
	No	6600 (73.5%)	1011 (83.8%)	337 (86%)
	Yes	2082 (23.2%)	196 (16.2%)	51 (13%)
	Unknown	243 (2.7%)		4 (1 %)
	Missing	56	0	1180
	Patient Douched Post-assault			
	No		1199 (99.3%)	
	Yes		8 (0.7%)	
30	**Patient brushed teeth or gargled Post-assault**			*Valid % (n=417)*

	No	4819 (53.7%)	85.4%	244 (58.5%)
	Yes	3896 (43.4%)	14.6%	168 (40.3%)
	Unknown	208 (2.3%)		5 (1.2%)
	Missing	58	0	1155
31	**Patient Ate or Drank Post-assault**			*Valid % (n=89)*
	No	1524 (17%)	61%	73 (78.5%)
	Yes	5521 (61.5%)	39%	16 (17.2%)
	Unknown	805 (9%)		
	Missing	1131	0	1479
	Patient Washed/Wiped Genital Area			Valid % (n=95)
	No	3127 (34.8%)	363 (30.1%)	70 (73.7%)
	Yes	5614 (62.5%)	844 (69.9%)	25 (26.3%)
	Unknown	186 (2.1%)		
	Missing	54		1477
32	**Patient Bathed or Showered Post-assault**			
	No	5358 (59.7%)	746 (61.8%)	755 (48%)
	Yes	3419 (38.1%)	461 (38.2%)	487 (31%)
	Unknown	186 (2.1%)	0	51 (3.2%)
	Missing	54		279
33	**Patient removed/inserted tampon/pad/diaphragm Post-assault**			
	No	7644 (85.1%)		
	Yes	931 (10.4%)		
	Unknown	161 (1.8%)		
	Not Included	183 (2%)		
	Missing	62		
	Patient changed clothing Post-assault			
	No		480 (39.8%)	
	Yes		727 (60.2%)	
	Unknown			
	Missing			
	Physical Injury			
	No	2511 (28%)	431 (35.9%)	
	Yes	6372 (70.9%)	770 (63.8%)	
	Unknown	35 (0.4%)		
	Missing	63	6	
	Number of Physical Injuries			

	Mean Median Mode Std. deviation Minimum Maximum Percentiles 25% 50% 75% Missing	6.39 3.00 0 10.607 0 185 .00 3.00 8.00 146		
	Location of Physical Injury: Head No Yes Unknown Missing	 7428 (82.7%) 1451 (16.2%) 42 (0.5%) 60		
	Location of Physical Injury: Neck No Yes Unknown Missing	 7274 (81%) 1596 (17.8%) 49 (0.5%) 62		
	Location of Physical Injury: Breasts No Yes Unknown Missing	 7601 (84.6%) 1202 (13.4%) 67 (0.7%) 111		
	Location of Physical Injury: Chest/Back No Yes Unknown Missing	 6828 (76%) 1993 (22.2%) 59 (0.7%) 101		
	Location of Physical Injury: Abdomen No Yes Unknown Missing	 8118 (90.4%) 701 (7.8%) 62 (0.7%) 100		
	Location of Physical Injury: Extremities No Yes Unknown	 3396 (37.8%) 5447 (60.7%) 50 (0.6%)		

	Missing	88		
	Type of Physical Injury: Laceration No Yes Unknown Missing	 8270 (92.1%) 601 (6.7%) 49 (0.5%) 61		
	Type of Physical Injury: Abrasion No Yes Unknown Missing	 5320 (59.2%) 3555 (39.6%) 44 (0.5%) 62		
	Type of Physical Injury: Bruise No Yes Unknown Missing	 4042 (45%) 4832 (53.8%) 48 (0.5%) 59		
	Type of Physical Injury: Redness/Erythema No Yes Unknown Missing	 6886 (76.7%) 1989 (22.1%) 48 (0.5%) 58		
	Type of Physical Injury: Ecchymosis No Yes Unknown Missing	 8660 (96.4%) 211 (2.3%) 49 (0.5%) 61		
	Type of Physical Injury: Swelling No Yes Unknown Missing	 8009 (89.2%) 863 (9.6%) 47 (0.5%) 62		
	Type of Physical Injury: Petechiae No Yes Unknown Missing	 7926 (88.3%) 946 (10.5%) 49 (0.5%) 60		
	Type of Physical Injury: Incision No Yes	 8834 (98.4%) 35 (0.4%)		

	Unknown Missing	49 (0.5%) 63		
	Type of Physical Injury: Avulsion No Yes Unknown Missing	 8828 (98.3%) 41 (0.5%) 48 (0.5%) 64		
	Type of Physical Injury: Discolored Mark No Yes Unknown Missing	 8042 (89.5%) 829 (9.2%) 48 (0.5%) 62		
	Type of Physical Injury: Puncture Wound No Yes Unknown Missing	 8749 (97.4%) 124 (1.4%) 48 (0.5%) 60		
	Type of Physical Injury: Fracture No Yes Unknown Missing	 8852 (98.6%) 20 (0.2%) 49 (0.5%) 60		
	Type of Physical Injury: Bite Mark No Yes Unknown Missing	 8676 (96.6%) 197 (2.2%) 46 (0.5%) 62		
	Type of Physical Injury: Burn No Yes Unknown Missing	 8808 (98.1%) 66 (0.7%) 45 (0.5%) 62		
	Type of Physical Injury: Missing or broken tooth or teeth No Yes Unknown	 8845 (98.5%) 36 (0.4%) 43 (0.5%)		

	Missing	47		
	Type of Physical Injury: Conjunctival Hemorrhage No Yes Unknown Missing	 8804 (98%) 74 (0.8%) 44 (0.5%) 59		
36	**Genital Injury** No Yes Unknown Missing	 4501 (50.1%) 4111 (45.8%) 106 (1.2%) 263	 684 (56.7%) 413 (42.5%) 10	
	Number of Genital Injuries Mean Median Mode Std. deviation Minimum Maximum Percentiles 25% 50% 75% Missing	 1.5 .00 0 2.863 0 50 .00 .00 2.00		
	Location of Genital Injury: Inner Thighs No Yes Unknown Missing	 8165 (90.9%) 469 (5.2%) 104 (1.2%) 242		
	Location of Genital Injury: Clitoral Hood/Clitoris No Yes Unknown NA/male patient Missing	 8019 (89.3%) 127 (1.4%) 104 (1.2%) 430 (5%) 283		
	Location of Genital Injury: Labia Majora No Yes Unknown NA/male patient	 7610 (84.7%) 550 (6.1%) 103 (1.1%) 430 (5%)		

	Missing	268		
	Location of Genital Injury: Labia Minora No Yes Unknown NA/male patient Missing	 7308 (81.4%) 837 (9.3%) 110 (1.2%) 430 (5%)		
	Location of Genital Injury: Peri-urethral tissue/urethra No Yes Unknown NA/male patient Missing	 7997 (89%) 114 (1.3%) 120 (1.3%) 430 (5%) 300		
	Location of Genital Injury: Peri-hymenal tissue No Yes Unknown NA/male patient Missing	 7775 (86.6%) 322 (3.6%) 123 (1.4%) 430 (5%) 310		
	Location of Genital Injury: Hymen No Yes Unknown NA/male patient Missing	 7779 (86.6%) 300 (3.3%) 132 (1.5%) 430 (5%) 319		
	Location of Genital Injury: Vagina No Yes Unknown NA/male patient Missing	 7415 (82.6%) 364 (4.1%) 256 (2.9%) 430 (5%) 495		
	Location of Genital Injury: Cervix No Yes Unknown NA/male patient Missing	 7232 (80.5%) 399 (4.4%) 297 (3.3%) 430 (5%) 596		

	Location of Genital Injury: Fossa Navicularis No Yes Unknown NA/male patient Missing	 6108 (68%) 1993 (22.2%) 122 (1.4%) 430 (5%) 305		
	Location of Genital Injury: Posterior Fourchette No Yes Unknown NA/male patient Missing	 7219 (80.4%) 887 (9.9%) 123 (1.4%) 430 (5%) 302		
	Location of Genital Injury: Perineum No Yes Unknown NA/male patient Missing	 7753 (86.3%) 369 (4.1%) 117 (1.3%) 430 (5%) 291		
	Location of Genital Injury: Anal/Rectal No Yes Unknown NA/male patient Missing	 7176 (79.9%) 510 (5.7%) 184 (2%) 430 (5%) 673		
	Location of Genital Injury: Buttocks No Yes Unknown Missing	 5847 (65.1%) 307 (3.4%) 109 (1.2%) 2718		
	Location of Genital Injury: Male Perianal or perineum No Yes Unknown Missing	 383 (89.1%) 22 (5.1%) 8 (1.9%) 17		
	Location of Genital Injury: Male Anus No Yes Unknown	 306 (71.2%) 91 (21.2%) 10 (2.3%)		

	Missing	23		
	Location of Genital Injury: Male Rectum No Yes Unknown Missing	338 (78.6%) 17 (4%) 22 (5.1%) 53		
	Location of Genital Injury: Scrotum No Yes Unknown Missing	393 (91.4%) 8 (1.9%) 10 (2.3%) 19		
	Location of Genital Injury: Male Urethral Meatus No Yes Unknown Missing	399 (92.8%) 3 (0.7%) 10 (2.3%) 18		
	Location of Genital Injury: Penile Shaft No Yes Unknown Missing	387 (90%) 15 (3.5%) 10 (2.3%) 18		
	Location of Genital Injury: Glans Penis No Yes Unknown Missing	392 (91.2%) 9 (2.1%) 10 (2.3% 19		
	Type of Genital Injury: Laceration No Yes Unknown Missing	6656 (74.1%) 1954 (21.8%) 123 (1.4%) 248		
	Type of Genital Injury: Abrasion No Yes Unknown Missing	6597 (73.5%) 2010 (22.4%) 123 (1.4%) 251		

	Type of Genital Injury: Redness with tenderness			
	No	7545 (84%)		
	Yes	1065 (11.9%)		
	Unknown	122 (1.4%)		
	Missing	249		
	Type of Genital Injury: Bruise			
	No	7998 (89.1%)		
	Yes	613 (6.8%)		
	Unknown	119 (1.3%)		
	Missing	251		
	Type of Genital Injury: Swelling			
	No	8262 (92%)		
	Yes	344 (3.8%)		
	Unknown	123 (1.4%)		
	Missing	252		
	Type of Genital Injury: Ecchymosis			
	No	8570 (95.4%)		
	Yes	37 (0.4%)		
	Unknown	122 (1.4%)		
	Missing	252		
	Type of Genital Injury: Petechiae			
	No	8479 (94.4%)		
	Yes	129 (1.4%)		
	Unknown	124 (1.4%)		
	Missing	249		
	Type of Genital Injury: Discolored			
	mark	8497 (94.6%)		
	No	111 (1.2%)		
	Yes	122 (1.4%)		
	Unknown			
		251		
	Missing			
	Type of Genital Injury: Avulsion			
	No	8593 (95.7%)		
	Yes	16 (0.2%)		
	Unknown	123 (1.4%)		
	Missing	249		
	Type of Genital Injury: Puncture			
	Wound	8600 (95.8%)		
	No	8 (0.1%)		

	Yes Unknown Missing	125 (1.4%) 248		
	Descriptive Findings of Crime Lab Data *Percentages listed as valid percent based upon the denominator of the submitted sexual assaults*			
38	**Location Of Analysis** (n=6834) UBFS Outsourced to BODE Private lab	 3798 (55.6%) 3033 (44.4%) 3 (0%)		
	Number of items* analyzed <3 or 4 or more within submitted evidence per case 3 or less items analyzed 4 or more items analyzed *Items defined as swabs from distinct body area, clothing, bedding, or other items with evidence collection	(n=6865) 5244 (76.4%) 1621 (23.6%)	(n=1207) 1169 (96.9%) 38 (3.1%)	(n=1543) 531 (34.4%) 1012 (64.4%)
	Swabs selected for male quant (Y-screen) or initial DNA screening Perianal* Vaginal Breast(s) Rectal Cervical Oral Body area, not including neck/breast Neck Underwear Other clothing Other Items not clothing/bedding Bedding Tampon Condom *For males, this includes all external genitalia swabs	(N=6865) (n=3574) 52% (n=3273) 47.7% (n=1503) 21.9% (n=1031) 15% (n=772) 11.8% (n=442) 6.7% (n=908) 13.2% (n=925) 13.5% (n=59) 0.9% (n=51) 0.8% (n=20) 0.3% (n=14) 0.2% (n=6) 0.09% (n=18) 0.2%	(N=1207) (n=546) 45.2% (n=282) 23.4% (n=252) 20.9% (n=32) 2.7% (n=57) 4.7% (n=63) 5.2% (n=126) 10.4% (n=112) 9.3% (n=11) 0.9% (n=5) 0.4% (n=2) 0.2% XXXX XXXX XXXX	(N=1543) (n=455) 28.3% (n=734) 46.7% (n=204) 13% (n=390) 24.8% (n=35) 2.2% (n=313) 19.9% (n=199) 12.7% (n=185) 11.8% (n=16) 1% (n=8) 0.5% (n=11) 0.7% (n=2) 0.13% (n=3) 0.19% (n=8) 0.5%
40	**Swabs with positive male quant (Y) DNA screening** Perianal* Vaginal Rectal	(n=6865) 2926 (42.6%) 2544 (37.1%) 656 (9.6%)		(n=1572) 415 (26.4%) 675 (42.9%) 320 (20.4%)

	Breast(s)	1179 (17.2%)		184 (11.7%)
	Cervical	770 (11.2%)		33 (2.1%)
	Oral	234 (3.6%)		172 (11.3%)
	Body area, not including neck/breast	726 (11.1%)		187 (12.2%)
	Neck	779 (11.8%)		179 (11.7%)
	Underwear	56 (0.9%)		15 (1%)
	Other clothing	51 (0.8%)		7 (0.5%)
	Other Items not clothing/bedding	18 (0.3%)		11 (0.7%)
	Bedding	14 (0.2%)		1 (0.07%)
	Tampon	6 (0.09%)		2 (0.13%)
	Condom	18 (0.3%)		7 (0.5%)
	*For males, this includes all external genitalia swabs			
	Swabs with full or partial STR DNA profile of foreign contributor (denominator is all SAKs rather than the # of tested swabs per site as noted in Tables 2 and 3. Refer to Tables 2 and 3.)	(n=6865)	(n=1207)	(n=1572)
	Perianal*	1317 (19.1%)	531 (44%)	137 (8.7%)
	Vaginal	1206 (17.6%)	265 (22%)	310 (19.7%)
	Rectal	253 (3.8%)	30 (2.5%)	107 (6.8%)
	Breast(s)	607 (9.2%)	234 (19.4%)	91 (5.8%)
	Cervical	336 (5.1%)	54 (4.5%)	14 (0.9%)
	Oral	54 (2.5%)	61 (5.1%)	8 (0.5%)
	Body area, not including neck/breast	278 (4.2%)	120 (9.9%)	70 (4.5%)
	Neck	351 (10.8%)	110 (9.1%)	93 (5.9%)
	Underwear	32 (0.5%)	9 (0.7%)	10 (0.6%)
	Other clothing	28 (0.4%)	1 (0.1%)	3 (0.2%)
	Other Items not clothing/bedding	12 (0.18%)	XXXX	4 (0.25%)
	Bedding	6 (0.09%)	XXXX	1 (0.07%)
	Tampon	3 (0.05%)	XXXX	0 (0%)
	Condom	12 (0.18%)	XXXX	7 (0.5%)
	*For males, this includes all external genitalia swabs			
41	**Number OF items/Swabs Tested**	Neck (n=925)		
	Mean	13.5%		4.26
	Median			4.00
	Mode			4
	Std. Deviation			1.739
	Variance			
	Min			0
	Max			16
	Percentiles			

	25 50 75 Missing			3.00 4.00 5.00 29
58	**Serology Done Before DNA at case level** (n=6865) No Yes, negative results Yes, positive for amylase Yes, positive for micro Yes, positive for PSA Yes, positive for amylase and SF	Body area, not including neck/breasts (n=908) 13.2%		700 (44.5%) 373 (23.7%) 67 (4.3%) 229 (14.6%) 34 (2.2%) 94 (6%)
59	**Male Quant DNA Found at case level** (n=6821) No Yes, female victim Male victim Female on female assault	Cervical (n=772) 11.8%		
122	**Swab From Suspect with Victims' DNA** (n=5905) No suspect exam noted Yes, penile suspect swab or other body location	Oral (n=442) 6.7%		1480 (96.6%) 52 (3.4%)
123	**Excluded Suspect by DNA Analysis** (n=5933) No Yes	Underwear (n=59) 0.9%		1426 (93.8%) 95 (6.2%)
124	**Suspect Standard Submitted** (n=6809) No Yes Missing	Other clothing (n=51) 0.8%		908 (57.8%) 622 (39.6%) 42
125	**Consensual Partner Standard Submitted** (n=6809) No Yes Missing	Other items, not clothing or bedding (n=20) 0.3%		1470 (93.5%) 62 (3.9%) 40
	Case Level STR DNA findings	Condom (n=18) 0.2%		

	(n=6810) STR DNA testing not completed Full or partial STR DNA foreign contributor profile developed Low Level of STR DNA of foreign contributor or complex mixture, inconclusive			
127	**STR DNA Profile Entered into CODIS NDIS** No Yes, uploaded	Bedding (n=14) 0.2%		 1028 (65.4%) 503 (32%)
128	**STR DNA Profile Entered into CODIS SDIS** No Yes, uploaded	Tampon (n=6) 0.09%	 648 (53.7%) 559 (46.3%)	 1009 (64.2%) 524 (33.3%)
	CODIS Profile Hit No Yes		 963 (79.8%) 244 (20.2%)	
Orange County Crime Lab Data *ONLY*				
94	Swab1VaginalLoci (n=282) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 22.50 24.00 24 5.10 0 25 24.00 24.00 24.00	
95	Swab2CervicalLoci (n=57) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 22.53 24.00 24 5.389 0 24 24.00 24.00 24.00	

96	Swab3PerianalExtGenitaliaLoci (n=546) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 23.08 24.00 24 2.329 0 25 22.00 24.00 24.00	
97	Swab4RectalLoci (n=32) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 22.81 24.00 24 4.425 0 24 24.00 24.00 24.00	
98	Swab5OralLoci (n=63) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 23.17 24.00 24 3.088 0 25 23.00 24.00 24.00	
99	Swab6BodyBreastLoci (n=252) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 22.75 24.00 24 3.971 0 25 22.00 24.00 24.00	

100	Swab7BodyNeckLoci (n=113) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 23.11 24.00 24 3.288 0 25 23.00 24.00 24.00	
101	Swab8BodyOtherLoci (n=126) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 23.00 24.00 24 3.705 0 25 23.00 24.00 24.00	
102	Swab9UnderwearLoci (n=10) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75 Missing		 21.20 24.00 24 7.495 0 24 22.00 24.00 24.00 1347	
103	Swab10OtherClothingLoci (n=6) Mean Median Mode Std. Deviation Min Max Percentiles 25 50 75		 7.67 .00 0 11.894 0 24 .00 .00 22.50	

	Missing			