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

Zhengli Wang[a], Kevin MacMillan[b], Mark Powell[b], and Lawrence M. Wein[a,1]

[a]Graduate School of Business, Stanford University, Stanford, CA 94305; and [b]Criminalistics Laboratory, San Francisco Police Department, San Francisco, CA 94158

Edited by Charles F. Manski, Northwestern University, Evanston, IL, and approved April 29, 2020 (received for review January 19, 2020)

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 (based on covariates gleaned from sexual assault kit questionnaires) which samples are most probative. This model is embedded within an optimization framework that selects which samples to test from each kit to maximize the Combined DNA Index System (CODIS) yield (i.e., the number of kits that generate at least one DNA profile for the criminal DNA database) subject to a budget constraint. Our analysis predicts that, relative to a policy that tests only the samples deemed probative by the sexual assault forensic examiners, the proposed policy increases the CODIS yield by 45.4% without increasing the cost. Full testing of all samples has a slightly lower cost-effectiveness than the selective policy based on forensic examiners, but more than doubles the yield. In over half of the sexual assaults, a sample was not collected during the forensic medical exam from the body location deemed most probative by the machine learning model. Our results suggest that electronic forensic records coupled with machine learning and optimization models could enhance the effectiveness of criminal investigations of sexual assaults.

forensic science | sexual assaults | crime solving | machine learning | optimization

A sexual assault kit (SAK) contains biological evidence collected from a victim during a forensic medical examination that occurs after a sexual assault. The SAK is transferred to law enforcement personnel, who are then responsible for submitting the SAK to a forensic laboratory for testing. The forensic laboratory attempts to recover DNA from the SAK, which is then uploaded into the Combined DNA Index System (CODIS), a national database of DNA profiles from known offenders/arrestees of both sexual assaults and nonsexual crimes. If a recently uploaded DNA profile from a SAK matches an existing DNA profile in CODIS, then it can provide a promising lead to law enforcement as to the identity of the sexual offender.

Although this crime-solving approach appears to have considerable potential, >200,000 SAKs in the United States have been held in law enforcement storage facilities without ever being submitted to a forensic laboratory for DNA testing (1, 2). Recently, several municipalities have tested their backlog of SAKs (3, 4) and these undertakings appear to be cost-effective (4–6).

The Bureau of Justice Assistance with the US Department of Justice has initiated the Sexual Assault Kit Initiative (SAKI), which provides funding to jurisdictions for testing their SAK backlog and investigating the subsequent CODIS hits (7). The Sexual Assault Forensic Evidence Reporting Act of 2017 (8) and the Manhattan District Attorney's Sexual Assault Kit Backlog Elimination Grant Program (9) also provide funding for testing the SAK backlog nationwide.

SAKs typically contain a number of samples from various body locations and clothing, such as undergarments, and the number of samples and their body locations vary across SAKs. Due to a combination of factors—including limited budgets, heavy workloads at forensic laboratories, and legislatively mandated turnaround time guidelines for submitting and processing SAKs—many municipalities are using selective testing strategies, where only a few (usually, up to three) (9–11) of the samples from the SAK are actually tested. In these cases, the tested samples are typically those that are deemed most probative (i.e., most likely to contain foreign DNA) based on information from police reports (if the SAK had been backlogged) or SAK questionnaires (10, 11), although sometimes are chosen randomly (ref. 9, p. 11). This approach is consistent with recommendation 25 in ref. 12, which calls for the prioritization of evidentiary items when processing SAKs.

It is not clear whether these selective testing strategies are preferable to full testing of all samples in a SAK. There are two competing forces at play: Selective testing has the potential to avoid wasting resources on testing less probative samples, but—in addition to the variable cost per sample tested—there is a fixed cost associated with processing a SAK that is independent of the number of samples tested within a SAK. That is, testing only the most probative samples does not take advantage of the economies of scale inherent in testing a SAK.

In this study, we use data from 868 SAKs tested by the San Francisco Police Department Criminalistics Laboratory during 2017 to 2019. For each of these SAKs, sexual assault forensic

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).

Author contributions: Z.W. and L.M.W. designed research; Z.W. and L.M.W. performed research; Z.W., K.M., M.P., and L.M.W. analyzed data; and Z.W., K.M., M.P., and L.M.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: lwein@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2001103117/-/DCSupplemental.

First published June 1, 2020.

examiners (SAFEs) (if these examiners are nurses, they are often referred to as sexual assault nurse examiners [SANEs]) recommended the most probative 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 for covariates associated with each sexual assault. In addition, we estimate the fixed (i.e., independent of the number of samples in a SAK) and 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*.

This 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 SAFEs chose more probative samples to obtain in the first place?

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
Victim gender at birth (M/F)	150/718
Loss of memory (Y/N/U)	492/332/44
No. of offenders (1/>1/U)	654/100/114
Consensual sex in prior 5 d (Y/N/U)	201/605/62
Consensual sex in prior 24 h (Y/N/U)	89/155/624
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)	331/217/320
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)	86/343/439
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)	170/278/420
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)	225/621/22
Other injury (Y/N/U)	343/507/18

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

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 868 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 (*SI Appendix*, 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).

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 L^1 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

Table 2. In total and broken down by location, the number of samples tested, the number of probative samples (as deemed by the SAFE), and the number of CODIS-uploadable samples

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

maximizing the data likelihood while keeping the set of nonzero elements in the estimated parameters to be small. This is achieved by adding an L^1 regularization term, commonly known as the LASSO penalty term, to the likelihood function. 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 separately for $i = 1, \ldots, 6$, using all of the covariates in *SI Appendix*, Table S1,

$$(\hat{\beta}_{0,LASSOLR}^{(i)}, \hat{\beta}_{LASSOLR}^{(i)}) = \arg \max_{\beta_0^{(i)}, \beta^{(i)}} - \sum_{j=1}^{n_i} \log \left[h\left(y_j^{(i)} (\beta_0^{(i)} + \beta^{(i)\prime} x_j^{(i)}) \right) \right] + \lambda \sum_{k=1}^{d} |\beta_k^{(i)}|, \quad [1]$$

where λ is known as the regularization parameter. Usually, the larger λ is, the fewer nonzero β 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 λ, use the validation set for choosing the best λ, and then use the test set to measure performance. Our performance metric used to maximize λ and measure performance is a normalized area under the curve (AUC), which is computed as follows. 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 test all samples in the SAK if there are fewer 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 AUC of this plot. This procedure is repeated 100 times and the average result is presented.

Cost Estimation. If n samples are tested in a SAK, then we assume the cost is $F + Vn$, where F is the fixed cost and V is the variable cost per sample. The costs include out-of-pocket material costs and the labor costs associated with the processing times. Only the marginal cost of testing the backlog is relevant within the context of our decision problem (i.e., how many samples to test in a SAK). Hence, we ignore equipment and overhead costs, thereby implicitly assuming that—relative to selective testing—full testing would not require new equipment or facilities; this is in contrast to the cost estimates from Project FORESIGHT (14), which includes all costs and allows one to compare cost efficiencies for various laboratory sizes. We begin by ignoring the costs associated with prescreening for biological materials, which was discontinued in San Francisco on November 1, 2017, and return to this issue at the end of this subsection.

All material costs are variable and include the screening cost per sample (scalpels, tubes) of \$1, the extraction cost per sample (EZ1, Qiacube, tubes) of \$11.35, the quantitation cost per sample (plate, adhesive, reagents quantitation trio) of \$6.20, the amplification cost per sample (plate, caps, reagents, globalfiler) of \$22.20, and the capillary electrophoresis (CE) cost per sample (array, buffer, plate, septa, size standard, formamide) of \$1.80, for a total of \$42.55 per sample.

Time estimates (fixed and variable, in minutes) include the time to screen the SAK (picture and inventory) of $30 + 5n$, DNA extraction time of $10n$, DNA quantitation time of $15 + 0.5n$, DNA amplification time of 15, CE time of 10, analysis and report writing time of $30 + 15n$, and review time of $30 + 10n$, for a total time of $130 + 40.5n$ min. Including benefits, the current salary at the San Francisco Police Department Criminalistics Laboratory is approximately \$170,000 (15), or \$1.417/min, assuming 2,000 h/y. Therefore, $F = 130 \times 1.417 = \$184.57$ and $V = 42.55 + (40.5 \times 1.417) = \$99.92$ per sample.

In our sensitivity analysis, we also investigate three alternative cost functions. First, we assume that screening for biological materials is undertaken prior to DNA processing. This increases the variable time for screening from 5 to 45 min, changing V to $42.55 + (80.5 \times 1.417) = \$156.62$ per sample. Next, because San Francisco salaries are considerably higher than the national average, we consider a smaller salary under both the prescreening and no prescreening scenarios. A 2017 to 2018 survey of US crime laboratories associated with Project FORESIGHT yields a median salary of \$118,648 (table 12 in ref. 14). San Francisco was part of this survey and the median salary reported by Project FORESIGHT for 2017 to 2018 was \$150,800. Hence, we use $\frac{118,648}{150,800} \times 170,000 = \$133,754$ as an estimate of the typical annual US salary, which converts to \$1.115/min. Substituting \$133,754 for \$170,000 yields $F = 130 \times 1.115 = \$144.90$ regardless of whether or not prescreening is undertaken, $V = 42.55 + (40.5 \times 1.115) = \$87.69$ per sample with no prescreening, and $42.55 + (80.5 \times 1.115) = \$132.28$ per sample with prescreening. The F/V ratios for the four scenarios are displayed in Table 3.

Optimization Problem. In *SI Appendix*, section 2, we mathematically formulate the problem of choosing which samples to test from each SAK to maximize the probability that a SAK is CODIS uploadable subject to a constraint on the mean cost per SAK, where we require—so as not to violate the spirit of testing the backlog—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*, Fig. S2) 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). Possible underlying reasons include incentive heteroskedasticity (i.e., SAFEs varying in their level of aggressiveness at identifying samples as probative) and information asymmetry (i.e., SAFEs having access to additional information beyond the list of covariates in *SI Appendix*, Table S1) (18). However, we do not believe that this phenomenon plays a significant role in our analysis for two reasons. First, the great majority of SAFEs operated on a region of

SOCIAL SCIENCES

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

$$\rho_{(y^{(i_1)}, y^{(i_2)})|x} = \frac{\sum_j (y_j^{(i_1)} - E[y_j^{(i_1)}|x_j])(y_j^{(i_2)} - E[y_j^{(i_2)}|x_j])}{\sqrt{\sum_j (y_j^{(i_1)} - E[y_j^{(i_1)}|x_j])^2}\sqrt{\sum_j (y_j^{(i_2)} - E[y_j^{(i_2)}|x_j])^2}}, \quad [2]$$

with standard error

$$\sqrt{\frac{1 - \rho_{(y^{(i_1)}, y^{(i_2)})|x}}{n-2}}, \quad [3]$$

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 (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 (5) using data from 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 $\frac{0.229/397}{0.466/912} 81.34 = 91.77$, 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 $\frac{81.34}{0.466/912}$. 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

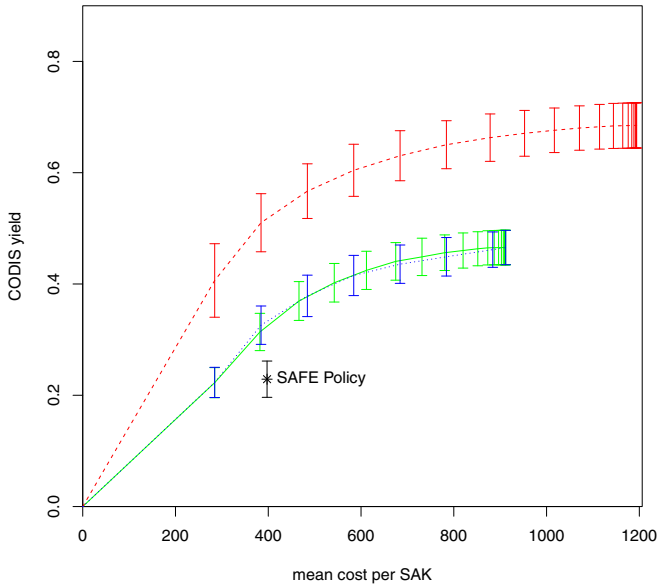

Fig. 1. Under the LASSO-LR model, the CODIS yield (i.e., the probability that a SAK generates at least one CODIS-uploadable sample) vs. the mean cost per SAK, under the SAFE policy (*), the priority policy (green solid line), the nonlinear priority policy (blue dotted line), and the priority policy with 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.

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 suboptimal. 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

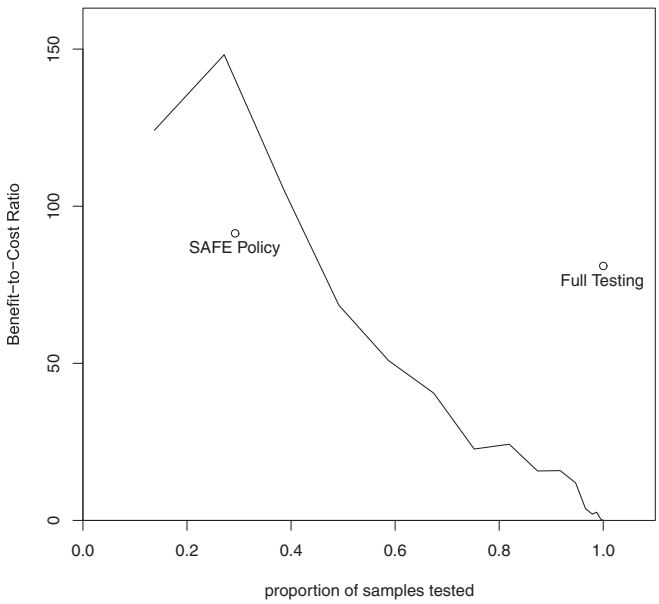

Fig. 2. Under the LASSO-LR policy, the marginal benefit-to-cost ratio of the priority policy vs. the proportion of samples tested. The benefit-to-cost ratio is 91.77 for the SAFE policy and 81.34 for full testing.

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 $\frac{723}{1468} = 0.493$ (5), and the CODIS yield in the Manhattan District Attorney's Office's Sexual Assault Kit backlog Elimination Grant Program, which allowed one random sample from a SAK to be tested, was less than $\frac{18,803}{55,252} = 0.340$, 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 three most probative samples per SAK, was 0.460 and was $\frac{2,934}{4,966} = 0.591$ in Cuyahoga County, OH under full testing (23), which are higher than the yields derived here. Moving forward, it is important to understand the factors (other than the number of samples tested) affecting the variation in CODIS yield.

One key covariate that is missing in our model is the distinction between a stranger assault and a nonstranger assault. The stranger vs. nonstranger relationship did not affect the probability of CODIS upload in ref. 24. An alternative analysis could consider the number of CODIS hits (i.e., the number of CODIS uploads that match an existing DNA profile in CODIS) rather than the number of CODIS entries as the dependent variable 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. *Criminol. 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. *Criminol. 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.

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* **1**, 45–55 (2019).
7. Bureau of Justice Assistance, US Department of Justice, Sexual assault kit initiative (SAKI). https://www.bja.gov/ProgramDetails.aspx?Program_-ID=117. Accessed 29 July 2019.
8. US Congress, S.1766–SAFER Act of 2017. https://www.congress.gov/bill/115thcongress/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-Eliminaton-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 (2017). www.endthebacklog.org/blog/guest-post-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. Luminais, 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).