A cost-effectiveness analysis of the number of samples to collect and test from a sexual assault

Zhengli Wang, Kevin MacMillan, Mark Powell, and Lawrence M. Wein

Although the backlog of untested sexual assault kits in the United States is starting to be addressed, many municipalities are opting for selective testing of samples within a kit, where only the most probative samples are tested. We use data from the San Francisco Police Department Criminalistics Laboratory, which tests all samples but also collects information on the samples flagged by sexual assault forensic examiners as most probative, to build a standard machine learning model that predicts which samples to test from each kit to maximize the Combined DNA Index System (CODIS) yield. Our results suggest that electronic records coupled with machine learning and optimization models could enhance the effectiveness of criminal investigations of sexual assaults.

Significance

Within the context of sexual assaults, we address a fundamental issue in criminal investigations: how much evidence to collect and process. Using data from the San Francisco Police Department, we show that machine learning algorithms outperform sexual assault forensic examiners at identifying probative samples. Relative to selective testing of samples, testing all DNA samples in a sexual assault kit more than doubles the number of sexual assault kits generating a DNA profile that can be entered into the criminal DNA database, at only a slightly lower benefit-to-cost ratio. Our results suggest that the yield of DNA profiles for the database would increase another 47.2% by collecting samples from the three most probative locations (as deemed by the machine learning algorithm).
examiners (SAFEs) (if these examiners are nurses, they are often referred to as sexual assault nurse examiners [SANEs]) recommended the most probable samples for testing. However, despite being given this information, the forensic laboratory tested all samples in the kit, which allows us to assess the effectiveness of the choices made by the SAFEs. We manually code the SAK questionnaires associated with these SAKs to obtain values with material Appendix, leaving only one continuous variable (i.e., per sample) testing costs of a SAK. We construct a standard machine-learning model that predicts the probability of obtaining a DNA profile that is of sufficiently high quality to be uploaded into CODIS—we hereafter refer to such a DNA profile as being CODIS uploadable—from a given sample based on the covariates obtained from the associated police report and then propose a SAK testing policy that attempts to maximize the number of SAKs that yield at least one CODIS-uploadable DNA profile subject to a budget constraint. It is worth stressing that we are maximizing the number of SAKs that generate at least one DNA profile that can be uploaded into CODIS and do not attempt to maximize the number of matches, or hits, to existing DNA profiles in CODIS that these uploaded profiles generate. This latter, downstream metric is briefly addressed at the end of Discussion.

The analysis allows us to address three research questions: 1) Relative to a machine-learning model, how well do SAFEs predict the most probative samples? 2) What is an optimal SAK testing policy and how does its performance compare to the existing selective testing strategy? And 3) could performance be further enhanced if SAFE chose more probative samples to obtain in the first place?

Materials and Methods

Data. Our dataset consists of all 913 SAKs handled by the San Francisco Police Department Criminalistics Laboratory with sexual assault dates ranging from September 27, 2016 to May 25, 2019. For each sexual assault, we extract from the SAK questionnaire the values for 23 covariates listed in Table 1, which includes characteristics about the victim, the offender(s), and the assault. We discard 45 SAKs that have missing values for the time delay between assault and examination and/or victim age, leaving us to study 868 SAKs; the 45 SAKs do not appear to systematically vary (e.g., by date of sexual assault) from the other 883 SAKs. We convert these data into one-hot encoding format, as in SI Appendix, Table S1, leaving only one continuous variable (victim age), which is standardized so that it has zero mean and unit variance across the observations, as is common practice in data preprocessing.

Roughly half of the covariates suffer from unknown values for roughly half of the sexual assaults (Table 1), and unknown values are treated as a separate category (Table S1). A visual inspection of the proportion of covariate values that are missing versus the date of the sexual assault revealed no underlying temporal pattern. These data entries are unknown for two reasons. First, the victim could not recall, which affects variables related to, e.g., ejaculation and condom use. Second, the SAFE did not record the information either because the information was negative or due to oversight, which affects variables such as the type of injuries incurred.

A total of 6,318 samples were tested from the 868 SAKs (mean 7.28, range 1–36). We have data on the specific location of each of the 6,318 samples. We aggregate the different sample locations in the raw data into six categories (Table 2), using the aggregation scheme in SI Appendix, Table S2. In terms of frequency, these six locations fall into three buckets: There are more than twice as many samples from body surface and genital locations than from oral and anal locations, and there are relatively few samples from clothing and foreign material (Table 2).

For each of these SAKs, a subset of the tested samples was identified as probative by a SAFE during the forensic medical examination; we refer to these samples as probative. Overall, 1,848 of the 6,318 (29.2%) samples are probative, giving a mean of 2.13 probative samples per SAK and a range of [0,8], with 70 (8.1%) SAKs having no probative samples. More than half of the probative samples are from the genital location, with most of the remaining samples roughly evenly distributed among the anal, body surface, and oral locations (Table 2).

Table 1. List of covariates describing the sexual assault, along with the number of SAKs (out of 868 SAKs) that had these values for each covariate (and the average victim age)

| Covariate | Values |
|-----------|--------|
| Time delay between assault and examination (0 d/1 d/≥ 2 d) | 315/298/255 |
| Victim age (y) | Average = 31.6 |
| Loss of memory (Y/N/U) | 150/718 |
| Consensual sex in prior 5 d (Y/N/U) | 201/615/62 |
| Known ejaculation (Y/N/U) | 176/102/590 |
| Condom used (Y/N/U) | 71/327/470 |
| Shower or bath before examination (Y/N/U) | 287/510/71 |
| Vaginal penetration of victim by offender (Y/N/U) | 331/127/320 |
| Anal penetration of victim by offender (Y/N/U) | 153/304/411 |
| Oral penetration of victim by offender (Y/N/U) | 110/315/443 |
| Offender’s mouth on genitals (Y/N/U) | 109/318/441 |
| Offender’s mouth on breasts (Y/N/U) | 88/343/439 |
| Offender’s mouth on other body parts (Y/N/U) | 228/225/415 |
| Digital penetration of victim by offender (Y/N/U) | 170/237/461 |
| Oral penetration of offender by victim (Y/N/U) | 170/278/480 |
| Strangled (Y/N/U) | 86/732/50 |
| Punched (Y/N/U) | 92/722/54 |
| Stabbed (Y/N/U) | 3/837/28 |
| Vaginal injury (Y/N/U) | 22/652/22 |
| Other injury (Y/N/U) | 343/507/18 |

Abbreviations: M, male; F, female; Y, yes; N, no; and U, unknown.

For each sample of each SAK, we also know whether it satisfies the criteria for uploading into CODIS; we refer to such samples as CODIS uploadable. Overall, 1,159 of the 6,318 (18.3%) samples are CODIS uploadable, giving a mean of 1.34 CODIS-uploadable samples per SAK and a range of [0,12], with 461 (53.1%) SAKs having no CODIS-uploadable samples. Just over one-half and one-quarter of the CODIS-uploadable samples are from the body surface and genital locations, respectively, followed by clothing, anal, and oral locations and foreign material (Table 2).

Finally, we note one procedural change that occurred during the timeframe under study. Prior to November 2017 (which covers 341 of the 868 SAKs, or 39.3%), prescreening for biological fluid (semen and saliva testing) was performed before DNA processing, and if the prescreening results were negative (which occurred in 15 cases), then no DNA testing was performed. Starting in November 2017, all SAKs bypassed prescreening and went directly to DNA processing. Currently, most large laboratories bypass prescreening, but many small laboratories still do prescreening. We investigate prescreening in a sensitivity analysis in Results.

Machine-Learning Models. We assess three standard machine-learning models (13): logistic regression (LR), logistic regression with L1 regularizer (LASSO-LR), and classification and regression tree (CART). We have six aggregated locations in our model (Table 2), and we estimate the probability of obtaining a CODIS-uploadable profile from each location separately. We are not incorporating statistical dependence of the response variables (and thus not using methods such as generalized estimating equations) because it is not clear what the correlation structure might be, and introducing a general correlation structure into the model will introduce too many additional variables to be estimated given the limited amount of data we have. However, the probabilities of obtaining a CODIS-uploadable profile from different locations within a SAK are correlated due to having the same covariate values. Nonetheless, it is possible that other dependencies are present due to at least three factors: All samples within a SAK are obtained by the same SAFE and are processed together in the same batch at the crime laboratory, and some sample locations are in close physical proximity (e.g., genital and anal). We return to this issue in Results and Discussion.

Because the LASSO-LR model is used to present our main results and because all three models have been in use for decades, we describe the LASSO-LR model here and relegate the descriptions of the LR and CART models to SI Appendix, section 1. LASSO-LR formulates the problem of
maximizing the data likelihood while keeping the set of nonzero elements in the estimated parameters to be small. This is achieved by adding an ℓ1 regularizer term, commonly known as the LASSO penalty term, to the likelihood function. Assuming there are \( n \) tested samples from location \( i \) in the training set, and given \( n_i \) \( d \)-dimensional covariate vectors \( x_i^\alpha \) and their corresponding binary labels \( y_i^\beta \) (which equals one if the tested sample from location \( i \) is uploadable into CODIS and equals zero if it is not), we calculate the maximum-likelihood estimates separately for \( i = 1, \ldots, 6 \), using all of the covariates in SI Appendix, Table S1,

\[
\hat{\beta}_{\text{LASSO}}(\hat{\lambda}) = \arg \max_{\beta \geq 0} - \sum_{i=1}^{n} \log h(y_i^\beta \mid x_i^\beta) + \lambda \sum_{i=1}^{d} |\beta_i^\alpha|,
\]

where \( \lambda \) is known as the regularization parameter. Usually, the larger \( \lambda \) is, the fewer nonzero \( \beta \) values there are.

We randomly divide the entire dataset into a training set, a validation set, and a testing set, using the ratio 5:1:4. For each location \( i \), we perform the LASSO-LR on the training set using different values of \( \lambda \), use the validation set for choosing the best \( \lambda \), and then use the test set to measure performance. Our performance metric used to maximize the data likelihood while keeping the set of nonzero elements to be small is achieved by adding an ℓ1 constraint on the mean cost per SAK, where we require so as not to violate the spirit of testing the background—that at least one sample from each SAK be tested. This optimization problem uses sample estimates for the probability that a sample from each location of each SAK is CODIS uploadable, as predicted by our machine-learning model. An optimal solution to this problem is likely to be complex (16), and we resort to a simple greedy algorithm in SI Appendix, section 2. In Results, this greedy algorithm is referred to as the nonlinear priority policy.

### Data Availability

The data used in this study appear in SI Appendix. 

#### Results

**Machine-Learning Results.** The LR results for all six locations appear in SI Appendix, Tables S6–S11. The normalized AUC (and 95% confidence interval) of the plot of the number of SAKs in CODIS versus the number of samples per kit (SI Appendix, Fig. S1) is 0.796 ± 0.014.

The LASSO-LR results appear in SI Appendix, Tables S12–S17. The normalized AUC of the plot of the number of SAKs in CODIS versus the number of samples per kit (SI Appendix, Figs. S2, S3) is 0.800 ± 0.018.

The CART network for one of the 100 runs appears in SI Appendix, Figs. S3–S8 for each of the six locations, and the results are summarized in SI Appendix, Tables S18–S23. The normalized AUC of the plot of the number of SAKs in CODIS versus the number of samples per kit (SI Appendix, Fig. S9) is 0.786 ± 0.016.

In summary, all three models achieve nearly the same normalized AUC and all outperform the SAFE policy, which tests only the samples deemed probative by a SAFE (SI Appendix, Figs. S1, S2, and S9). All our cost-effectiveness results are presented using the LASSO-LR model, although we perform a sensitivity analysis using the other two models.

Before moving on to our main results, we address two issues: the possibility that the superiority of the machine-learning algorithms over the SAFES is due to the concavity of the curve in SI Appendix, Fig. S2 and possible correlation among samples within a SAK. It is known that even if the individual SAFES are all operating somewhere along the machine-learning curve in SI Appendix, Fig. S2, their aggregate performance would fall below the curve due to Jensen’s inequality and the concavity of the curve (e.g., refs. 17 and 18).

### Sample Location

| Sample location | No. of samples | No. of probative samples | No. of CODIS uploadable samples |
|-----------------|----------------|--------------------------|--------------------------------|
| Body surface    | 2,364          | 275                      | 594                            |
| Genital         | 1,932          | 1,014                     | 298                            |
| Oral            | 939            | 223                      | 71                             |
| Anal            | 732            | 310                      | 87                             |
| Clothing        | 287            | 19                       | 102                            |
| Foreign material| 64             | 7                        | 7                              |
| Total           | 6,318          | 1,848                     | 1,159                          |
Table 3. The ratio of the fixed cost to the variable cost, $F/V$, for the four scenarios

|                          | No prescreening | Prescreening |
|--------------------------|----------------|--------------|
| San Francisco salary     | 1.85           | 1.18         |
| Average salary           | 1.65           | 1.10         |

The upper left scenario corresponds to our base case.

the curve in SI Appendix, Fig. S2 that is nearly linear; e.g., 91.6% of SAKs had fewer than or equal to three samples identified as probative. Second, our data allow us to directly observe that the samples identified as probative by the SAFEs do not align well with the samples identified as most likely to be CODIS uploadable by the machine-learning model. More specifically, among SAKs for which SAFEs identified exactly one probative sample, 95.7% of these samples were not the top sample identified by the machine-learning model; among SAKs for which SAFEs identified exactly two probative samples, 76.9% of these samples were not among the top two samples identified by the machine-learning model; and among SAKs for which SAFEs identified exactly three probative samples, 51.1% of these samples were not among the top three samples identified by the machine-learning model.

To assess the correlation among samples within a SAK, we compute the partial correlation of the yield $y^{(i)}_j$ across locations $i = 1, \ldots, 6$ within a SAK, i.e., the correlation of $y^{(1)}_j, \ldots, y^{(6)}_j$ conditioned on the covariate vector $x^{(i)}_j$. Using a logistic regression model with all possible covariates (not just the covariates specified in SI Appendix, Tables S4 and S5, and so now $x^{(i)}_j$ is independent of $i$ and will be denoted by $x_j$), we compute the partial correlation between $y^{(1)}_j$ and $y^{(2)}_j$ for each pair $(i_1, i_2)$ by

$$P(y^{(i_1)}_j, y^{(i_2)}_j|x) = \frac{\sum (y^{(i_1)}_j - E[y^{(i_1)}_j|x]) (y^{(i_2)}_j - E[y^{(i_2)}_j|x])}{\sqrt{\sum (y^{(i_1)}_j - E[y^{(i_1)}_j|x])^2 \times \sum (y^{(i_2)}_j - E[y^{(i_2)}_j|x])^2}}.$$  

[2] with standard error

$$\sqrt{1 - P(y^{(i_1)}_j, y^{(i_2)}_j|x)} \times \frac{1}{n - 2},$$  

[3] where $n$ is the number of SAKs that have at least one sample from each of locations $i_1$ and $i_2$ (19). We find that there is statistically significant positive partial correlation for most pairs of locations (SI Appendix, Table S3). The largest values are between anal and foreign material (e.g., condoms) and between genital and anal, suggesting that the proximity of locations plays a role in these correlations. The implications of this correlation are addressed in Discussion.

Cost-Effectiveness Results. We compare the performance of the nonlinear priority policy derived earlier to that of two simpler policies. One is the SAFE policy, which tests only the probative samples as deemed by the SAFEs. The other is the priority policy, which ranks each sample in a SAK by its probability of being CODIS uploadable. For a given value of the parameter $n$, we test the top $n$ samples from each SAK by its probability of being CODIS uploadable. For a given value of the parameter $n$, we test the top $n$ samples from each SAK (if there are fewer than $n$ samples in the SAK, we test all samples in the SAK). By varying $n$ from 1 to 20, we generate a tradeoff curve of the probability a SAK is CODIS uploadable (i.e., the probability at least one CODIS-uploadable sample is tested) versus the average cost per SAK. This policy can be viewed as a simplification of the nonlinear priority policy, where we are restricting ourselves to exactly $n$ tested samples from each SAK and replacing the quantity in SI Appendix, section 2, Eq. 7 by $p_{ij}$, which would be the appropriate quantity if the objective function in SI Appendix, section 2, Eq. 3 was changed to the mean number of CODIS-uploadable samples per kit. The tradeoff curve for the nonlinear priority policy is generated by using the two-stage greedy algorithm derived in SI Appendix, section 2 for various values of the budget $B$. A numerical example to illustrate how we compute the CODIS yield appears in SI Appendix, section 3.

Our main results appear in Fig. 1, where full testing corresponds to the right endpoint of the nonlinear priority policy curve. Note that the fixed cost associated with testing a SAK causes the lower left portion of the nonlinear priority policy curve to be slightly convex, whereas the decreasing marginal returns to testing samples cause the upper right portion of the nonlinear priority policy curve to be concave. Relative to the SAFE policy, full testing increases the CODIS yield more than twofold, from 0.229 to 0.466, and also increases the mean cost per SAK by a slightly larger ratio, from $397$ to $912$. For a budget of $397$, the nonlinear priority policy increases the CODIS yield from the SAFE policy’s value of 0.229 to 0.333 (a 45.4% increase). The performance of the priority policy is nearly indistinguishable from the performance of the nonlinear priority policy (Fig. 1); because the former policy is much easier to implement than the latter, we hereafter consider the priority policy in lieu of the nonlinear priority policy.

To put the results in Fig. 1 into better perspective, we note that the benefit-to-cost ratio of full testing has been estimated to be 81.34 in Detroit, MI (3); i.e., every dollar spent on testing a SAK saves on average $81.34 in the cost associated with future sexual assaults that are averted due to testing. Using this value allows us to convert from a cost-effectiveness analysis to a cost-benefit analysis. In Fig. 2, we equate the benefit-to-cost ratio of full testing to 81.34, which generates a benefit-to-cost ratio for the SAFE policy of $\frac{0.229 \times 81.34}{0.333} = 61.77$, and then compute the marginal benefit-to-cost ratio of the priority policy by taking the derivative of the priority policy curve in Fig. 1 and multiplying it by $\frac{81.34}{91.77}$. We also transform the horizontal axis from the mean cost per SAK to the proportion of samples tested. This marginal benefit-to-cost ratio increases from 125 to nearly 150 at approximately the cost of the SAFE policy and then drops below 91.77 when 44% of samples are tested. However, the marginal benefit-to-cost ratio, although steadily decreasing after this point, remains large in absolute terms throughout most of the testing: e.g., it is 24.3 when 81.9% of samples are tested and is 15.9 when 91.7% of samples are tested.

A comparison of these policies has revealed how much improvement is possible by using machine learning to predict the likelihood that each sample will end up in CODIS. However, we note that as predicted by the machine-learning model, 57.8% of the SAKs did not have any samples from its best location, 38.4% did not have any samples from its second-best location, and 33.9% did not have any samples from its third-best location. These high omission rates occur for two main reasons. First, 30% of SAKs with oral penetration of the victim by the offender did not have any oral samples, and 12% of SAKs with anal penetration did not have any anal samples. These omissions are unlikely to be due to victims failing to provide informed consent to the testing of certain body locations because only 1% of SAKs with vaginal penetration failed to obtain a genital sample. Second, 51% of SAKs had clothing as the best location, and a clothing sample was not obtained in 69% of these SAKs.

To assess how much further improvement could be achieved if SAFEs had obtained a sample from these top locations during
the forensic medical examination, we recompute the priority policy under the hypothetical assumption that a sample from each of the top three locations, respectively, was available. That is, we generate a synthetic sample for each of the top three locations of each SAK that had no samples, use the LASSO-LR model to compute the probability that the synthetic sample is CODIS uploadable, and then recompute the performance of the priority policy. We assume that collecting these synthetic samples is free: While the labor and material costs of collecting additional samples are minuscule compared to the cost of processing these samples, we are also ignoring any marginal psychological costs associated with collecting additional samples. The addition of these synthetic samples (top curve in Fig. 1) increases the yield of full testing by 47.2% (from 0.466 to 0.685) while increasing the cost of full testing by only 30.1% (from $912 to $1,194), leading to a benefit-to-cost ratio of 91.43, which is almost identical to the benefit-to-cost ratio of the SAFE policy. Relative to the SAFE policy, the full testing policy with synthetic samples increases the yield by 3-fold to 0.685, and the priority policy with synthetic samples increases the yield by 2.26-fold, from 0.229 to 0.517 at the SAFE cost of $397.

We perform two types of sensitivity analyses. First, we recompute Fig. 1 using the LR and CART models, and the results are very similar (SI Appendix, Figs. S10 and S11). Next we recompute Fig. 1 using the other three cost scenarios in Table 3 (SI Appendix, Figs. S12–S14). Our results are quite insensitive with respect to the four scenarios in Table 3. In fact, because the fixed and variable costs transform only the horizontal axis in Fig. 1, the increase in the CODIS yield achieved by the priority policy relative to the SAFE policy (at the same cost) is 41.5%, regardless of the values of $F$ and $V$.

**Discussion**

Our analysis provides quantitative answers to our three main research questions. First, following in a long line of research by psychologists showing the superiority of model-based judgment over expert-based judgment (20), a standard machine-learning algorithm appears to outperform the SAFEs at choosing the most probative samples: For the same average number of samples tested, the number of CODIS entries increases by 22.0% (SI Appendix, Fig. S2). Second, the priority policy, which performs nearly identically to the much more complex nonlinear priority policy, outperforms the SAFE policy, which tests only the samples flagged by the SAFEs. For the same cost as the SAFE policy, the priority policy increases the number of SAKs that are entered into CODIS by 41.5%. Part of this improvement is due to the superiority of the machine-learning algorithm over the SAFEs’ decisions, and part is due to optimally exploiting the economies of scale inherent in DNA processing. Moreover, the benefit-to-cost ratio is somewhat similar for full testing (which is estimated to be 81.34 in ref. 5) and the SAFE policy (estimated to be 91.77), although the former policy more than doubles the CODIS yield; i.e., the additional effectiveness achieved by testing only samples deemed probative by SAFEs is mostly offset by the lack of economies of scale associated with testing so few samples per kit. Taken together, these results provide strong support for testing all samples in a SAK, as is currently done in the San Francisco Police Department Criminalistics Laboratory.

The testing of samples is a multistep batch process, where results (i.e., uploadable vs. not uploadable) are not obtained until all samples have been processed through all steps; this makes sequential testing unattractive. For example, a policy that is sometimes used in practice is a two-stage policy that tests only the probative samples in the first stage and—for the SAKs that do not yield a CODIS-uploadable profile among its probative samples—tests the remaining samples in the second stage (thereby incurring an additional fixed cost $F$). It is clear from Fig. 1 that this policy would achieve the same CODIS yield as full testing, but at a higher mean cost, and hence is clearly sub-optimal. Nonetheless, conditioned on having already performed first-stage testing, our results suggest that many more CODIS entries could be generated by performing follow-up testing of the remaining samples. We are unaware of any published results on the amount of follow-up testing that is being performed

**Fig. 1.** Under the LASSO-LR model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the mean cost per SAK, under the SAFE policy (*), the priority policy (green solid line), the nonlinear priority policy (blue dotted line), and the priority policy with additional synthetic samples (red dashed line). The 95% CIs are depicted for each integer value of the parameter $n$ for the priority policy and for each integer value of the mean number of samples tested per SAK for the nonlinear priority policy and the priority policy with additional synthetic samples.

**Fig. 2.** Under the LASSO-LR policy, the marginal benefit-to-cost ratio of the priority policy vs. the proportion of samples tested. The benefit-to-cost ratio is 91.77 for the SAFE policy and 81.34 for full testing.
or its effectiveness, but this is an important issue for future research.

Our model predicts that significant further improvements could be achieved if SAFEs obtained more of the samples deemed most probative by the machine-learning model. Under full testing, an additional sample (if it is currently missing from the SAK) from the three most probative locations increases the CODIS yield by 47.2% and represents a threefold increase over the CODIS yield of the SAFE policy. In particular, for SAKs where the victim has bathed or showered prior to the examination, an effort should be made to obtain unwashed clothing samples, and oral and anal samples should be taken if there is oral or anal penetration.

With respect to implementing our results, the coordinated team approach advocated by the US Department of Justice notes that the purposes of the forensic medical examination “are to address patients’ health care needs and collect evidence suitable for possible use by the criminal justice system” (ref. 21, p. 4). It is important that any use of machine-learning or optimization models does not interfere with the focus on the patients’ healthcare needs and their right of informed consent regarding evidence collection.

Recall that our machine-learning models ignore any correlation across samples within a SAK, except for the conditioning on the common covariates. If one were to cluster at the SAK level, the standard errors would probably increase, and hence the CIs in SI Appendix, Tables S6–S17, which were computed under an independent and identically distributed assumption, should be taken with care. However, because we measure the performance of various policies using the actual SAKs, the performance of these specific policies incorporates this omitted correlation. This is not true of the top curve in Fig. 1 and SI Appendix, Figs. S10–S14 because the synthetic samples are generated using the LASSO-LR model, which assumes independence across the samples in a SAK (after conditioning on the covariates). We hypothesize that the omitted correlation in the synthetic samples leads to a slight overestimate of the performance of this top curve because the usefulness of additional samples is likely to be smaller under positive correlation. Moreover, it is possible that using a more sophisticated machine-learning model that incorporates the omitted correlation would lead to improved, albeit more complicated, policies, although—as noted earlier—such a model would require a larger dataset than we have here. We hypothesize that incorporating the positive correlation would lead to proposed policies that test slightly fewer samples per SAK (again, because the usefulness of additional samples is likely to be smaller under positive correlation). Nonetheless, we suspect that the impact of ignoring correlation is small for both of these issues (i.e., the overestimate of the top curve in Fig. 1 and the suboptimality of our proposed policy) because the normalized AUC of the LASSO-LR model is quite large (0.800) and hence the correlated residuals are relatively small in magnitude.

Because it is difficult to assess how generalizable our findings are, it is important to repeat this analysis using data from other municipalities, ideally with a larger number of SAKs. One variable that may vary across municipalities is the number of samples obtained and tested per kit. The mean of 7.28 samples tested per SAK is very similar to the 7.5 samples tested per kit in Oakland, CA, which also used full testing (22). A survey of US crime laboratories associated with Project FORESIGHT suggests that the total number of DNA samples tested per criminal case is 4.29 (tables 6 and 9 in ref. 14). We also note that the CODIS yield under full testing in Detroit, MI (3) was $\frac{2,934}{18,803} = 0.155$, and the CODIS yield in the Manhattan District Attorney’s Office’s Sexual Assault Kit backlog Elimination Grant Program, which which used the three most probative samples per SAK, was $\frac{0.460}{0.340} = 0.340$, which is the mean number of CODIS entries per SAK. While these two quantities are not inconsistent with our findings, the CODIS yield in ref. 10, which used the SAKs from the San Francisco Police Department Criminalistics Laboratory, where probative samples were flagged by SAFEs but all samples in a SAK and highlight the potential benefit of the machine-learning model. Other analyses suggest that the hit rate is higher for SAKs associated with stranger sexual assaults (and assaults that involve weapons) (3, 5), which could lead to less aggressive testing of samples in nonstranger SAKs under the nonlinear priority policy (which allows the number of samples tested per SAK to vary) but not under the priority policy (which tests the same number of samples from each SAK). We note that if there is a CODIS entry but no CODIS hit in a nonstranger SAK, the entry may still deter the offender from committing future offenses (25).

Conclusion

Within the context of sexual assaults, we address a fundamental issue in criminal investigations: how much evidence to collect and process. Using machine learning, optimization, and a dataset from the San Francisco Police Department Criminalistics Laboratory, where probative samples were flagged by SAFEs but all samples were tested, we show that standard machine-learning algorithms outperform SAFEs at identifying probative samples, that accounting for the economies of scale in DNA processing allows for a more cost-effective testing strategy, and that full testing of all DNA samples in a SAK has a slightly lower benefit-to-cost ratio than testing only the samples deemed most probative by the SAFEs, but more than doubles the CODIS yield. Moreover, our results suggest that the CODIS yield would increase another 47.2% by collecting samples from the three most probative locations (as deemed by the machine-learning algorithm). Taken together, our results support the testing of all samples in a SAK and highlight the potential benefit of the real-time use of machine learning and optimization algorithms during a sexual assault forensic medical examination; however, similar analyses in other municipalities are needed to assess the generalizability of our findings.

Acknowledgments This research was supported by the Graduate School of Business, Stanford University. We thank Rebecca Cotterman and Kelley Fracchia for manually coding the sexual assault kit questionnaires and Julie Valentine and Stefan Wager for helpful discussions.

1. N. P. Lovrich et al., National DNA Study Report, Final Report (US Department of Justice, Washington, DC, 2004).
2. K. J. Strom, M. J. Hickman, Unanalyzed evidence in law-enforcement agencies: A national examination of forensic processing in police departments. Criminal. Publ. Pol. 9, 381–404 (2010).
3. R. Campbell, S. J. Pierce, D. B. Sharma, H. Feeney, G. Fehler-Cabral, Should rape kit testing be prioritized by victim-offender relationship? Empirical comparison of forensic testing outcomes for stranger and nonstranger sexual assaults. Criminal. Publ. Pol. 15, 555–583 (2016).
4. M. Singer, R. Lovell, D. Flannery, Cost savings and cost effectiveness of the Cuyahoga County sexual assault kit task force (Begun Center for Violence Prevention Research and Education, Case Western Reserve University, Cleveland, OH, 2016). http://begun. case.edu/wp-content/uploads/2016/06/Cost-Savings-and-Cost-Effectiveness-Brief-1. Accessed 6 November 2017.

6 of 7  www.pnas.org/cgi/doi/10.1073/pnas.2001103117 Wang et al.
5. C. Wang, L. M. Wein, Analyzing approaches to the backlog of untested sexual assault kits in the U.S.A. J. Forensic Sci. 63, 1110–1121 (2018).

6. P. J. Speaker, The jurisdictional return on investment from processing the backlog of untested sexual assault kits. Forensic Sci. Int. Synergy 63, 1110–1121 (2018).

7. Bureau of Justice Assistance, US Department of Justice, Sexual assault kit initiative (SAKI). https://www.bja.gov/ProgramDetails.aspx?Program_ID=11. Accessed 29 July 2019.

8. US Congress, S.1766–SAFER Act of 2017. https://www.congress.gov/bill/115th-congress/senate-bill/1766. Accessed 29 July 2019.

9. Office of Manhattan District Attorney, C. R. Vance Jr., Test every kit: Results from Manhattan district attorney's office's sexual assault kit backlog elimination grant program (2019). https://www.manhattanda.org/wp-content/uploads/2019/03/Test-Every-Kit-Results-from-the-Manhattan-District-Attorneys-Offices-Sexual-Assault-Kit-Backlog-Elimination-Grant-Program.pdf. Accessed 9 December 2019.

10. J. Valentine, S. Miles, Utah Quick Kit (UQuik): A Collaborative program on the sexual assault kit analysis process in Proceedings, American Academy of Forensic Sciences 70th Annual Scientific Meeting (American Academy of Forensic Sciences, Colorado Springs, CO, 2018), Abstract E67, p. 530.

11. CA Department of Justice Jan Bashinski DNA Laboratory, California expands rapid DNA analysis system. Accessed 29 July 2019.

12. National Institute of Justice, Office of Justice Programs, US Department of Justice, National Best Practices for Sexual Assault Kits: A Multidisciplinary Approach (US Department of Justice, Washington, DC, 2017).

13. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction (Springer, New York, NY, ed. 2, 2009).

14. Project FORESIGHT, Project FORESIGHT Annual Report, 2017-2018. Forensic Science Initiative (College of Business & Economics, West Virginia University, Morgantown, WV, 2018).

15. Transparent California, https://transparentcalifornia.com/salaries/all. Accessed 31 July 2019.

16. D. Wojtczak, “On strong NP-completeness of rational problems” in Proceedings of the 13th International Computer Science Symposium in Russia, CSR, 2018, F. V. Fomin, V. V. Podolskii, Eds. (Springer International Publishing, New York, NY, 2018), pp. 308–320.

17. C. Manski, Interpreting point predictions: Some logical issues. Foundations Trends Accounting 10, 238–261 (2016).

18. F. Feng, H. Hong, K. Tang, J. Wang, Decision making with machine learning and ROC curves (2019). https://ssrn.com/abstract=3382962. Accessed 26 March 2020.

19. J. Cohen, P. Cohen, S. G. West, L. S. Aiken, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Routledge Press, Abingdon-on-Thames, UK, ed. 3, 2002).

20. R. M. Dawes, D. Faust, P. E. Meehl, Clinical versus actuarial judgment. Science 243, 1668–1674 (1989).

21. Office on Violence against Women, US Department of Justice, A National Protocol for Sexual Assault Medical Forensic Examinations: Adults/Adolescents (US Department of Justice, Washington, DC, ed. 2, 2013), NCJ 228119.

22. J. S. Mihalovich, E. Kingsbury, “Victim sexual assault evidence kits-The OPD crime lab and Alameda County district attorney's office teamwork” in The 129th Semi-Annual Seminar of the California Association of Criminalists (San Francisco, CA, 2017).

23. R. Lovell, M. Luminas, D. J. Flannery, R. Bell, B. Kyker, Describing the process and quantifying the outcomes of the Cuyahoga County sexual assault kit initiative. J. Crim. Justice 57, 106–115 (2018).

24. J. Valentine, S. Miles, L. Miles, L. Mabey, “Testing sexual assault kits supports the principle of 'justice for all'' in Proceedings, American Academy of Forensic Sciences 71st Annual Scientific Meeting (American Academy of Forensic Sciences, 2019), Abstract E74, p. 574.

25. J. L. Doleac, The effects of DNA databases on crime. Am. Econ. J. Appl. Econ. 9, 165–201 (2016).
Supporting Information

Tables S1-S3 and S6-S23 and Figs. S1-S14 are discussed in the main text. Two alternative machine learning models are described in §1, the optimization problem is formulated and analyzed in §2, and a numerical example is provided in §3.

1 Two Alternative Machine Learning Models

We describe the logistic regression (LR) model and the Classification and Regression Tree (CART) model in §1.1 and §1.2, respectively.

1.1 Logistic Regression

To restrict the number of covariates in the LR analysis, we use different sets of covariates relevant to each of the six locations: all six locations use the 14 covariates in Table S4 and additional covariates appearing in Table S5 are used for the first four locations.

Assuming there are $n_i$ tested samples from location $i$ in the training set, and given $n_i$ $d$-dimensional covariate vectors $x_j^{(i)}$ and their corresponding binary labels $y_j^{(i)}$ (which equals one if the tested sample from location $i$ is uploadable into CODIS, and equals zero if it is not), we calculate the maximum likelihood estimates for $i = 1, \ldots, 6$ via

$$
(\hat{\beta}_0^{(i)}, \hat{\beta}_{LR}^{(i)}) = \arg \max_{\beta_0^{(i)}, \beta_{LR}^{(i)}} - \sum_{j=1}^{n_i} \log \left[ h\left( y_j^{(i)} (\beta_0^{(i)} + \beta_{LR}^{(i)} x_j^{(i)}) \right) \right],
$$

where

$$
h(a) = \frac{1}{1 + e^{-a}}.
$$

We randomly divide the entire data set into a training set and a testing set, using the ratio 6:4. For each location $i$, we perform the LR on the training set, and then use the test set to measure performance. We first prioritize the samples by their probability of being CODIS-uploadable and test the top $n$ samples in each SAK for various values of $n$ (and testing all samples in the SAK if there are less than $n$ samples). Defining a SAK as being
“in CODIS” if it has at least one tested sample that is CODIS-uploadable, we then plot the number of SAKs in CODIS versus the number of samples tested, and use as our performance metric the normalized Area Under the Curve (AUC) of this plot. This procedure is repeated 100 times and the average result (and 95% confidence interval) is presented.

1.2 Classification and Regression Tree

The CART approach relies on recursive splitting of the data into smaller and smaller subgroups. In the implementation of CART, the dataset is split into the two subgroups that achieve the most gain in the objective function. This procedure is continued on each subgroup until some minimum subgroup size (which is also known as the minimum bucket size of tree leaf node) is reached.

We randomly divide the entire data set into a training set, a validation set and a testing set, using the ratio 5:1:4. We perform CART on the training set using different values of the minimum bucket size, and use the validation set to choosing the best bucket size, using the performance metric of the normalized AUC of the number of SAKs in CODIS versus the number of samples tested. This procedure is repeated 100 times and the average result is presented.

2 Optimization Problem

Let $q_j$ be the probability that a SAK is of type $j = 1, \ldots, J$, where a SAK type is characterized by its set of covariate values and its number of samples at each location, and $p_{ij}$ be the probability that a sample from location $i$ of a type $j$ SAK is CODIS-uploadable, as predicted by our machine learning model. We ignore any statistical imprecision in $q_j$ and $p_{ij}$, and plug our sample estimates into the following optimization problem. Let $n_{ij}$ be the number of samples from location $i$ of a type $j$ SAK, which are known values. Our decision variables are denoted by $x_{ij}$, which is the number of samples tested from location $i$ of a type
Let \( z_{ij} = 1 \) if \( x_{ij} \geq 1 \), and \( z_{ij} = 0 \) otherwise. We require that \( \sum_{i=1}^{6} x_{ij} z_{ij} \geq 1 \), so that at least one sample from each SAK is tested. Ignoring this requirement might lead to slightly better performance but would violate the spirit of testing the backlog. The problem of maximizing the probability that a SAK is CODIS-uploadable subject to a constraint on the mean cost per SAK is

\[
\max_{x_{ij}, z_{ij}} \sum_{j=1}^{J} q_j \left( 1 - \prod_{i=1}^{6} (1 - p_{ij})^{x_{ij}} \right) 
\]

subject to

\[
\sum_{j=1}^{J} q_i (F + v \sum_{i=1}^{6} x_{ij} z_{ij}) \leq B, 
\]

\[
\sum_{i=1}^{6} x_{ij} z_{ij} \geq 1, 
\]

\[
0 \leq x_{ij} \leq n_{ij},
\]

where \( x_{ij} \) is integer-valued and \( B \) is the allowable budget for the mean cost per SAK.

An optimal solution to (3)-(6) is likely to be complex \[1\], and we resort to a simple greedy algorithm. Because constraint (5) eliminates the importance of the fixed cost \( F \), a natural algorithm for this problem is a two-stage variation of the greedy algorithm for the unbounded linear knapsack problem \[2\]. In the first stage, for \( j = 1, \ldots, J \), set \( x_{i^*j} = 1 \) for \( i^* = \arg \max_i \{ p_{ij} | n_{ij} > 0 \} \), which guarantees that constraint (5) is satisfied, and let \( x_{ij} = 0 \) otherwise. Then the remaining problem reduces to a nonlinear bounded knapsack problem. Given the current \( x_{ij} \) values, the marginal cost from increasing \( x_{ij} \) to \( x_{ij} + 1 \) is \( v \) for all \( i \) and \( j \), and the marginal benefit is

\[
\left( 1 - \prod_{l=1}^{6} (1 - p_{lj})^{x_{lj}} \right) - \left( 1 - \prod_{l=1}^{6} (1 - p_{lj})^{x_{lj}} (1 - p_{ij}) \right) = p_{ij} \prod_{l=1}^{6} (1 - p_{lj})^{x_{lj}}.
\]

Hence, starting with the \( x_{ij} \) values from the first stage, the second stage of our algorithm repeatedly ranks the \((i, j)\) pairs that have \( x_{ij} < n_{ij} \) by \( p_{ij} \prod_{l=1}^{6} (1 - p_{lj})^{x_{lj}} \) and increases \( x_{ij} \) by 1 for \( \arg \max_{i,j} p_{ij} \prod_{l=1}^{6} (1 - p_{lj})^{x_{lj}} \) until the budget \( B \) is used up. Notice that this policy captures the decreasing returns toward the CODIS yield of testing samples for a SAK;
i.e., the quantity in (7) is decreasing in the $x_{ij}$ values. In the Results section, this greedy algorithm is referred to as the Nonlinear Priority Policy.

3 A Numerical Example for Computing the CODIS Yield

To illustrate how we compute the curves in Fig. 1 in the main text, we drop the subscript $j$ and consider a hypothetical SAK. The expected yield from the machine learning algorithm (based on the SAK’s covariate values) is $(p_1, \ldots, p_6) = (0.8, 0.6, 0.7, 0.1, 0.3, 0.9)$. The SAK contains ten samples, where $(n_1, \ldots, n_6) = (4, 3, 2, 0, 1, 0)$. Four of these ten samples generated CODIS-uploadable DNA profiles and the number of these in each location is $(2, 2, 0, 0, 0, 0)$; i.e., two of the four samples from location 1 and two of the three samples from location 2 generated profiles for CODIS.

Now we compute the CODIS yield for the SAK – i.e., the probability that at least one of the tested samples generates a CODIS-uploadable DNA profile – under the Priority Policy (described in the Results section of the main text) when $k$ samples are tested, for $k = 1, \ldots, 5$. When $k = 1$, we test a sample from location $\arg \max_i p_i | n_i > 0$, which is location 1. The yield is $2/4 = 0.5$. When $k = 2$, we test two samples without replacement from location 1, which has yield $5/6$ (i.e., the probability that both samples are not CODIS-uploadable is $1 - \left(\frac{2}{4}\right) \left(\frac{1}{3}\right)$). When $k = 3$, we test three samples without replacement from location 1, and the yield is 1. Similarly, when $k = 4$, we test four samples from location 1 and the yield is 1. When $k = 5$, we test four samples from location 1 and one sample from location 3 and the yield is 1.

Finally, we compute the CODIS yield for the SAK for $k = 1, 2, 3$ when we add synthetic samples (i.e., the top curve in Fig. 1) to the top three locations, which in this case are locations 6, 1 and 3. Because $n_6 = 0$, $n_1 > 0$ and $n_3 > 0$, we add one synthetic sample from location 6 to the SAK. When $k = 1$, we test the synthetic sample from location 6 and the yield is $p_6 = 0.9$. When $k = 2$, we test the synthetic sample from location 6 and a sample
from location 1, and the yield is $1 - (0.1)(0.5) = 0.95$. When $k = 3$, we test the synthetic sample from location 6 and two samples without replacement from location 1, and the yield is $1 - (0.1) \left( \frac{2}{3} \right) \left( \frac{1}{3} \right) = \frac{59}{60}$. 
References

[1] Wojtczak D. On strong NP-completeness of rational problems. Proceedings of the 13th International Computer Science Symposium in Russia, CSR 2018, Eds. Fomin FV, Podolskii VV. 2018:308-320.

[2] Dantzig GB. Discrete-variable extremum problems, Operations Research 1957;5:266-288.
| Time delay between assault and exam = 0 days | Time delay between assault and exam = 1 day |
|---------------------------------------------|---------------------------------------------|
| Victim age (yr)                             | Victim age (yr)                             |
| Victim gender M                             | Victim gender M                             |
| Loss of memory Y                            | Loss of memory Y                            |
| Number of offenders = 1                     | Number of offenders > 1                     |
| Consensual sex in prior 5 days U            | Consensual sex in prior 5 days U            |
| Consensual sex in prior 24 hours U          | Consensual sex in prior 24 hours U          |
| Known ejaculation U                         | Known ejaculation Y                         |
| Condom used U                               | Condom used Y                               |
| Shower or bath before exam U                | Shower or bath before exam Y                |
| Vaginal penetration of victim by offender U | Vaginal penetration of victim by offender Y |
| Anal penetration of victim by offender U    | Anal penetration of victim by offender Y    |
| Oral penetration of victim by offender U    | Oral penetration of victim by offender Y    |
| Offender’s mouth on genitals U              | Offender’s mouth on genitals Y              |
| Offender’s mouth on breasts U               | Offender’s mouth on breasts Y               |
| Offender’s mouth on other body parts U      | Offender’s mouth on other body parts Y      |
| Digital penetration of victim by offender U | Digital penetration of victim by offender Y |
| Oral penetration of offender by victim U    | Oral penetration of offender by victim U    |
| Strangled U                                 | Strangled Y                                 |
| Punched U                                   | Punched Y                                   |
| Stabbed U                                   | Stabbed Y                                   |
| Vaginal injury U                            | Vaginal injury Y                            |
| Other injury U                              | Other injury Y                              |

**Table S1.** List of covariates used in the training set, in the one-hot encoding format. M means Male, Y means Yes and U means Unknown.
| Keywords                                                                 | Location          |
|------------------------------------------------------------------------|-------------------|
| abdo, umbili, umbilicus, umbilical, stomach, abd-mons, belly, flank,   | Body surface      |
| side of body, Abdomen, Adbomen, adbomen, abdmoen, abdomen, finger,    |                   |
| palm, hand, nail, forearm, lower arm, wrist, bicep, arm, elbow,        |                   |
| antecubital, shoulder, scapula, collar bone, leg, foot, thigh, thing,  |                   |
| thigs, calf, knee, hair, head hair, ear, facial, face, eye, orbital,    |                   |
| neck, suction injury, nose, back of head, chin, neck, brest, brest,    |                   |
| reast, breasts, chest, nipple, apple, breast, low back, lower back,    |                   |
| upper back, back, WL, WL+, touch, ABD, R index, L index, pubic, corona,|                   |
| navel, hip, groin, buttock                                             |                   |
| vag, vagina, vagina, Vagina, caginal, vulva, vulvar, vulva, cervix,    | Genital           |
| cervix, cervix, cervix, cervical, cerival, cervical, peni, hymen,     |                   |
| scrotum, scrotal, semen, labia, glans, glands, vestibul, vestible,     |                   |
| mons, mon, mos, OS, clitoris, urethral, testicle, fourchette,          |                   |
| Fourchette, gentalia, genetalia, G-1, G-2, V1-4, IUD, genital, tampon  |                   |
| cheek, lip, buccal, mouth, saliva, pral, oal, periora, neum, ora, oral,| Oral              |
| nasal                                                                  |                   |
| recta, rectal, rectum, anus, aal, nanl, perinani, gluteal, anua, coccyx,| Anal              |
| sacrum, perineal, A-1, A-2, anal, perineum                             |                   |
| unerwear, underwaer, panty, Panties, Panites, panites, thong,         | Clothing          |
| underwear, sweatpant, glove, suit, shirt, cloth, clothing, bra, legging |                   |
| slide, scrapping, scratch, urine, blood, fiber, debris, blood card,    | Foreign material   |
| paper, container with fluid, forensic material, gauze, tissue,         |                   |
| catepillar, foreign, plastic thread, condom, pad, sanitary pad,       |                   |
| menstrual pad, menstrual cup, self collected, self collected pad,      |                   |
| sanitary napkin, diva cup, diaper                                      |                   |

**Table S2.** Aggregation of keywords in police reports into six locations.
| Location1         | Body Surface | Genital | Oral | Anal | Clothing | Foreign Material |
|------------------|--------------|---------|------|------|----------|-----------------|
| Body Surface     | 0.276,0.351  | 0.241   | 0.302| 0.218| 0.081    |                 |
| Genital          | 0.143,0.225  | 0.512   |      | 0.422|          | 0.284           |
| Oral             |              | 0.169   |      |      | -0.011   | 0.013           |
| Anal             |              |         |      |      |          | 0.329,0.459     |
| Clothing         |              |         |      |      |          | 0.614           |
| Foreign Material |              |         |      |      |          | 0.204,0.507     |

**Table S3.** Partial correlations, along with 95% confidence intervals, for each pair of locations, computed via equations (9)-(10) in the main text.
| Time delay between assault and exam (0 days/1 day/ ≥ 2 days) |
|-------------------------------------------------------------|
| Victim age (yrs)                                            |
| Victim gender at birth (M/F)                                |
| Loss of memory (Y/N/U)                                      |
| Number of offenders (1/> 1/U)                               |
| Consensual sex in prior 5 days (Y/N/U)                      |
| Consensual sex in prior 24 hours (Y/N/U)                    |
| Known ejaculation (Y/N/U)                                   |
| Condom used (Y/N/U)                                         |
| Shower or bath before exam (Y/N/U)                          |
| Strangled (Y/N/U)                                           |
| Punched (Y/N/U)                                             |
| Stabbed (Y/N/U)                                             |
| Other injury (Y/N/U)                                        |

**Table S4.** List of 14 covariates used for all six locations in the Logistic Regression analysis. M/F means that the possible values are Male or Female, and Y/N/U means that the possible values are Yes, No or Unknown.
| Location 1         | Location 2                     | Location 3                                | Location 4                                |
|-------------------|--------------------------------|-------------------------------------------|-------------------------------------------|
| Offender’s mouth on breasts | Vaginal injury               | Oral penetration of victim by offender   | Anal penetration of victim by offender   |
| Offender’s mouth on other body parts | Vaginal penetration of victim by offender | Oral penetration of offender by victim |                                |
|                   | Oral penetration of victim by offender |                                         |                                |
|                   | Offender’s mouth on genitals   |                                          |                                |
|                   | Digital penetration of victim by offender |                                         |                                |

**Table S5.** Additional covariates used for individual locations in the Logistic Regression analysis.
| Covariate                                           | Coefficient | 95% CI             |
|-----------------------------------------------------|-------------|-------------------|
| (Intercept)                                         | -2.406      | (-3.433,-1.379)   |
| Time delay between assault and exam 0 Days          | 0.808       | (0.418,1.198)     |
| Time delay between assault and exam 1 Day           | 0.657       | (0.338,0.976)     |
| Victim age (yr)                                     | -0.164      | (-0.294,-0.033)   |
| Victim gender M                                     | -0.289      | (-0.760,0.181)    |
| Loss of memory U                                   | -1.182      | (-4.126,1.763)    |
| Loss of memory Y                                   | 0.042       | (-0.331,0.415)    |
| Number of offenders One                            | 0.297       | (-0.082,0.677)    |
| Number of offenders U                              | 0.053       | (-0.365,0.471)    |
| Consensual sex in prior 5 days U                   | -0.35       | (-1.091,0.392)    |
| Consensual sex in prior 5 days Y                   | 0.244       | (-0.207,0.695)    |
| Consensual sex in prior 24 hours U                 | 0.053       | (-0.475,0.582)    |
| Consensual sex in prior 24 hours Y                 | -0.166      | (-0.678,0.346)    |
| Known ejaculation U                                | 0.282       | (-0.189,0.754)    |
| Known ejaculation Y                                | 0.52        | (0.022,1.019)     |
| Condom used U                                      | 0.132       | (-0.273,0.537)    |
| Condom used Y                                      | -0.172      | (-0.746,0.401)    |
| Shower or bath before exam U                       | -0.734      | (-1.338,-0.131)   |
| Shower or bath before exam Y                       | -0.772      | (-1.15,-0.393)    |
| Strangled U                                        | 0.249       | (-1.282,1.779)    |
| Strangled Y                                        | -0.141      | (-0.689,0.406)    |
| Punched U                                          | -0.21       | (-1.645,1.224)    |
| Punched Y                                          | -0.339      | (-0.758,0.079)    |
| Stabbed U                                          | -2.001      | (-12.221,8.219)   |
| Stabbed Y                                          | -2.946      | (-14.484,8.592)   |
| Other Injury U                                     | 2.043       | (-8.498,12.583)   |
| Other Injury Y                                     | -0.184      | (-0.431,0.063)    |
| Offender’s mouth on breast U                       | -0.507      | (-1.16,0.146)     |
| Offender’s mouth on breast Y                       | 0.315       | (-0.025,0.654)    |
| Offender’s mouth on other body parts U             | 0.767       | (0.168,1.367)     |
| Offender’s mouth on other body parts Y             | 0.768       | (0.258,1.278)     |

**Table S6.** Results for the logistic regression model for location 1 (body surface).
| Covariate                                                                 | Coefficient | 95% CI             |
|---------------------------------------------------------------------------|-------------|--------------------|
| (Intercept)                                                               | -3.613      | (-5.45,-1.776)     |
| Time delay between assault and exam 0 Days                                | 1.45        | (0.967,1.932)      |
| Time delay between assault and exam 1 Day                                 | 1.095       | (0.543,1.647)      |
| Victim age (yr)                                                           | 0.039       | (-0.134,0.212)     |
| Victim gender M                                                           | 1.586       | (0.492,2.68)       |
| Loss of memory U                                                          | -3.856      | (-13.61,5.898)     |
| Loss of memory Y                                                          | -0.273      | (-0.726,0.18)      |
| Number of offenders One                                                   | 0.161       | (-0.486,0.807)     |
| Number of offenders U                                                     | -0.481      | (-1.16,0.198)      |
| Consensual sex in prior 5 days U                                          | 0.354       | (-0.533,1.24)      |
| Consensual sex in prior 5 days Y                                          | 0.361       | (0.022,0.701)      |
| Consensual sex in prior 24 hours U                                       | -0.651      | (-1.075,-0.226)    |
| Consensual sex in prior 24 hours Y                                       | 0.095       | (-0.511,0.701)     |
| Known ejaculation U                                                       | 0.467       | (-0.026,0.961)     |
| Known ejaculation Y                                                       | 0.767       | (0.177,1.357)      |
| Condom used U                                                             | -0.149      | (-0.863,0.565)     |
| Condom used Y                                                             | -1.247      | (-1.995,-0.499)    |
| Shower or bath before exam U                                             | -0.066      | (-0.931,0.8)       |
| Shower or bath before exam Y                                             | -0.6        | (-1.108,-0.092)    |
| Strangled U                                                               | -0.075      | (-1.79,1.64)       |
| Strangled Y                                                               | -0.249      | (-0.779,0.282)     |
| Punched U                                                                 | 0.562       | (-1.37,2.494)      |
| Punched Y                                                                 | -0.55       | (-1.039,-0.061)    |
| Stabbed U                                                                 | -14.912     | (-20.668,-9.157)   |
| Stabbed Y                                                                 | -14.687     | (-16.043,-13.331)  |
| Other Injury U                                                            | 13.207      | (3.599,22.815)     |
| Other Injury Y                                                            | -0.129      | (-0.385,0.126)     |
| Vaginal Injury U                                                          | -0.032      | (-9.161,9.097)     |
| Vaginal Injury Y                                                          | 0.413       | (-0.076,0.903)     |
| Vaginal penetration of victim by offender U                              | 1.091       | (-0.187,2.368)     |
| Vaginal penetration of victim by offender Y                               | 1.392       | (0.243,2.541)      |
| Oral penetration of victim by offender U                                  | -3.161      | (-20.787,14.465)   |
| Oral penetration of victim by offender Y                                  | -1.567      | (-17.597,14.462)   |
| Offender's mouth on genitals U                                            | 2.918       | (-14.723,20.56)    |
| Offender's mouth on genitals Y                                            | 1.974       | (-14.104,18.053)   |
| Digital penetration of victim by offender U                               | -0.13       | (-0.64,0.379)      |
| Digital penetration of victim by offender Y                               | -0.261      | (-0.825,0.304)     |

**Table S7.** Results for the logistic regression model for location 2 (genital).
| Covariate                                           | Coefficient | 95% CI          |
|-----------------------------------------------------|-------------|-----------------|
| (Intercept)                                         | -3.301      | (-6.835, 0.233) |
| Time delay between assault and exam 0 Days          | 0.877       | (-0.075, 1.829) |
| Time delay between assault and exam 1 Day           | 0.939       | (0.083, 1.794)  |
| Victim age (yr)                                     | -0.356      | (-0.697, -0.016)|
| Victim gender M                                     | -0.365      | (-1.205, 0.475) |
| Loss of memory U                                    | -16.814     | (-17.997, -15.631) |
| Loss of memory Y                                    | -0.375      | (-1.294, 0.544) |
| Number of offenders One                             | 0.998       | (-2.466, 4.462) |
| Number of offenders U                               | 0.334       | (-6.464, 7.132) |
| Consensual sex in prior 5 days U                    | -7.17       | (-22.987, 8.647) |
| Consensual sex in prior 5 days Y                    | 0.493       | (-0.356, 1.343) |
| Consensual sex in prior 24 hours U                  | 0.243       | (-0.389, 0.875) |
| Consensual sex in prior 24 hours Y                  | 0.541       | (-0.583, 1.166) |
| Known ejaculation U                                 | -0.427      | (-1.158, 0.303) |
| Known ejaculation Y                                 | -0.797      | (-1.783, 0.19)  |
| Condom used U                                       | 0.254       | (-0.614, 1.122) |
| Condom used Y                                       | -0.935      | (-6.676, 4.805) |
| Shower or bath before exam U                        | -16.391     | (-17.28, -15.501) |
| Shower or bath before exam Y                        | -0.648      | (-1.364, 0.069) |
| Strangled U                                         | -5.876      | (-22.645, 10.893) |
| Strangled Y                                         | 0.154       | (-0.769, 1.077) |
| Punched U                                           | -0.329      | (-12.989, 12.331) |
| Punched Y                                           | -0.625      | (-6.441, 5.191) |
| Stabbed U                                           | 1.043       | (-25.339, 27.424) |
| Stabbed Y                                           | 2.742       | (-23.604, 29.087) |
| Other Injury U                                      | -7.211      | (-33.452, 19.03) |
| Other Injury Y                                      | -0.338      | (-0.899, 0.224) |
| Oral penetration of victim by offender U            | -0.379      | (-2.108, 1.349) |
| Oral penetration of victim by offender Y            | 0.135       | (-0.994, 1.264) |
| Oral penetration of offender by victim U            | -0.2        | (-2.098, 1.698) |
| Oral penetration of offender by victim Y            | -0.364      | (-1.149, 0.42)  |

**Table S8.** Results for the logistic regression model for location 3 (oral).
| Covariate                                                   | Coefficient | 95% CI          |
|-------------------------------------------------------------|-------------|-----------------|
| (Intercept)                                                 | -8.528      | (-23.102, 6.046) |
| Time delay between assault and exam 0 Days                  | 4.133       | (-4.327, 12.592) |
| Time delay between assault and exam 1 Day                   | 3.769       | (-4.643, 12.18)  |
| Victim age (yr)                                             | 0.019       | (-0.298, 0.335)  |
| Victim gender M                                             | -0.531      | (-1.345, 0.283)  |
| Loss of memory U                                            | -1.827      | (-12.221, 8.567) |
| Loss of memory Y                                            | -0.145      | (-0.964, 0.674)  |
| Number of offenders One                                     | 0.826       | (-0.139, 1.79)   |
| Number of offenders U                                       | -0.341      | (-3.837, 3.155)  |
| Consensual sex in prior 5 days U                            | -0.714      | (-5.408, 3.98)   |
| Consensual sex in prior 5 days Y                            | 0.536       | (-0.624, 1.696)  |
| Consensual sex in prior 24 hours U                          | -0.135      | (-1.088, 0.819)  |
| Consensual sex in prior 24 hours Y                          | 0.026       | (-1.227, 1.278)  |
| Known ejaculation U                                         | 2.881       | (-7.988, 13.751) |
| Known ejaculation Y                                         | 3.885       | (-7.122, 14.891) |
| Condom used U                                               | -0.137      | (-1.052, 0.777)  |
| Condom used Y                                               | -1.03       | (-2.159, 0.1)    |
| Shower or bath before exam U                                | 1.016       | (-5.274, 7.305)  |
| Shower or bath before exam Y                                | -0.712      | (-1.361, 0.063)  |
| Strangled U                                                 | -0.329      | (-6.423, 5.765)  |
| Strangled Y                                                 | -0.159      | (-1.391, 1.071)  |
| Punched U                                                   | 0.817       | (-5.407, 7.041)  |
| Punched Y                                                   | -0.412      | (-1.632, 0.808)  |
| Stabbed U                                                   | -14.692     | (-21.529, -7.854) |
| Stabbed Y                                                   | -15.055     | (-17.228, -12.882) |
| Other Injury U                                              | 7.357       | (-11.664, 26.378) |
| Other Injury Y                                              | -0.227      | (-0.843, 0.388)  |
| Anal penetration of victim by offender U                    | -0.34       | (-1.504, 0.823)  |
| Anal penetration of victim by offender Y                    | -0.109      | (-1.031, 0.813)  |

**Table S9.** Results for the logistic regression model for location 4 (anal).
| Covariate                                                                 | Coefficient | 95% CI              |
|---------------------------------------------------------------------------|-------------|---------------------|
| (Intercept)                                                               | -2.339      | (-6.235,1.558)      |
| Time delay between assault and exam 0 Days                                | 0.898       | (0.075,1.72)        |
| Time delay between assault and exam 1 Day                                 | 0.45        | (-0.468,1.367)      |
| Victim age (yr)                                                           | 0.069       | (-0.326,0.464)      |
| Victim gender M                                                           | -0.134      | (-0.812,0.545)      |
| Loss of memory U                                                          | 5.43        | (-9.865,20.725)     |
| Loss of memory Y                                                          | 0.11        | (-0.712,0.931)      |
| Number of offenders One                                                   | -0.248      | (-0.828,0.332)      |
| Number of offenders U                                                     | -0.634      | (-2.099,0.831)      |
| Consensual sex in prior 5 days U                                          | -2.605      | (-17.13,11.92)      |
| Consensual sex in prior 5 days Y                                          | 1.107       | (0.059,2.155)       |
| Consensual sex in prior 24 hours U                                        | 0.342       | (-0.793,1.478)      |
| Consensual sex in prior 24 hours Y                                        | -0.074      | (-0.958,0.811)      |
| Known ejaculation U                                                       | 0.604       | (-2.937,4.144)      |
| Known ejaculation Y                                                       | 2.265       | (-1.313,5.844)      |
| Condom used U                                                             | 0.378       | (-0.212,0.967)      |
| Condom used Y                                                             | -0.961      | (-1.875,-0.047)     |
| Shower or bath before exam U                                              | -6.787      | (-24.366,10.791)    |
| Shower or bath before exam Y                                              | -0.006      | (-0.687,0.675)      |
| Strangled U                                                               | -15.017     | (-26.083,-3.951)    |
| Strangled Y                                                               | -0.895      | (-1.87,0.081)       |
| Punched U                                                                 | -14.246     | (-23.562,-4.93)     |
| Punched Y                                                                 | -0.023      | (-1.178,1.132)      |
| Stabbed U                                                                 | 17.721      | (-2.96,38.401)      |
| Stabbed Y                                                                 | NA          | (NA,NA)             |
| Other Injury U                                                            | 11.588      | (-15.846,39.022)    |
| Other Injury Y                                                            | -0.27       | (-0.919,0.38)       |

**Table S10.** Results for the logistic regression model for location 5 (clothing).
| Covariate                                      | Coefficient | 95% CI            |
|------------------------------------------------|-------------|-------------------|
| (Intercept)                                    | -54.414     | (-226.003,117.175)|
| Time delay between assault and exam 0 Days     | 35.959      | (-66.516,138.434) |
| Time delay between assault and exam 1 Day      | -1.125      | (-127.575,125.324)|
| Victim age (yr)                                | 4.142       | (-47.686,55.969)  |
| Victim gender M                                | -11.816     | (-165.526,141.894)|
| Loss of memory U                               | 10.049      | (-269.693,289.791)|
| Loss of memory Y                               | -41.015     | (-132.006,49.975) |
| Number of offenders One                        | 22.851      | (-43.448,89.15)   |
| Number of offenders U                          | 34.948      | (-67.667,137.564) |
| Consensual sex in prior 5 days U               | -12.451     | (-189.892,164.989)|
| Consensual sex in prior 5 days Y               | -15.833     | (-101.135,69.469) |
| Consensual sex in prior 24 hours U             | 13.246      | (-62.102,88.594)  |
| Consensual sex in prior 24 hours Y             | 25.382      | (-119.065,169.828)|
| Known ejaculation U                            | 31.519      | (-139.583,202.621)|
| Known ejaculation Y                            | 37.411      | (-112.006,186.827)|
| Condom used U                                  | -17.647     | (-133.386,98.093) |
| Condom used Y                                  | -34.391     | (-158.556,89.775) |
| Shower or bath before exam U                   | 12.304      | (-219.786,244.394)|
| Shower or bath before exam Y                   | -27.224     | (-85.275,30.827)  |
| Strangled U                                    | 7.532       | (-145.439,160.503)|
| Strangled Y                                    | 0.304       | (-181.658,182.267)|
| Punched U                                      | -41.507     | (-214.041,131.028)|
| Punched Y                                      | -32.94      | (-173.175,107.295)|
| Stabbed U                                      | 16.624      | (-96.284,129.531) |
| Stabbed Y                                      | NA          | (NA,NA)           |
| Other Injury U                                 | -96.504     | (-535.861,342.852)|
| Other Injury Y                                 | 16.771      | (-51.985,85.528)  |

**Table S11.** Results for the logistic regression model for location 6 (foreign material).
| Covariate                                      | Coefficient | 95% CI            | No. of nonzeros |
|-----------------------------------------------|-------------|-------------------|-----------------|
| (Intercept)                                   | -2.254      | (-3.295,-1.213)   | 100             |
| Time delay between assault and exam 0 Days    | 0.778       | (0.321,1.237)     | 100             |
| Time delay between assault and exam 1 Day     | 0.567       | (0.149,0.984)     | 100             |
| Victim age (yr)                               | -0.111      | (-0.241,0.02)     | 91              |
| Victim gender M                               | -0.267      | (-0.804,0.271)    | 77              |
| Loss of memory U                              | -0.9        | (-2.083,0.283)    | 100             |
| Loss of memory Y                              | 0.163       | (-0.201,0.526)    | 73              |
| Number of offenders One                       | 0.304       | (-0.055,0.663)    | 94              |
| Number of offenders U                         | 0.129       | (-0.2,0.457)      | 44              |
| Consensual sex in prior 5 days U              | -0.157      | (-0.789,0.474)    | 67              |
| Consensual sex in prior 5 days Y              | 0.146       | (-0.232,0.524)    | 75              |
| Consensual sex in prior 24 hours U            | -0.012      | (-0.383,0.358)    | 66              |
| Consensual sex in prior 24 hours Y            | -0.127      | (-0.509,0.255)    | 63              |
| Known ejaculation U                           | 0.165       | (-0.233,0.564)    | 45              |
| Known ejaculation Y                           | 0.3         | (-0.133,0.734)    | 83              |
| Condom used U                                 | 0.215       | (-0.206,0.636)    | 67              |
| Condom used Y                                 | -0.244      | (-0.718,0.23)     | 73              |
| Shower or bath before exam U                  | -0.635      | (-1.26,-0.009)    | 99              |
| Shower or bath before exam Y                  | -0.742      | (-1.144,-0.34)    | 100             |
| Vaginal penetration of victim by offender U   | -0.055      | (-0.465,0.356)    | 42              |
| Vaginal penetration of victim by offender Y   | 0.245       | (-0.171,0.662)    | 90              |
| Anal penetration of victim by offender U      | -0.147      | (-0.606,0.313)    | 47              |
| Anal penetration of victim by offender Y      | -0.024      | (-0.395,0.348)    | 60              |
| Oral penetration of victim by offender U      | -0.836      | (-2.507,0.834)    | 47              |
| Oral penetration of victim by offender Y      | 0.22        | (-0.781.22)       | 49              |
| Offender’s mouth on genitals U                | 0.923       | (-0.914,2.76)     | 37              |
| Offender’s mouth on genitals Y                | -0.23       | (-1.456,0.997)    | 50              |
| Offender’s mouth on breast U                  | -0.359      | (-1.145,0.427)    | 45              |
| Offender's mouth on breast Y                  | 0.305       | (0.005,0.605)     | 98              |
| Offender's mouth on other body parts U        | 0.382       | (-0.263,1.027)    | 60              |
| Offender’s mouth on other body parts Y        | 0.468       | (-0.085,1.02)     | 95              |
| Digital penetration of victim by offender U   | 0.379       | (-0.177,0.936)    | 93              |
| Digital penetration of victim by offender Y   | 0.169       | (-0.213,0.552)    | 64              |
| Oral penetration of offender by victim U      | -0.146      | (-0.738,0.445)    | 48              |
| Oral penetration of offender by victim Y      | 0.09        | (-0.323,0.502)    | 64              |
| Strangled U                                   | 0.248       | (-0.588,1.084)    | 59              |
| Strangled Y                                   | -0.185      | (-0.637,0.267)    | 76              |
| Punched U                                     | -0.186      | (-1.08,0.708)     | 60              |
| Punched Y                                     | -0.341      | (-0.714,0.032)    | 95              |
| Stabbed U                                     | -0.246      | (-2.813,2.321)    | 49              |
| Stabbed Y                                     | -0.19       | (-2.871,2.491)    | 56              |
| Vaginal Injury U                              | 0.168       | (-2.228,2.563)    | 74              |
| Vaginal Injury Y                              | 0.237       | (-0.055,0.528)    | 90              |
| Other Injury U                                | 0.546       | (-2.021,3.114)    | 77              |
| Other Injury Y                                | -0.173      | (-0.402,0.057)    | 90              |

**Table S12.** Results for the LASSO logistic regression model for location 1 (body surface).

The last column gives the number of instances out of 100 runs that the covariate had a nonzero value.
| Covariate                                                                 | Coefficient | 95% CI              | No. of nonzeros |
|---------------------------------------------------------------------------|-------------|---------------------|-----------------|
| (Intercept)                                                               | -2.492      | (-3.965,-1.019)     | 100             |
| Time delay between assault and exam 0 Days                                | 1.158       | (0.545,1.771)       | 100             |
| Time delay between assault and exam 1 Day                                 | 0.734       | (0.066,1.402)       | 100             |
| Victim age (yr)                                                           | 0.05        | (-0.123,0.222)      | 71              |
| Victim gender M                                                           | 0.866       | (-0.293,2.025)      | 97              |
| Loss of memory U                                                          | -1.262      | (-2.959,0.436)      | 100             |
| Loss of memory Y                                                          | -0.186      | (-0.565,0.193)      | 73              |
| Number of offenders One                                                   | 0.165       | (-0.28,0.609)       | 76              |
| Number of offenders U                                                     | -0.372      | (-1.02,0.276)       | 78              |
| Consensual sex in prior 5 days U                                          | 0.338       | (-0.483,1.159)      | 68              |
| Consensual sex in prior 5 days Y                                          | 0.311       | (0.012,0.61)        | 99              |
| Consensual sex in prior 24 hours U                                        | -0.506      | (-0.95,-0.063)      | 99              |
| Consensual sex in prior 24 hours Y                                        | 0.197       | (-0.364,0.758)      | 72              |
| Known ejaculation U                                                       | 0.363       | (-0.119,0.844)      | 46              |
| Known ejaculation Y                                                       | 0.53        | (-0.111,1.172)      | 99              |
| Condom used U                                                             | -0.042      | (-0.553,0.469)      | 64              |
| Condom used Y                                                             | -0.967      | (-1.678,-0.257)     | 100             |
| Shower or bath before exam U                                              | -0.294      | (-0.893,0.304)      | 77              |
| Shower or bath before exam Y                                              | -0.537      | (-0.946,-0.127)     | 99              |
| Vaginal penetration of victim by offender U                               | 0.74        | (-0.572,2.053)      | 40              |
| Vaginal penetration of victim by offender Y                               | 0.522       | (-0.392,1.437)      | 98              |
| Anal penetration of victim by offender U                                  | -0.23       | (-0.61,0.149)       | 65              |
| Anal penetration of victim by offender Y                                  | -0.563      | (-1.179,0.054)      | 99              |
| Oral penetration of victim by offender U                                  | -0.16       | (-0.945,0.625)      | 29              |
| Oral penetration of victim by offender Y                                  | 0.242       | (-0.637,1.121)      | 62              |
| Offender’s mouth on genitals U                                            | 0.114       | (-0.81,1.038)       | 24              |
| Offender’s mouth on genitals Y                                            | 0.319       | (-0.799,1.436)      | 47              |
| Offender’s mouth on breast U                                              | 0.085       | (-0.506,0.676)      | 34              |
| Offender’s mouth on breast Y                                              | -0.07       | (-0.491,0.351)      | 50              |
| Offender’s mouth on other body parts U                                    | -0.313      | (-0.968,0.343)      | 57              |
| Offender’s mouth on other body parts Y                                    | 0.28        | (-0.129,0.69)       | 90              |
| Digital penetration of victim by offender U                               | -0.035      | (-0.41,0.34)        | 49              |
| Digital penetration of victim by offender Y                               | -0.263      | (-0.672,0.146)      | 80              |
| Oral penetration of offender by victim U                                  | -0.246      | (-0.717,0.225)      | 62              |
| Oral penetration of offender by victim Y                                  | -0.078      | (-0.509,0.352)      | 62              |
| Strangled U                                                                | 0.087       | (-0.921,1.095)      | 39              |
| Strangled Y                                                                | -0.127      | (-0.67,0.415)       | 59              |
| Punched U                                                                 | 0.339       | (-0.625,1.302)      | 56              |
| Punched Y                                                                 | -0.347      | (-0.796,0.102)      | 79              |
| Stabbed U                                                                 | -2.025      | (-5.529,1.479)      | 29              |
| Stabbed Y                                                                 | -1.279      | (-3.727,1.17)       | 58              |
| Vaginal Injury U                                                          | 0.052       | (-3.07,3.174)       | 47              |
| Vaginal Injury Y                                                          | 0.367       | (-0.141,0.876)      | 93              |
| Other Injury U                                                            | 1.369       | (-1.871,4.61)       | 82              |
| Other Injury Y                                                            | -0.066      | (-0.307,0.174)      | 44              |

**Table S13.** Results for the LASSO logistic regression model for location 2 (genital). The last column gives the number of instances out of 100 runs that the covariate had a nonzero value.
| Covariate                                        | Coefficient | 95% CI          | No. of nonzeros |
|------------------------------------------------|-------------|-----------------|-----------------|
| (Intercept)                                     | -3.273      | (-4.647,-1.898) | 100             |
| Time delay between assault and exam 0 Days      | 0.479       | (-0.283,1.242)  | 81              |
| Time delay between assault and exam 1 Day       | 0.541       | (-0.248,1.33)   | 90              |
| Victim age (yr)                                 | -0.176      | (-0.593,0.24)   | 81              |
| Victim gender M                                 | -0.39       | (-1.195,0.415)  | 61              |
| Loss of memory U                               | -1.686      | (-4.506,1.134)  | 100             |
| Loss of memory Y                               | -0.293      | (-0.987,0.402)  | 70              |
| Number of offenders One                        | 0.447       | (-0.251,1.145)  | 97              |
| Number of offenders U                          | 0.463       | (-0.833,1.758)  | 31              |
| Consensual sex in prior 5 days U               | -0.892      | (-3.058,1.275)  | 50              |
| Consensual sex in prior 5 days Y               | 0.475       | (-0.077,1.027)  | 84              |
| Consensual sex in prior 24 hours U             | 0           | (-0.514,0.513)  | 45              |
| Consensual sex in prior 24 hours Y             | 0.854       | (0.011,1.698)   | 41              |
| Known ejaculation U                            | -0.422      | (-1.287,0.433)  | 48              |
| Known ejaculation Y                            | -0.563      | (-1.493,0.367)  | 71              |
| Condom used U                                  | 0.204       | (-0.378,0.785)  | 43              |
| Condom used Y                                  | -0.46       | (-2.036,1.116)  | 66              |
| Shower or bath before exam U                   | -1.805      | (-4.214,0.603)  | 100             |
| Shower or bath before exam Y                   | -0.579      | (-1.199,0.042)  | 98              |
| Vaginal penetration of victim by offender U    | -0.519      | (-1.424,0.385)  | 70              |
| Vaginal penetration of victim by offender Y    | 0.084       | (-0.661,0.828)  | 75              |
| Anal penetration of victim by offender U       | 0.517       | (-0.521,1.555)  | 63              |
| Anal penetration of victim by offender Y       | 0.129       | (-0.631,0.89)   | 51              |
| Oral penetration of victim by offender U       | -0.613      | (-2.24,1.014)   | 50              |
| Oral penetration of victim by offender Y       | -0.094      | (-1.1,0.911)    | 48              |
| Offender’s mouth on genitals U                 | 0.713       | (-1.264,2.691)  | 20              |
| Offender’s mouth on genitals Y                 | -0.285      | (-1.634,1.064)  | 40              |
| Offender’s mouth on breast U                   | 0.075       | (-1.1,1.251)    | 36              |
| Offender’s mouth on breast Y                   | 0.261       | (-0.485,1.008)  | 66              |
| Offender’s mouth on other body parts U         | 0.935       | (-0.617,2.488)  | 35              |
| Offender’s mouth on other body parts Y         | 0.752       | (-0.21,1.714)   | 99              |
| Digital penetration of victim by offender U    | 0.432       | (-0.803,1.667)  | 48              |
| Digital penetration of victim by offender Y    | 0.479       | (-0.137,1.096)  | 98              |
| Oral penetration of offender by victim U       | -0.62       | (-1.948,0.708)  | 35              |
| Oral penetration of offender by victim Y       | -0.189      | (-0.69,0.312)   | 67              |
| Strangled U                                    | -0.884      | (-3.212,1.445)  | 43              |
| Strangled Y                                    | 0.096       | (-0.548,0.74)   | 46              |
| Punched U                                      | 0.407       | (-1.595,2.408)  | 50              |
| Punched Y                                      | -0.125      | (-0.941,0.691)  | 56              |
| Stabbed U                                      | 0.706       | (-2.022,3.435)  | 13              |
| Stabbed Y                                      | 2.376       | (-2.267,7.74)   | 49              |
| Vaginal Injury U                               | -0.116      | (-3.597,3.365)  | 42              |
| Vaginal Injury Y                               | -0.205      | (-0.633,0.224)  | 59              |
| Other Injury U                                 | -1.775      | (-5.37,1.821)   | 40              |
| Other Injury Y                                 | -0.384      | (-0.89,0.123)   | 86              |

**Table S14.** Results for the LASSO logistic regression model for location 3 (oral). The last column gives the number of instances out of 100 runs that the covariate had a nonzero value.
| Covariate                                                                 | Coefficient | 95% CI        | No. of nonzeros |
|---------------------------------------------------------------------------|-------------|---------------|-----------------|
| (Intercept)                                                               | -4.286      | (-7.296, -1.276) | 100             |
| Time delay between assault and exam 0 Days                                 | 2.094       | (0.179, 4.011) | 100             |
| Time delay between assault and exam 1 Day                                  | 1.737       | (-0.046, 3.507) | 100             |
| Victim age (yr)                                                           | 0.069       | (-0.247, 0.385) | 66              |
| Victim gender M                                                           | -0.549      | (-1.879, 0.78) | 80              |
| Loss of memory U                                                          | -0.406      | (-3.289, 2.351) | 63              |
| Loss of memory Y                                                          | -0.011      | (-0.771, 0.748) | 57              |
| Number of offenders One                                                   | 0.692       | (-0.223, 1.607) | 97              |
| Number of offenders U                                                     | 0.111       | (-1.265, 1.487) | 58              |
| Consensual sex in prior 5 days U                                          | 0.053       | (-1.143, 1.248) | 51              |
| Consensual sex in prior 5 days Y                                          | 0.395       | (-0.736, 1.527) | 88              |
| Consensual sex in prior 24 hours U                                        | -0.187      | (-1.167, 0.793) | 69              |
| Consensual sex in prior 24 hours Y                                        | 0.297       | (-1.034, 1.628) | 61              |
| Known ejaculation U                                                       | 0.889       | (-0.964, 2.741) | 45              |
| Known ejaculation Y                                                       | 1.241       | (-0.524, 3.006) | 99              |
| Condom used U                                                             | -0.096      | (-0.894, 0.703) | 54              |
| Condom used Y                                                             | -0.713      | (-1.699, 0.272) | 86              |
| Shower or bath before exam U                                              | 0.448       | (-1.712, 2.608) | 71              |
| Shower or bath before exam Y                                              | -0.559      | (-1.143, 0.024) | 100             |
| Vaginal penetration of victim by offender U                               | 0.367       | (-0.855, 1.587) | 61              |
| Vaginal penetration of victim by offender Y                               | -0.082      | (-1.261, 0.907) | 60              |
| Anal penetration of victim by offender U                                  | 0.041       | (-1.419, 1.501) | 59              |
| Anal penetration of victim by offender Y                                  | -0.117      | (-1.047, 0.813) | 61              |
| Oral penetration of victim by offender U                                 | 0.649       | (-1.469, 2.768) | 35              |
| Oral penetration of victim by offender Y                                  | -0.336      | (-1.176, 0.503) | 48              |
| Offender’s mouth on genitals U                                            | 0.664       | (-0.585, 1.914) | 50              |
| Offender’s mouth on genitals Y                                            | -0.145      | (-1.182, 0.892) | 42              |
| Offender’s mouth on breast U                                             | -0.771      | (-2.479, 0.937) | 68              |
| Offender’s mouth on breast Y                                              | -0.638      | (-1.679, 0.403) | 84              |
| Offender’s mouth on other body parts U                                    | -0.379      | (-1.511, 0.752) | 56              |
| Offender’s mouth on other body parts Y                                    | 0.617       | (-0.429, 1.664) | 83              |
| Digital penetration of victim by offender U                               | 0.319       | (-0.663, 1.302) | 72              |
| Digital penetration of victim by offender Y                               | -0.295      | (-1.253, 0.664) | 75              |
| Oral penetration of offender by victim U                                  | -0.901      | (-2.382, 0.58) | 69              |
| Oral penetration of offender by victim Y                                  | 0.189       | (-0.695, 1.072) | 72              |
| Strangled U                                                               | 0.298       | (-0.787, 1.383) | 46              |
| Strangled Y                                                               | -0.192      | (-1.444, 1.06) | 72              |
| Punched U                                                                 | 0.320       | (-1.096, 1.734) | 55              |
| Punched Y                                                                 | -0.380      | (-1.592, 0.833) | 76              |
| Stabbed U                                                                 | -1.621      | (-5.123, 1.88) | 40              |
| Stabbed Y                                                                 | -1.784      | (-4.784, 1.217) | 41              |
| Vaginal Injury U                                                          | -0.744      | (-4.298, 2.811) | 38              |
| Vaginal Injury Y                                                          | 0.132       | (-0.721, 0.985) | 70              |
| Other Injury U                                                            | 0.493       | (-3.308, 4.293) | 45              |
| Other Injury Y                                                            | -0.240      | (-0.866, 0.386) | 75              |

**Table S15.** Results for the LASSO logistic regression model for location 4 (anal). The last column gives the number of instances out of 100 runs that the covariate had a nonzero value.
| Covariate                                             | Coefficient | 95% CI            | No. of nonzeros |
|-------------------------------------------------------|-------------|-------------------|-----------------|
| (Intercept)                                           | -2.898      | (-5.811,0.016)    | 100             |
| Time delay between assault and exam 0 Days            | 0.897       | (-0.107,1.9)      | 99              |
| Time delay between assault and exam 1 Day             | 0.411       | (-0.408,1.23)     | 80              |
| Victim age (yr)                                       | 0.103       | (-0.264,0.469)    | 85              |
| Victim gender M                                       | 0.077       | (-1.25,1.404)     | 82              |
| Loss of memory U                                      | 1.775       | (-0.631,4.181)    | 99              |
| Loss of memory Y                                      | 0.234       | (-0.614,1.082)    | 80              |
| Number of offenders One                               | -0.134      | (-0.84,0.572)     | 74              |
| Number of offenders U                                 | -0.337      | (-1.555,0.882)    | 86              |
| Consensual sex in prior 5 days U                      | -0.768      | (-3.159,1.624)    | 88              |
| Consensual sex in prior 5 days Y                      | 1.09        | (-0.025,2.206)    | 100             |
| Consensual sex in prior 24 hours U                    | -0.065      | (-1.136,1.007)    | 87              |
| Consensual sex in prior 24 hours Y                    | -0.467      | (-1.587,0.653)    | 82              |
| Known ejaculation U                                   | 0.145       | (-1.039,1.328)    | 66              |
| Known ejaculation Y                                   | 1.871       | (-0.096,3.839)    | 100             |
| Condom used U                                         | 0.663       | (-0.512,1.837)    | 94              |
| Condom used Y                                         | -0.979      | (-2.44,0.482)     | 94              |
| Shower or bath before exam U                          | -1.371      | (-4.249,1.507)    | 87              |
| Shower or bath before exam Y                          | 0.066       | (-0.72,0.852)     | 84              |
| Vaginal penetration of victim by offender U           | 0.337       | (-1.186,1.861)    | 69              |
| Vaginal penetration of victim by offender Y           | 1.023       | (-0.283,2.33)     | 99              |
| Anal penetration of victim by offender U              | -0.146      | (-1.144,0.853)    | 71              |
| Anal penetration of victim by offender Y              | 0.422       | (-0.649,1.493)    | 86              |
| Oral penetration of victim by offender U              | -2.018      | (-6.763,2.728)    | 58              |
| Oral penetration of victim by offender Y              | -0.115      | (-2.951,2.721)    | 67              |
| Offender’s mouth on genitals U                        | 1.952       | (-2.714,6.617)    | 50              |
| Offender’s mouth on genitals Y                        | 1.169       | (-2.017,4.356)    | 60              |
| Offender’s mouth on breast U                          | 0.575       | (-1.761,2.91)     | 72              |
| Offender’s mouth on breast Y                          | -0.79       | (-1.988,0.409)    | 89              |
| Offender’s mouth on other body parts U                | 0.167       | (-1.288,1.622)    | 66              |
| Offender’s mouth on other body parts Y                | -0.67       | (-1.903,0.563)    | 90              |
| Digital penetration of victim by offender U           | 1.02        | (-0.537,2.577)    | 88              |
| Digital penetration of victim by offender Y           | 0.622       | (-0.441,1.683)    | 84              |
| Oral penetration of offender by victim U              | -0.866      | (-2.936,1.203)    | 75              |
| Oral penetration of offender by victim Y              | 1.223       | (-0.287,2.733)    | 95              |
| Strangled U                                           | -3.933      | (-9.251,1.384)    | 88              |
| Strangled Y                                           | -1.493      | (-3.215,0.229)    | 97              |
| Punched U                                             | -1.9        | (-4.09,0.29)      | 88              |
| Punched Y                                             | 0.023       | (-1.176,1.222)    | 75              |
| Stabbed U                                             | 3.997       | (3.741,4.253)     | 2               |
| Stabbed Y                                             | NA          | (NA,NA)           | NA              |
| Vaginal Injury U                                      | 3.891       | (-1.7,9.482)      | 17              |
| Vaginal Injury Y                                      | -0.223      | (-1.2,0.755)      | 65              |
| Other Injury U                                        | 2.202       | (-4.95,9.354)     | 23              |
| Other Injury Y                                        | -0.227      | (-1.072,0.617)    | 64              |

**Table S16.** Results for the LASSO logistic regression model for location 5 (clothing). The last column gives the number of instances out of 100 runs that the covariate had a nonzero value.
| Covariate                                      | Coefficient | 95% CI          | No. of nonzeros |
|------------------------------------------------|-------------|-----------------|-----------------|
| (Intercept)                                    | -11.36      | (-22.335,-0.385)| 100             |
| Time delay between assault and exam 0 Days     | 5.562       | (-0.968,12.092) | 91              |
| Time delay between assault and exam 1 Day      | -6.884      | (-18.643,4.875) | 4               |
| Victim age (yr)                                | 1.76        | (-2.465,5.98)   | 49              |
| Victim gender M                                | -2.229      | (-4.863,0.405)  | 17              |
| Loss of memory U                               | -0.698      | (NA,NA)         | NA              |
| Loss of memory Y                               | -2.384      | (-6.471,1.702)  | 36              |
| Number of offenders One                        | 3.244       | (-1.694,8.181)  | 72              |
| Number of offenders U                          | -0.586      | (-1.605,0.433)  | 4               |
| Consensual sex in prior 5 days U               | -2.278      | (-8.928,4.372)  | 14              |
| Consensual sex in prior 5 days Y               | -2.367      | (-7.291,2.557)  | 46              |
| Consensual sex in prior 24 hours U             | 0.242       | (-4.768,5.252)  | 14              |
| Consensual sex in prior 24 hours Y             | 3.117       | (-2.502,8.736)  | 28              |
| Known ejaculation U                            | 1.4         | (-0.833,3.633)  | 5               |
| Known ejaculation Y                            | 0.815       | (-4.591,6.22)   | 19              |
| Condom used U                                  | -0.643      | (-1.690,0.403)  | 10              |
| Condom used Y                                  | -2.532      | (-6.284,1.221)  | 31              |
| Shower or bath before exam U                   | -2.111      | (-6.242,2.018)  | 19              |
| Shower or bath before exam Y                   | -2.874      | (-8.032,2.283)  | 44              |
| Vaginal penetration of victim by offender U    | 4.178       | (-0.558,8.913)  | 63              |
| Vaginal penetration of victim by offender Y    | 0.038       | (NA,NA)         | NA              |
| Anal penetration of victim by offender U       | 2.568       | (-3.027,8.163)  | 2               |
| Anal penetration of victim by offender Y       | 2.505       | (-2.225,7.236)  | 54              |
| Oral penetration of victim by offender U       | 0.72        | (-0.599,2.04)   | 8               |
| Oral penetration of victim by offender Y       | 1.269       | (-5.264,7.802)  | 20              |
| Offender’s mouth on genitals U                 | 0.191       | (-0.393,0.776)  | 7               |
| Offender’s mouth on genitals Y                 | 0.065       | (-0.221,0.352)  | 5               |
| Offender’s mouth on breast U                   | 1.998       | (-2.053,6.048)  | 22              |
| Offender’s mouth on breast Y                   | 4.613       | (-2.799,12.025) | 37              |
| Offender’s mouth on other body parts U         | 0.81        | (-1.399,3.019)  | 17              |
| Offender’s mouth on other body parts Y         | 3.581       | (-2.058,9.22)   | 44              |
| Digital penetration of victim by offender U    | 1.08        | (-0.948,3.108)  | 8               |
| Digital penetration of victim by offender Y    | 1.937       | (-4.378,8.243)  | 34              |
| Oral penetration of offender by victim U       | 3.181       | (-8.142,14.504) | 6               |
| Oral penetration of offender by victim Y       | 2.257       | (-0.971,5.485)  | 34              |
| Strangled U                                    | -0.842      | (-2.446,0.762)  | 8               |
| Strangled Y                                    | 2.291       | (-5.378,9.96)   | 32              |
| Punched U                                      | -3.042      | (-7.391,1.307)  | 6               |
| Punched Y                                      | -0.133      | (-6.383,6.118)  | 32              |
| Stabbed U                                      | NA          | (NA,NA)         | NA              |
| Stabbed Y                                      | NA          | (NA,NA)         | NA              |
| Vaginal Injury U                               | -0.648      | (NA,NA)         | NA              |
| Vaginal Injury Y                               | -2.071      | (-7.463,3.317)  | 28              |
| Other Injury U                                 | -1.44       | (-4.845,1.966)  | 5               |
| Other Injury Y                                 | 1.152       | (-1.869,4.174)  | 20              |

**Table S17.** Results for the LASSO logistic regression model for location 6 (foreign material). The last column gives the number of instances out of 100 runs that the covariate had a nonzero value.
| Covariate                                                                 | Count |
|---------------------------------------------------------------------------|-------|
| Time delay between assault and exam = 0 days                              | 89    |
| Time delay between assault and exam = 1 day                               | 90    |
| Victim age (yr)                                                           | 100   |
| Victim gender M                                                           | 22    |
| Loss of memory U                                                          | 0     |
| Loss of memory Y                                                          | 26    |
| Number of offenders = 1                                                   | 69    |
| Number of offenders = U                                                   | 13    |
| Consensual sex in prior 5 days U                                          | 3     |
| Consensual sex in prior 5 days Y                                          | 81    |
| Consensual sex in prior 24 hours U                                        | 59    |
| Consensual sex in prior 24 hours Y                                        | 30    |
| Known ejaculation U                                                       | 45    |
| Known ejaculation Y                                                       | 17    |
| Condom used U                                                             | 41    |
| Condom used Y                                                             | 2     |
| Shower or bath before exam U                                              | 11    |
| Shower or bath before exam Y                                              | 97    |
| Vaginal penetration of victim by offender U                               | 39    |
| Vaginal penetration of victim by offender Y                               | 65    |
| Anal penetration of victim by offender U                                  | 16    |
| Anal penetration of victim by offender Y                                  | 10    |
| Oral penetration of victim by offender U                                  | 17    |
| Oral penetration of victim by offender Y                                  | 14    |
| Offender’s mouth on genitals U                                            | 6     |
| Offender’s mouth on genitals Y                                            | 0     |
| Offender’s mouth on breast U                                              | 23    |
| Offender’s mouth on breast Y                                              | 44    |
| Offender’s mouth on other body parts U                                    | 20    |
| Offender’s mouth on other body parts Y                                    | 65    |
| Digital penetration of victim by offender U                               | 30    |
| Digital penetration of victim by offender Y                               | 31    |
| Oral penetration of offender by victim U                                  | 28    |
| Oral penetration of offender by victim Y                                  | 17    |
| Strangled U                                                               | 10    |
| Strangled Y                                                               | 9     |
| Punched U                                                                 | 16    |
| Punched Y                                                                 | 10    |
| Stabbed U                                                                 | 0     |
| Stabbed Y                                                                 | 0     |
| Vaginal Injury U                                                          | 0     |
| Vaginal Injury Y                                                          | 55    |
| Other Injury U                                                            | 0     |
| Other Injury Y                                                            | 78    |

**Table S18.** Results for the CART model for location 1 (body surface). The last column gives the number of instances out of 100 runs that the covariate is used in the splitting of tree branches.
| Covariate                                           | Count |
|----------------------------------------------------|-------|
| Time delay between assault and exam = 0 days       | 98    |
| Time delay between assault and exam = 1 day        | 84    |
| Victim age (yr)                                    | 100   |
| Victim gender M                                    | 32    |
| Loss of memory U                                   | 0     |
| Loss of memory Y                                   | 32    |
| Number of offenders = 1                            | 26    |
| Number of offenders = U                            | 16    |
| Consensual sex in prior 5 days U                   | 0     |
| Consensual sex in prior 5 days Y                   | 58    |
| Consensual sex in prior 24 hours U                 | 77    |
| Consensual sex in prior 24 hours Y                 | 35    |
| Known ejaculation U                                | 10    |
| Known ejaculation Y                                | 28    |
| Condom used U                                      | 27    |
| Condom used Y                                      | 9     |
| Shower or bath before exam U                       | 4     |
| Shower or bath before exam Y                       | 78    |
| Vaginal penetration of victim by offender U        | 15    |
| Vaginal penetration of victim by offender Y        | 38    |
| Anal penetration of victim by offender U           | 21    |
| Anal penetration of victim by offender Y           | 13    |
| Oral penetration of victim by offender U           | 16    |
| Oral penetration of victim by offender Y           | 12    |
| Offender’s mouth on genitals U                     | 3     |
| Offender’s mouth on genitals Y                     | 4     |
| Offender’s mouth on breast U                       | 25    |
| Offender’s mouth on breast Y                       | 12    |
| Offender’s mouth on other body parts U             | 18    |
| Offender’s mouth on other body parts Y             | 31    |
| Digital penetration of victim by offender U        | 23    |
| Digital penetration of victim by offender Y        | 26    |
| Oral penetration of offender by victim U           | 21    |
| Oral penetration of offender by victim Y           | 28    |
| Strangled U                                        | 0     |
| Strangled Y                                        | 3     |
| Punched U                                          | 4     |
| Punched Y                                          | 4     |
| Stabbed U                                          | 0     |
| Stabbed Y                                          | 0     |
| Vaginal Injury U                                   | 0     |
| Vaginal Injury Y                                   | 48    |
| Other Injury U                                     | 0     |
| Other Injury Y                                     | 61    |

**Table S19.** Results for the CART model for location 2 (genital). The last column gives the number of instances out of 100 runs that the covariate is used in the splitting of tree branches.
| Covariate                                         | Count |
|--------------------------------------------------|-------|
| Time delay between assault and exam = 0 days     | 54    |
| Time delay between assault and exam = 1 day      | 59    |
| Victim age (yr)                                  | 100   |
| Victim gender M                                  | 18    |
| Loss of memory U                                 | 0     |
| Loss of memory Y                                 | 30    |
| Number of offenders One                          | 30    |
| Number of offenders = U                          | 27    |
| Consensual sex in prior 5 days U                 | 0     |
| Consensual sex in prior 5 days Y                 | 46    |
| Consensual sex in prior 24 hours U               | 17    |
| Consensual sex in prior 24 hours Y               | 36    |
| Known ejaculation U                              | 6     |
| Known ejaculation Y                              | 19    |
| Condom used U                                    | 23    |
| Condom used Y                                    | 8     |
| Shower or bath before exam U                     | 8     |
| Shower or bath before exam Y                     | 67    |
| Vaginal penetration of victim by offender U      | 14    |
| Vaginal penetration of victim by offender Y      | 36    |
| Anal penetration of victim by offender U         | 5     |
| Anal penetration of victim by offender Y         | 6     |
| Oral penetration of victim by offender U         | 8     |
| Oral penetration of victim by offender Y         | 6     |
| Offender’s mouth on genitals U                   | 0     |
| Offender’s mouth on genitals Y                   | 6     |
| Offender’s mouth on breast U                     | 10    |
| Offender’s mouth on breast Y                     | 15    |
| Offender’s mouth on other body parts U           | 16    |
| Offender’s mouth on other body parts Y           | 58    |
| Digital penetration of victim by offender U      | 2     |
| Digital penetration of victim by offender Y      | 48    |
| Oral penetration of offender by victim U         | 10    |
| Oral penetration of offender by victim Y         | 25    |
| Strangled U                                      | 3     |
| Strangled Y                                      | 9     |
| Punched U                                        | 0     |
| Punched Y                                        | 6     |
| Stabbed U                                        | 0     |
| Stabbed Y                                        | 0     |
| Vaginal Injury U                                 | 0     |
| Vaginal Injury Y                                 | 27    |
| Other Injury U                                   | 0     |
| Other Injury Y                                   | 22    |

**Table S20.** Results for the CART model for location 3 (oral). The last column gives the number of instances out of 100 runs that the covariate is used in the splitting of tree branches.
| Covariate | Count |
|-----------|-------|
| Time delay between assault and exam = 0 days | 72 |
| Time delay between assault and exam = 1 day | 57 |
| Victim age (yr) | 89 |
| Victim gender M | 20 |
| Loss of memory U | 0 |
| Loss of memory Y | 13 |
| Number of offenders = 1 | 24 |
| Number of offenders = U | 12 |
| Consensual sex in prior 5 days U | 2 |
| Consensual sex in prior 5 days Y | 49 |
| Consensual sex in prior 24 hours U | 25 |
| Consensual sex in prior 24 hours Y | 2 |
| Known ejaculation U | 11 |
| Known ejaculation Y | 58 |
| Condom used U | 11 |
| Condom used Y | 3 |
| Shower or bath before exam U | 14 |
| Shower or bath before exam Y | 39 |
| Vaginal penetration of victim by offender U | 21 |
| Vaginal penetration of victim by offender Y | 56 |
| Anal penetration of victim by offender U | 4 |
| Anal penetration of victim by offender Y | 12 |
| Oral penetration of victim by offender U | 4 |
| Oral penetration of victim by offender Y | 3 |
| Offender’s mouth on genitals U | 0 |
| Offender’s mouth on genitals Y | 4 |
| Offender’s mouth on breast U | 4 |
| Offender’s mouth on breast Y | 0 |
| Offender’s mouth on other body parts U | 15 |
| Offender’s mouth on other body parts Y | 27 |
| Digital penetration of victim by offender U | 6 |
| Digital penetration of victim by offender Y | 5 |
| Oral penetration of offender by victim U | 10 |
| Oral penetration of offender by victim Y | 9 |
| Strangled U | 0 |
| Strangled Y | 0 |
| Punched U | 0 |
| Punched Y | 3 |
| Stabbed U | 0 |
| Stabbed Y | 0 |
| Vaginal Injury U | 0 |
| Vaginal Injury Y | 22 |
| Other Injury U | 0 |
| Other Injury Y | 32 |

**Table S21.** Results for the CART model for location 4 (anal). The last column gives the number of instances out of 100 runs that the covariate is used in the splitting of tree branches.
| Covariate                                                                 | Count |
|---------------------------------------------------------------------------|-------|
| Time delay between assault and exam = 0 days                              | 32    |
| Time delay between assault and exam = 1 day                              | 2     |
| Victim age (yr)                                                           | 46    |
| Victim gender M                                                           | 0     |
| Loss of memory U                                                          | 0     |
| Loss of memory Y                                                          | 2     |
| Number of offenders = 1                                                   | 35    |
| Number of offenders = U                                                   | 0     |
| Consensual sex in prior 5 days U                                          | 0     |
| Consensual sex in prior 5 days Y                                          | 84    |
| Consensual sex in prior 24 hours U                                       | 7     |
| Consensual sex in prior 24 hours Y                                       | 0     |
| Known ejaculation U                                                       | 0     |
| Known ejaculation Y                                                       | 84    |
| Condom used U                                                             | 4     |
| Condom used Y                                                             | 0     |
| Shower or bath before exam U                                              | 0     |
| Shower or bath before exam Y                                              | 13    |
| Vaginal penetration of victim by offender U                              | 1     |
| Vaginal penetration of victim by offender Y                              | 28    |
| Anal penetration of victim by offender U                                  | 5     |
| Anal penetration of victim by offender Y                                  | 3     |
| Oral penetration of victim by offender U                                 | 0     |
| Oral penetration of victim by offender Y                                 | 0     |
| Offender’s mouth on genitals U                                           | 0     |
| Offender’s mouth on genitals Y                                           | 0     |
| Offender’s mouth on breast U                                             | 1     |
| Offender’s mouth on breast Y                                             | 0     |
| Offender’s mouth on other body parts U                                   | 1     |
| Offender’s mouth on other body parts Y                                   | 3     |
| Digital penetration of victim by offender U                              | 4     |
| Digital penetration of victim by offender Y                              | 0     |
| Oral penetration of offender by victim U                                 | 0     |
| Oral penetration of offender by victim Y                                 | 9     |
| Strangled U                                                               | 0     |
| Strangled Y                                                               | 0     |
| Punched U                                                                 | 0     |
| Punched Y                                                                 | 0     |
| Stabbed U                                                                 | 0     |
| Stabbed Y                                                                 | 0     |
| Vaginal Injury U                                                          | 0     |
| Vaginal Injury Y                                                          | 0     |
| Other Injury U                                                            | 0     |
| Other Injury Y                                                            | 7     |

**Table S22.** Results for the CART model for location 5 (clothing). The last column gives the number of instances out of 100 runs that the covariate is used in the splitting of tree branches.
| Covariate                                      | Count |
|------------------------------------------------|-------|
| Time delay between assault and exam = 0 days   | 46    |
| Time delay between assault and exam = 1 day    | 3     |
| Victim age (yr)                                | 16    |
| Victim gender M                                | 0     |
| Loss of memory U                               | 0     |
| Loss of memory Y                               | 5     |
| Number of offenders = 1                        | 7     |
| Number of offenders = U                        | 0     |
| Consensual sex in prior 5 days U               | 0     |
| Consensual sex in prior 5 days Y               | 0     |
| Consensual sex in prior 24 hours U             | 0     |
| Consensual sex in prior 24 hours Y             | 0     |
| Known ejaculation U                            | 2     |
| Known ejaculation Y                            | 4     |
| Condom used U                                  | 0     |
| Condom used Y                                  | 0     |
| Shower or bath before exam U                   | 0     |
| Shower or bath before exam Y                   | 0     |
| Vaginal penetration of victim by offender U    | 12    |
| Vaginal penetration of victim by offender Y    | 0     |
| Anal penetration of victim by offender U       | 0     |
| Anal penetration of victim by offender Y       | 1     |
| Oral penetration of victim by offender U       | 3     |
| Oral penetration of victim by offender Y       | 2     |
| Offender’s mouth on genitals U                 | 0     |
| Offender’s mouth on genitals Y                 | 1     |
| Offender’s mouth on breast U                   | 1     |
| Offender’s mouth on breast Y                   | 1     |
| Offender’s mouth on other body parts U         | 0     |
| Offender’s mouth on other body parts Y         | 4     |
| Digital penetration of victim by offender U    | 0     |
| Digital penetration of victim by offender Y    | 1     |
| Oral penetration of offender by victim U       | 2     |
| Oral penetration of offender by victim Y       | 1     |
| Strangled U                                    | 0     |
| Strangled Y                                    | 0     |
| Punched U                                      | 0     |
| Punched Y                                      | 0     |
| Stabbed U                                      | 0     |
| Stabbed Y                                      | 0     |
| Vaginal Injury U                               | 0     |
| Vaginal Injury Y                               | 0     |
| Other Injury U                                 | 0     |
| Other Injury Y                                 | 0     |

**Table S23.** Results for the CART model for location 6 (foreign material). The last column gives the number of instances out of 100 runs that the covariate is used in the splitting of tree branches.
**Figure S1.** The number of SAKs that generate at least one DNA profile for CODIS vs. the number of samples tested per kit, under the logistic regression model and under the SAFE Policy, along with 95% CIs.
**Figure S2.** The number of SAKs that generate at least one DNA profile for CODIS vs. the number of samples tested per kit, under the LASSO-LR model and under the SAFE Policy, along with 95% CIs.
Figure S3. An illustration of the CART model for one iteration for location 1 (body surface). The three numbers in each node represent the predicted value of the dependent variable (1 if the sample is CODIS-uploadable, and 0 otherwise), the proportion of the test set samples that are CODIS-uploadable in this branch, and the proportion of the test set samples that are in this branch. For example, in the lower right node in Fig. S7, 39% of the test set samples have the time delay between assault and exam = 0 days. Within this 39% of the test set samples, 55% are CODIS-uploadable, and so the model (because 55% > 50%) predicts that the sample is CODIS-uploadable.
Figure S4. An illustration of the CART model for one iteration for location 2 (genital). The three numbers in each node represent the predicted value of the dependent variable (1 if the sample is CODIS-uploadable, and 0 otherwise), the proportion of the test set samples that are CODIS-uploadable in this branch, and the proportion of the test set samples that are in this branch. For example, in the lower right node in Fig. S7, 39% of the test set samples have the time delay between assault and exam = 0 days. Within this 39% of the test set samples, 55% are CODIS-uploadable, and so the model (because 55% > 50%) predicts that the sample is CODIS-uploadable.
Figure S5. An illustration of the CART model for one iteration for location 3 (oral). The three numbers in each node represent the predicted value of the dependent variable (1 if the sample is CODIS-uploadable, and 0 otherwise), the proportion of the test set samples that are CODIS-uploadable in this branch, and the proportion of the test set samples that are in this branch. For example, in the lower right node in Fig. S7, 39% of the test set samples have the time delay between assault and exam = 0 days. Within this 39% of the test set samples, 55% are CODIS-uploadable, and so the model (because 55% > 50%) predicts that the sample is CODIS-uploadable.
**Figure S6.** An illustration of the CART model for one iteration for location 4 (anal). The three numbers in each node represent the predicted value of the dependent variable (1 if the sample is CODIS-uploadable, and 0 otherwise), the proportion of the test set samples that are CODIS-uploadable in this branch, and the proportion of the test set samples that are in this branch. For example, in the lower right node in Fig. S7, 39% of the test set samples have the time delay between assault and exam = 0 days. Within this 39% of the test set samples, 55% are CODIS-uploadable, and so the model (because 55% > 50%) predicts that the sample is CODIS-uploadable.
Figure S7. An illustration of the CART model for one iteration for location 5 (clothing).

The three numbers in each node represent the predicted value of the dependent variable (1 if the sample is CODIS-uploadable, and 0 otherwise), the proportion of the test set samples that are CODIS-uploadable in this branch, and the proportion of the test set samples that are in this branch. For example, in the lower right node, 39% of the test set samples have the time delay between assault and exam = 0 days. Within this 39% of the test set samples, 55% are CODIS-uploadable, and so the model (because 55% > 50%) predicts that the sample is CODIS-uploadable.
Figure S8. An illustration of the CART model for one iteration for location 6 foreign material). The three numbers in each node represent the predicted value of the dependent variable (1 if the sample is CODIS-uploadable, and 0 otherwise), the proportion of the test set samples that are CODIS-uploadable in this branch, and the proportion of the test set samples that are in this branch. For example, in the lower right node in Fig. S7, 39% of the test set samples have the time delay between assault and exam = 0 days. Within this 39% of the test set samples, 55% are CODIS-uploadable, and so the model (because 55% > 50%) predicts that the sample is CODIS-uploadable.
Figure S9. The number of SAKs that generate at least one DNA profile for CODIS vs. the number of samples tested per kit, under the CART model and under the SAFE Policy, along with 95% CIs.
Figure S10. Under the logistic regression model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the expected cost per SAK, under the SAFE Policy (*), the Priority Policy (green -), the Nonlinear Priority Policy (blue -·-), and the Priority Policy with additional synthetic samples (red - -). The 95% CIs are depicted for each integer value of the parameter $n$ for the Priority Policy and for each integer value of the mean number of samples tested per SAK for the Nonlinear Priority Policy and the Priority Policy with additional synthetic samples.
**Figure S11.** Under the CART model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the expected cost per SAK, under the SAFE Policy (*), the Priority Policy (green –), the Nonlinear Priority Policy (blue ·), and the Priority Policy with additional synthetic samples (red - -). The 95% CIs are depicted for each integer value of the parameter $n$ for the Priority Policy and for each integer value of the mean number of samples tested per SAK for the Nonlinear Priority Policy and the Priority Policy with additional synthetic samples.
Figure S12. Under the (San Francisco Salary, Prescreening) scenario (Table 3) and the LASSO-LR model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the expected cost per SAK, under the SAFE Policy (*), the Priority Policy (green –), the Nonlinear Priority Policy (blue ·), and the Priority Policy with additional synthetic samples (red - -). The 95% CIs are depicted for each integer value of the parameter n for the Priority Policy and for each integer value of the mean number of samples tested per SAK for the Nonlinear Priority Policy and the Priority Policy with additional synthetic samples.
Figure S13. Under the (Average Salary, No Prescreening) scenario (Table 3) and the LASSO-LR model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the expected cost per SAK, under the SAFE Policy (*), the Priority Policy (green –), the Nonlinear Priority Policy (blue ·), and the Priority Policy with additional synthetic samples (red - -). The 95% CIs are depicted for each integer value of the parameter $n$ for the Priority Policy and for each integer value of the mean number of samples tested per SAK for the Nonlinear Priority Policy and the Priority Policy with additional synthetic samples.
Figure S14. Under the (Average Salary, Prescreening) scenario (Table 3) and the LASSO-LR model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the expected cost per SAK, under the SAFE Policy (*), the Priority Policy (green -), the Nonlinear Priority Policy (blue ·), and the Priority Policy with additional synthetic samples (red - -). The 95% CIs are depicted for each integer value of the parameter $n$ for the Priority Policy and for each integer value of the mean number of samples tested per SAK for the Nonlinear Priority Policy and the Priority Policy with additional synthetic samples.