JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2018, Vol. 63, No. 4
doi: 10.1111/1556-4029.13739
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL; CRIMINALISTICS

Can Wang,[1] *M.S.; and Lawrence M. Wein*,[2] *Ph.D.*

Analyzing Approaches to the Backlog of Untested Sexual Assault Kits in the U.S.A.

ABSTRACT: Motivated by the debate over how to deal with the huge backlog of untested sexual assault kits in the U.S.A., we construct and analyze a mathematical model that predicts the expected number of hits (i.e., a new DNA profile matches a DNA sample in the criminal database) as a function of both the proportion of the backlog that is tested and whether the victim–offender relationship is used to prioritize the kits that are tested. Refining the results in Ref. (Criminol Public Policy, 2016, 15, 555), we use data from Detroit, where government funding was used to process ≈15% of their backlog, to predict that prioritizing stranger kits over nonstranger kits leads to only a small improvement in performance (a 0.034 increase in the normalized area under the curve of the hits vs. proportion of backlog tested curve). Two rough but conservative cost-benefit analyses—one for testing the entire backlog and a marginal one for testing kits from nonstranger assaults—suggest that testing all sexual assault kits in the backlog is quite cost-effective: for example, spending ≈$1641 to test a kit averts sexual assaults costing ≈$133,484 on average.

KEYWORDS: forensic science, forensic DNA, sexual assaults, crime solving, probabilistic modeling, statistics

After a sexual assault, victims are typically advised to undergo a forensic medical examination, where biological evidence is collected as part of a sexual assault kit (SAK). The SAK is transferred to law enforcement personnel, who are responsible for submitting the SAK to a forensic laboratory. If DNA can be recovered from the SAK, it can be entered into CODIS (Combined DNA Index System), which is a national database of DNA profiles from known offenders/arrestees (referred to as the offender file) and from crime scene evidence (referred to as the forensic file, where the identities are unknown); both files contain DNA samples obtained from sexual assaults (SAs) and from nonsexual crimes. If the new CODIS entry generates a hit (i.e., its DNA matches the DNA of an existing DNA sample in CODIS), then it can provide either a promising lead if it hits a sample in the offender file, or a link to an earlier crime if it hits a sample in the forensic file.

Despite the potential value of this forensic infrastructure, >200,000 SAKs have never been submitted to a laboratory for DNA testing and instead reside in law enforcement storage facilities (1,2). Testing backlogged SAKs, some of which may be many years old, may generate other benefits aside from CODIS hits (3–6): It helps populate CODIS, sends positive messages to victims, moves the justice system toward better responses to sexual assaults, holds offenders accountable through arrests and convictions, and prevents wrongful prosecutions. The National Institute of Justice funded studies in Los Angeles (7), New Orleans (8), Detroit (3,9) and Houston (10) to process some of their untested SAKs and to gain a better understanding of how to address this immense national backlog. Although the hit results from all four studies are roughly comparable (table 13 in (10)), different conclusions were drawn in the two largest studies: SAs are characterized as either stranger SAs or nonstranger SAs based on the victim–offender relationship, and researchers in Detroit (3) recommend that all untested SAKs be tested, while researchers in Los Angeles (7) recommend not to test all SAKs but rather to focus on stranger SAKs (i.e., SAKs from stranger SAs). We investigate the reason for this divergence in the Discussion, but make two points now: The Detroit recommendation focuses on hit performance and is based at least partly on the adoption of a liberal definition of equivalent hit probabilities between stranger SAKs and nonstranger SAKs (to have different hit probabilities required the 90% confidence interval [CI] of the odds ratio to fall outside [0.4,2.50], and the estimated odds ratio was 1.78), and the LA recommendation is based on downstream metrics such as arrests, charges, and convictions. Hence, a study's choice of performance measures (e.g., CODIS updates, CODIS hits, or convictions) may influence its conclusions.

Jurisdictions facing large SAK backlogs have four main options: (i) test no SAKs in the backlog, (ii) test all SAKs in the backlog with no prioritization, (iii) test all SAKs in the backlog with prioritization (e.g., process all stranger SAKs within the statute of limitations before processing any SAKs beyond the statute of limitations or any nonstranger SAKs), and (iv) processing only some of the SAKs (e.g., only the stranger SAKs within the statute of limitations). Option (ii) is referred to as the "forklift" approach and was implemented by New York City, where it was deemed too time-consuming (due to the limitations of their information systems) to sort and prioritize the SAKs before processing (11).

We build a mathematical model and calibrate it using the Detroit SAK data (LA's data are not detailed enough to support a mathematical model) to address two questions: Should all

[1]Electrical Engineering Department, Stanford University, Stanford, CA 94305, USA.
[2]Graduate School of Business, Stanford University, Stanford, CA 94305, USA.

Received 2 Aug. 2017; and in revised form 28 Nov. 2017; accepted 29 Nov. 2017.

SAKs be tested, and if not, which SAKs should be tested, and in particular, should stranger SAKs be prioritized over nonstranger SAKs? Our model incorporates and quantifies the extent to which offenders specialize in either stranger or nonstranger SAs, and if they specialize, which type of offender is more likely to generate recoverable DNA, more likely to commit more additional crimes that are in CODIS or the backlog, and more likely to have these additional crimes end up in the backlog. After estimating the model parameters, we generate curves of the expected number of hits versus the proportion of the backlog that is processed, for policies that do—and do not—prioritize stranger SAKs over nonstranger SAKs. As recommended in Ref. (7,12), we also use this analysis as a basis for two cost-benefit analyses: One for testing the entire backlog relative to testing none of the backlog, and a marginal analysis for testing nonstranger SAKs. Hence, referring back to the four options, we explicitly compare options (i) and (ii), and also option (iv) with respect to stranger vs. nonstranger SAKs, and recommend that all backlog SAKs be tested. Although we perform other analyses that shed some light on option (iii) (e.g., prioritization based on stranger vs. nonstranger SAK, and on weapon use), we do not make explicit recommendations regarding prioritization and in particular perform no prioritization analysis that incorporates whether a SAK is beyond the statute of limitation. That is, we address the strategic issue of which SAKs should be tested, but do not fully address the operational issue of what order to test them in.

Materials and Methods

Data

The Detroit SAK data are publicly available from the National Archive of Criminal Justice Data (13) and are described in detail in Ref. (3,9). The census undertaken in Ref. (9) uncovered 11,219 SAKs in Detroit police property from SAs occurring during 1980 to November 1, 2009. Of these 11,219 SAKs, 2512 had been submitted for testing, leaving 8707 in the backlog. From this unsubmitted backlog, the researchers tested a sample of 1595 SAKs in four testing groups: 445 stranger SAKs (testing group 1), 449 nonstranger SAKs (testing group 2), 351 SAKs that were beyond the statutes of limitations (testing group 3), and 350 SAKs used to compare DNA testing methods (testing group 4). For most of these SAKs, we know whether the SA type was stranger or nonstranger, whether or not a weapon was used, whether or not DNA was recovered, whether there were any hits to the offender file in CODIS, the forensic file in CODIS or to the other 1594 SAKs tested. For most of the CODIS hits, we know only the most recent qualifying offense (e.g., if there was a hit to an offender in CODIS that had committed three crimes—a SA followed by a burglary followed by a homicide—we only know about the homicide) for hits to the offender file, and no crime-type information for hits to the forensic file. In addition, we have publicly available data (13) from a pilot project where 400 randomly selected SAKs from Detroit's backlog of 11,219 SAKs were tested and their stranger vs. nonstranger SAK status were reported based on police reports (Shaw J. Justifying injustice: How the criminal justice system explains its response to sexual assault. Unpublished doctoral dissertation. East Lansing, MI: Michigan State University, 2014); these data are used to weight the 1595 samples so that they are representative of the entire Detroit population.

A Model of the Data-generating Process

We develop a model of the data-generating process, where the observed data are for offenders who have at least one SAK with recovered DNA in the tested portion of the backlog. That is, we develop a probabilistic model for an offender who has at least one SAK with recovered DNA in the tested portion of the backlog. This model serves several purposes: Its estimated parameter values shed light on the behavior of sexual offenders, it allows us to compare the hit performance of various policies (e.g., the forklift approach or the priority of stranger SAKs over nonstranger SAKs), and its output provides key input to the cost-benefit analyses.

We assume that a SAK in the backlog is a stranger SAK with probability q_s and a nonstranger SAK with probability q_{ns}, and that DNA can be recovered from each stranger SAK with probability d_s and can be recovered from each nonstranger SAK with probability d_{ns}, independent of any additional information about the offender. The values of d_s and d_{ns} could differ if the nature of the assault, the time delay in victim reporting or the extent of victim cleanup prior to the forensic medical examination differs systematically between stranger and nonstranger SAs. To allow heterogeneous specialization of SAs by an offender, we let S be a beta random variable with parameters α and β. Each offender is assigned a different value of S from this distribution, and this value determines the probability that a SA by this offender is a stranger SA. The possible values s of S vary between 0 and 1, where, for example, $s = 1$ corresponds to an offender who commits only stranger SAs, $s = 0$ corresponds to an offender who commits only nonstranger SAs, and $s = 0.5$ corresponds to an offender who is equally likely to commit stranger or nonstranger SAs.

Let X be the random number of additional (i.e., aside from the conditioned SAK with recovered DNA in the tested portion of the backlog) crimes—sexual and nonsexual—associated with an offender, either related to a DNA profile in CODIS (our model does not differentiate between samples in the offender file and samples in the forensic file in CODIS, and we return to this issue in the Discussion) or related to a SAK with recoverable DNA in the tested and untested portions of the SAK backlog. To allow for possible overdispersion, we assume that X conditioned on $S = s$ is a negative binomial random variable with parameters r and $p = 1 - \left(\frac{s}{1-p_s} + \frac{1-s}{1-p_{ns}}\right)^{-1}$, implying that the mean of X is $\frac{rp}{1-p} = s\frac{rp_s}{1-p_s} + (1-s)\frac{rp_{ns}}{1-p_{ns}}$, where $\frac{rp_s}{1-p_s}$ is the mean number of additional crimes for an offender who commits only stranger SAs, and $\frac{rp_{ns}}{1-p_{ns}}$ is the mean number of additional crimes for an offender who commits only nonstranger SAs. That is, the mean number of additional crimes an offender commits depends in a linear way on the offender's level of specialization between stranger and nonstranger SAs. In addition, the variance-to-mean ratio of the number of additional crimes, which is $\frac{1}{1-p} = s\frac{1}{1-p_s} + (1-s)\frac{1}{1-p_{ns}}$, also depends in a linear way on the offender's level of specialization. Before settling on this set of assumptions for X, we considered two alternative models. First, X given $S = s$ is negative binomial with $r = sr_s + (1-s)r_{ns}$ and $p = p_s = p_{ns}$, which implies that the mean of X varies linearly in s and the variance-to-mean ratio is independent of s; this alternative gave very similar results to the chosen model (e.g., the hit probability of the two models differed by 10^{-4}). Second, X given $S = s$ is a Poisson random variable with a mean that varies linearly in s; this simpler model, which does not allow for

overdispersion, generated qualitatively similar results to the chosen model.

Let Z be the total number of SAKs with recoverable DNA in the backlog associated with an offender. Conditioned on $X = x$ and $S = s$, we assume that Z is one plus a binomial random variable with parameters x and $sb_s + (1 - s)b_{ns}$, where b_s is the probability that an additional crime is in the SAK backlog for an offender who commits only stranger SAs, and b_{ns} is the probability that an additional crime is in the SAK backlog for an offender who commits only nonstranger SAs.

Of the Z SAKs in the backlog related to an offender, we need to model how many are stranger SAKs and how many are nonstranger SAKs. Conditioned on $Z = z$ and $S = s$, let Z_s be the total number of stranger SAKs with recoverable DNA in the backlog related to an offender, which we assume is a binomial random variable with parameters z and s. The total number of nonstranger SAKs with recoverable DNA in the backlog related to an offender is denoted by $Z_{ns} = Z - Z_s$.

Finally, associated with an offender we define T_s and T_{ns} to be the number of stranger SAKs and nonstranger SAKs with recoverable DNA in the backlog that were tested in the Detroit study (3). Conditioned on $Z_s = z_s$ and $Z_{ns} = z_{ns}$, we assume that T_s is binomial with parameters z_s and θ_s, and T_{ns} is binomial with parameters z_{ns} and θ_{ns}, where θ_s and θ_{ns} are the proportions of stranger and nonstranger SAKs in the backlog that were tested in Ref. (3). We let $T = T_s + T_{ns}$ be the total number of tested SAKs with recoverable DNA in the backlog associated with an offender.

In §1.1 of the Supporting Information (SI), we estimate the DNA recovery probabilities (d_s, d_{ns}), the proportion of the backlog that are stranger or nonstranger SAKs (q_s, q_{ns}), and the testing probabilities (θ_s, θ_{ns}). We then use maximum likelihood estimation (§1.2-1.5 of the SI) to jointly estimate the model's remaining seven parameters $(\alpha, \beta, r, p_s, p_{ns}, b_s, b_{ns})$ from the observed data. Because CODIS has been in place for over 20 years, we can safely assume that this system is in a steady state (e.g., even though CODIS is slowly growing, older offenders are retiring and new offenders are starting their offending careers), and hence the model parameters do not vary over time.

Results

Parameter Values

In the slightly u-shaped beta probability density function (PDF) associated with our point estimates of (α, β) (Table 1), approximately three times as many offenders specialize in nonstranger SAs than stranger SAs (Fig. 1). The remaining point estimates suggest that—relative to an offender who specializes in nonstranger SAs—an offender who specializes in stranger SAs has a slightly higher DNA recovery probability (the DNA recovery probabilities differ from the estimates in (3,9), where they are referred to as CODIS entry rates, because they include data from only testing groups 1 and 2, and we include data from all four testing groups), nearly the same testing probability (and hence the imbalance between the number of stranger SAs and nonstranger SAs associated with an offender is not caused by how the backlog was sampled for testing), a twofold higher mean and threefold higher variance of additional crimes, and a fivefold higher probability that these additional crimes are in the SAK backlog (Table 1). The point estimates of (r, p_s) and (r, p_{ns}) (Table 1) also suggests that the number of additional crimes by an offender is overdispersed relative to the Poisson distribution (where the variance equals the mean), with variance-to-mean ratios of 2.878 and 1.906 for offenders specializing in stranger SAs and nonstranger SAs, respectively (Fig. 2).

However, there is considerable uncertainty surrounding most of these parameters. To gain a better sense of the uncertainty in the beta PDF, we use the k-means algorithm (14) to cluster 1000 bootstrapped (α, β) pairs (15) into five groups (§1.6 in the SI),

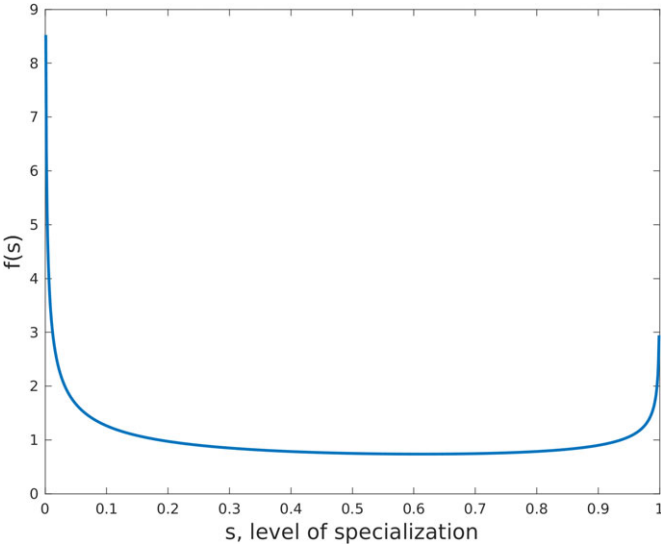


FIG. 1—*The probability density function of the beta distribution with the maximum likelihood estimates $\alpha = 0.5794$ and $\beta = 0.7333$. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 1—*Parameter values and 90% confidence intervals.*

Parameter	Description	Value	90% CI
d_s	probability DNA is recovered from stranger SAK	0.530	[0.487, 0.576]
d_{ns}	probability DNA is recovered from nonstranger SAK	0.463	[0.422, 0.507]
q_s	proportion of the backlog that are stranger SAKs	0.449	[0.366, 0.532]
q_{ns}	proportion of the backlog that are nonstranger SAK	0.551	[0.468, 0.634]
A	parameter of beta distribution for SA type specialization	0.579	[0.108, 3.221]
β	parameter of beta distribution for SA type specialization	0.733	[0.133, 4.065]
r	parameter of negative binomial distribution for stranger/ nonstranger SA offender	0.977	[0.255, 32.098]
p_s	parameter of negative binomial distribution for stranger SA offender	0.653	[0.042, 0.948]
p_{ns}	parameter of negative binomial distribution for nonstranger SA offender	0.475	$[2.129 \times 10^{-5}, 0.902]$
b_s	probability additional crime is in SAK backlog for stranger SA offender	0.294	[0.065, 0.489]
b_{ns}	probability additional crime is in SAK backlog for nonstranger SA offender	0.053	$[1.760 \times 10^{-9}, 0.234]$
θ_s	proportion of stranger SAKs in backlog that are tested	0.172	[0.142, 0.213]
θ_{ns}	proportion of nonstranger SAKs in backlog that are tested	0.181	[0.156, 0.214]

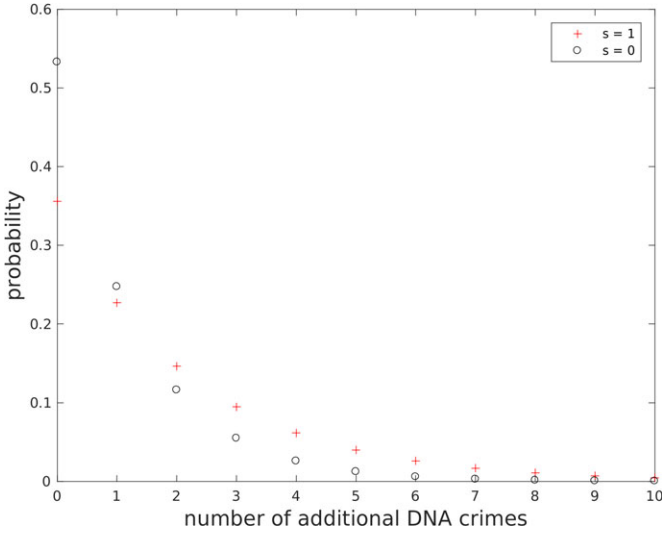


FIG. 2—*The probability mass functions of the negative binomial distribution with the maximum likelihood estimates for the stranger-specialized offenders* ($r = 0.977$, $p_s = 0.475$) *and nonstranger-specialized offenders* ($r = 0.977$, $p_{ns} = 0.294$). *[Color figure can be viewed at wileyonlinelibrary.com]*

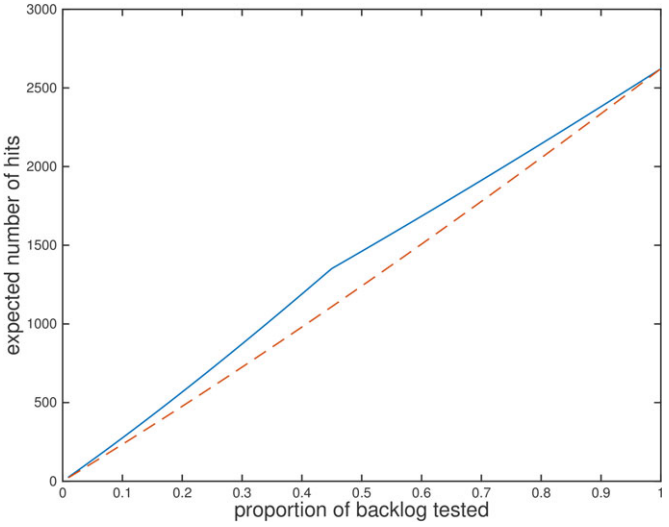


FIG. 3—*The main results. The expected number of hits vs. the proportion of the backlog tested, for the stranger-priority policy (–) and the no-priority policy (- -). [Color figure can be viewed at wileyonlinelibrary.com]*

and display the beta PDF associated with the center of each group (Fig. S3). This analysis suggests that 81.6% of the bootstrapped beta PDFs exhibit strong asymmetry associated with more specialization of nonstranger SAs, and 18.4% of the bootstrapped beta PDFs exhibit less specialization (i.e., are unimodal and roughly symmetric). The scatter plots (Fig. S4) of the 1000 pairs of bootstrapped estimates for (d_s, d_{ns}), (p_s, p_{ns}) and (b_s, b_{ns}) allow us to infer with statistical confidence that $d_s > d_{ns}$, $p_s > p_{ns}$ and $b_s > b_{ns}$.

Main Results

In our model, a SAK generates a hit if its offender has at least one match in CODIS or we test at least two SAKs associated with the offender. In §2 of the SI, we derive the expected number of hits as a function of the proportion of the backlog processed for both the stranger-priority policy (i.e., nonstranger SAKs are processed only after all stranger SAKs in the backlog are processed) and the no-priority policy (i.e., the victim–offender relationship is not used to prioritize the SAKs in the backlog). Under the point estimates in Table 1, the stranger-priority policy provides a modest improvement over the no-priority policy (Fig. 3). The maximum vertical distance between the two curves occurs at the point on the horizontal axis that represents the proportion of the backlog that is composed of stranger SAKs, 0.449. The expected total number of hits when the entire backlog of 8307 SAKs (8707 minus the 400 SAKs in the pilot project) is processed is 2621.8, which corresponds to a hit probability (i.e., the proportion of SAKs that generate at least one hit, which is referred to as the unconditional hit rate in Ref. (3)) of 0.316. A natural performance metric to compare the two curves in Fig. 3 is the normalized area under the curve (AUC), which is the area under the curve divided by 2621.8. The normalized AUCs are 0.527 for the stranger-priority policy and 0.482 for the no-priority policy.

A histogram of the normalized AUC for the stranger-priority policy minus the normalized AUC for the no-priority policy (Fig. S5) for the 1000 bootstrapped sets of parameters gives a mean of 0.031, a range of [0.006,0.056] and a 95% CI of [0.015,0.046]; that is, the stranger-priority policy always has a higher normalized AUC than the no-priority policy in this calculation.

Accounting for the Time to Sort the SAKs

One of the qualitative arguments put forth in Ref. (3) for not prioritizing stranger SAKs is that the modest increase in the hit probability would be offset by the additional time required to sort SAKs into stranger vs. nonstranger SAKs. It took 2958 h to review records to decide which 1600 SAKs to test in Detroit (16), which corresponds to 1.85 h per SAK. The review consisted of determining the adjudication status, the victim–offender relationship and the statute of limitations; the adjudication status took much more time than the other two criteria because multiple records (e.g., arrest records, court records) had to be pulled and checked. However, the majority of this 1.85 h consisted of finding and retrieving the paper files and sorting through the papers therein. As a conservative estimate (i.e., lower bound), we assume that it takes one hour of sorting time to obtain each stranger SAK. Although the Detroit review was undertaken by prosecutors, the victim–offender relationship data are likely to be in the police records and could be performed by a police officer. We assume a labor cost of $30/hr to search the records, which gives a sorting cost of $30/SAK to obtain each of the stranger SAKs in the backlog. Assuming a baseline testing cost of $1000/SAK (the reported range is between $400 and $1500 per SAK (17)), a plot (Fig. S6) of the expected number of hits vs. cost for the two policies using the parameters in Table 1 shows that the inclusion of the sorting cost reduces the difference in the normalized AUC between the two policies from 0.045 (i.e., 0.527–0.482) to 0.034.

A Cost-benefit Analysis

We undertake a rough, but conservative, cost-benefit analysis that compares the policy of testing all SAKs in the backlog to the policy of testing no SAKs in the backlog; in the Discussion, after introducing some additional information, we use this

analysis as a basis for a marginal cost-benefit analysis that focuses on testing the nonstranger SAKs in the backlog. The SAK testing cost and associated downstream costs will be compared to the expected cost and number of SAs averted by testing a SAK.

The testing cost has four components. We use table 7 in Ref. (18) to estimate the first three components, and to be conservative, we use costs from Los Angeles County, which are nearly twice as much as the average cost of the five cities studied in Ref. (18). It costs $980 for preliminary testing and the generation of the DNA profile. The CODIS entry cost is $167, which we multiply by the DNA recovery probability, $q_s d_s + q_{ns} d_{ns} = 0.491$, to get $82. The posthit processing cost is 2481–1147 = $1334, and we multiply this by the hit probability 0.316 to get $422. Finally, using table 1 in Ref. (19), we include the justice costs of $8503, which include the cost of investigation, legal defense, incarceration, parole and probation, and the offender productivity cost of $4610, which is the lost earnings due to incarceration. We multiply this sum, 8503 + 4610 = $13,113, by the estimated conviction probability 0.012 (the details yielding 0.012 are provided later) to get $157. Hence, the expected total testing cost is 980 + 82 + 422 + 157 = $1641. Note that this is an overestimate because there is some overlap (i.e., we are double counting) between the posthit processing costs in Ref. (18) and the justice costs in Ref. (19).

The expected number of SAs averted by testing a SAK is the probability that a SAK test leads to a conviction times the expected number of SAs potentially averted (i.e., averted if the test leads to a conviction); this approach conservatively assumes that a hit that results in something short of a conviction, such as an investigation or an arrest, does not avert any future SAs. The expected number of SAs potentially averted is the expected number of future SAs the offender commits before either desisting or being convicted, under the assumption that we do not test any SAKs in the backlog. The number of SAs per year of time at risk (e.g., not incarcerated) committed by a sexual offender varies widely across offenders, but the mean has been reported as 7.10 (table 2 in Ref. (20)); moreover, this rate decreases very slowly with the offender's age (20). This figure is largely based on self-reports, which experiments have suggested generate underestimates of the true offending rate (21). For adult offenders, there is a seven-year gap on average between the onset of SA activity and the time of the first conviction, and 20% of these offenders self-report that they had started to desist by the time of the first conviction (22); these figures allow us to estimate the time until an offender either desists or is convicted. We assume that DNA testing of SAKs in the backlog plays no role in the first conviction after the mean seven-year gap; that is, the seven-year gap is under the "business-as-usual" scenario where some SAKs are tested but many SAKs are left untested in a backlog. Because there can be many SA events for an offender over the gap between onset and first conviction, if the events happen regularly (i.e., at deterministic time intervals) over this gap and each event has the same probability of conviction, then the time until the first conviction is well modeled by a geometric distribution. Approximating the discrete geometric distribution by its continuous counterpart, the exponential distribution, we assume that the time until conviction (independent of whether or not he has desisted) is an exponential random variable C with rate δ, where $\delta^{-1} = 7$ years. In the absence of a first conviction, we assume an exponential lifetime L of a sexual offender, which is roughly consistent with the fact that many convicted offenders self-report that they have a short offending career (22), with rate γ. The 20% self-reported desistance probability in Ref. (22) can be interpreted as $P(L<C) = 0.2$; because $P(L<C) = \gamma/(\gamma+\delta)$, we have $\gamma^{-1} = 4\delta^{-1} = 28$ years.

Let A be the random age (i.e., the time interval between the time of the SA until the time the backlog is being tested) of a SAK in the backlog. The probability distribution of A is derived using the year that each SAK in the Detroit backlog was created (Fig. S1) and assuming that the entire backlog is tested in 2015, which is the publication date of Ref. (9). Now consider a SAK-generating SA that occurs randomly during a sexual offender's active offending career, and suppose that this SAK is in the backlog. Because the potentially averted SAs are only averted if testing the SAK leads to a conviction, we assume in this calculation that the offender has not been convicted before the time that we test the backlog (e.g., when New York City tested its entire SAK backlog, 21.0% of hits did not lead to a conviction because the offender had already been arrested (11)). That is, we let the offender's SA occur at time 0, and assume that $C > A$. If we do not test this SAK (or any future SAKs in the backlog) then this offender will continue to commit SAs for an amount of time equal to

$$\mathbb{E}[\max\{0, \min\{L, C\} - A\} | C > A], \quad (1)$$

where this expectation is with respect to the random variables L, C and A. In Eq. (1), we have used the fact (e.g., Chapter 5 in Ref. (23)) that the forward recurrence time of an exponential random variable has the same distribution as the original exponential random variable (i.e., the time interval between the offender's random SA at time 0 until the time the offender desists is exponential with parameter γ, and similarly for the time until first conviction). The calculation of this conditional expectation (§3 in the SI) gives 3.69 years. Hence, if we test this SAK, we have the potential to avert on average 7.10/yr × 3.69 \year}} = 26.22 SAs. In the Discussion, we use an alternative approach using the Detroit data—rather than the criminal career data in Ref. (20) and Ref. (22)—as the starting point, and obtain a quantity that is of the same order of magnitude. In addition, by setting $A = 0$ in the conditional expectation, we find that a SAK tested at the time of the SA has the potential to avert on average 39.76 SAs.

We now estimate the probability that a hit leads to a conviction. When New York City tested their entire backlog of SAKs, 49 of 1329 CODIS hits led to a conviction (11). We assume that the probability that a hit leads to a conviction is 49/1329 = 0.037, which is an underestimate because there were 44 open cases at the time of reporting (11) (and so the 49 convictions is a lower bound). We assume that the probability that a tested SAK leads to a conviction is our hit probability, 0.316, times 0.037, or 0.012, which implies that the mean number of SAs averted is 0.012 × 26.22 = 0.307. Los Angeles had two convictions of 371 cases (although the authors state that this DNA evidence was of questionable use in the convictions) and a hit probability of 0.178 (7), giving a conviction probability conditioned on a hit of 0.030, which is similar to New York City's conviction probability (conditioned on DNA recovery) of 0.037. Nonetheless, our use of the NYC conviction probability may be very conservative: In Denver, where there was close collaboration between the police department and the district attorney's office, 97 SAK hits from previously cold cases led to 48 convictions (page 38 of (24)).

We use two approaches to compare the $1641 testing cost to the 0.307 SAs averted. By table 1 of Ref. (19), the cost of a SA

equals the victim costs of $138,310 plus the willingness-to-pay (to avoid the SA) cost of $297,102 (the use of this cost component has been endorsed by a panel of experts (25)), which yields a total cost of $435,419 (to be conservative, we do not include the justice costs and offender productivity cost here). Hence, $1641 can be spent to save an expected $0.307 \times \$435{,}419 = \$133{,}484$. The other approach is to compute the marginal cost per averted SA of $1641/0.307 = $5353, which is less than the marginal cost per serious crime averted by lengthening sentences ($7600) or adding police officers (range of $26,300–$62,600) (26), but more than the estimated range of $59–$555 per serious crime averted by adding additional DNA profiles to CODIS (estimated by exploiting state-level legislative changes that expanded state databases) (26).

Incorporating Other Information into The Priority Policy

An investigation in Ref. (9) of other factors that are predictors of improved hit probabilities revealed two: the use of a weapon and the time delay between the assault and the victim's forensic medical examination. In §4 of the SI, we extend our model by considering eight classes of SAKs, characterized by stranger vs. nonstranger SA, weapon use vs. no weapon use, and delay ≤1 day vs. delay ≥2 days. Due to the additional complexity of this model, for simplicity we assume the number of additional crimes X conditioned on $S = s$ follows the Poisson distribution with mean $s\lambda_s + (1 - s)\lambda_{ns}$ instead of the negative binomial distribution as in the original model. Here, λ_s is the mean number of additional crimes for an offender who commits only stranger SAs, and λ_{ns} is the mean number of additional crimes for an offender who commits only nonstranger SAs. As with the parameter pairs $(\lambda_s, \lambda_{ns})$, (b_s, b_{ns}) and (θ_s, θ_{ns}), we allow the specialization level s of an offender to impact the probability of weapon use and the probability of a short delay via the parameter pairs (ω_s, ω_{ns}) and (τ_s, τ_{ns}), respectively. Our resulting parameter estimates suggest that just over two-thirds of stranger SAs and hardly any nonstranger SAs involve weapon use, and >90% of SAs have a delay ≤1 day, with a slightly higher probability for nonstranger SAs (Table S5). The optimal policy ranks the eight classes, with higher priority generally given to weapon use, short delays and stranger SAs (Table S6). The normalized AUC for the priority policy is 0.549, compared to 0.475 for the no-priority policy (Fig. S7); that is, the inclusion of weapon and delay information increases the difference in the normalized AUC between the priority policy and the no-priority policy from 0.045 to 0.074. While the preference for prioritizing stranger SAs and weapon use is due to a higher hit probability conditioned on DNA recovery, the desirability of prioritizing short examination delays is due to a higher probability of DNA recovery; this relationship between DNA recovery probability and the time delay between the assault and the examination was also observed in Los Angeles (7). Table S6 suggests that, of the three factors, stranger vs. nonstranger SAKs generates the biggest improvement in the hit performance, followed by weapon use, with the examination delay having the smallest impact.

Discussion

We make additional observations about the parameter estimates, put forth a six-step argument for testing the entire backlog of SAKs, and discuss the limitations of the data.

Observations about the Parameter Estimates

Fitting our model to the data reveals some new information about SAs—information that is independent of (and, by Ref. (21), perhaps less biased than) the typical approach of self-reported data (22), which is deemed necessary due to the hidden nature of the crime (27). The estimated beta distribution in Fig. 1 exhibits an asymmetry: nearly three times as many offenders commit only nonstranger SAs than commit only stranger SAs, whereas offenders who commit stranger SAs tend to specialize less and perform a mixture of both types. Fig. 1 has some interesting implications if it is accurate (Fig. S3): It confirms the concern raised in Ref. (3) and elsewhere that nonstranger serial rapists may go undetected if nonstranger SAKs are not entered into CODIS, for example, because many law enforcement personnel view nonstranger SAs as less serious. It also confirms the quote "one person's friend is another person's stranger" (3) and is consistent with the observation that over 25% of serial sexual offenders in a sample of previously untested SAKs in Cuyahoga County, Ohio assaulted both strangers and nonstrangers (28).

We also learn from the values of r, p_s and p_{ns} in Table 1 that a stranger offender commits on average almost three times as many additional DNA crimes as a nonstranger offender. Moreover, for a stranger offender, the likelihood is much higher than for a nonstranger offender that these additional crimes are in the backlog (i.e., $b_s > 5b_{ns}$). This is not inconsistent with the finding that 44% of law enforcement agencies in a nationally representative sample indicated that they did not submit SAKs to a crime laboratory if a suspect had not been identified (2). These point estimates imply that an offender who commits only stranger SAs has on average $1 + rp_s/(1 - p_s) = 2.83$ SAKs with recoverable DNA in the backlog, whereas an offender who commits only nonstranger SAs has an average of only $1 + rp_{ns}/(1 - p_{ns}) = 1.88$ SAKs with recoverable DNA in the backlog; among all offenders with at least one SAK in the backlog that has recoverable DNA, the mean number of SAKs with recoverable DNA in the backlog is 2.30. The mean number of DNA crimes in CODIS for an offender is increasing and concave in the specialization parameter s, and the mean number of SAKs in the backlog for an offender is increasing and convex in s (Fig. S8).

Of the 362 Detroit offenders who had associated qualifying offenses in CODIS, 44 offenses were SAs and 318 were nonsexual crimes. This small proportion (44/362 = 0.122) and the fact that the nonsexual crimes in CODIS are often serious crimes (e.g., armed robbery, aggravated assault) that have a cost > $100k (table 1 in Ref. (19)), implies that we are being extremely conservative in our cost-benefit analysis by ignoring the averted costs of nonsexual crimes in CODIS. Combining the proportion 0.122 with the parameter estimates in Table 1 allows us to decompose the probability that an additional crime with recoverable DNA is in the backlog (i.e., b_S and b_{ns}) into the product of two probabilities: the probability that an additional crime is a SA and the probability that an additional SA is in the backlog. (see §2 of the SI and Table 2 for details). The probability that an additional crime with recoverable DNA is a SA is 0.398 for a stranger offender (i.e., an offender who specializes in stranger SAs), 0.145 for a nonstranger offender (i.e., an offender who specializes in nonstranger SAs), and 0.264 when averaged over all offenders with at least one SAK with recoverable DNA in the backlog. The probability that an additional SA with recoverable DNA is in the backlog, which is approximately equal to the

TABLE 2—*Additional quantities derived in §2 of the SI.*

Quantity	Stranger Offenders ($s = 1$)	Nonstranger Offenders ($s = 0$)	Averaged Over Offenders
Pr (additional crime is in backlog)	0.294	0.056	0.161
Pr (additional crime is SA)	0.398	0.145	0.264
Pr (additional SA is in backlog)	0.738	0.388	0.612
Number of past SAs with SAK	3.264	2.437	2.740
Number of past SAs	15.618	13.540	14.359

probability that an additional SA is in the backlog because the DNA recovery probability is quite similar for stranger and nonstranger SAKs, is 0.738 for a stranger offender, 0.388 for a nonstranger offender, and 0.612 when averaged over all offenders with at least one SAK with recoverable DNA in the backlog. The 0.738 vs. 0.388 discrepancy perhaps suggests that Detroit was not often using SAKs as an investigative tool (i.e., not using them to attempt to solve cold stranger SA cases).

The above information also allows us to estimate the mean number of past SAs committed by an offender with at least one SAK with recoverable DNA in the backlog, as follows. We estimate that the mean number of past SAs for which a SAK is generated is 3.264 for a stranger offender, 2.437 for a nonstranger offender, and 2.740 when averaged over all offenders with at least one SAK with recoverable DNA in the backlog (§2 of the SI and Table 2). We do not know what proportion of reported SAs result in a SAK, although it is significantly less than one because victims face considerable hurdles in obtaining a SAK, including geographical access (29) and, until the recent Sexual Assault Survivors' Rights Act, cost (30), and victims sometimes report their SAs too late to undergo a forensic medical examination. The SA reporting probabilities have been estimated to be 0.209 for stranger SAs and 0.180 for nonstranger SAs (31), which may be underestimated for two reasons: These reporting estimates do not consider homeless people, people in institutions or people without phones, and Detroit—due to the sometimes tense relationship between its police and civilians (3)—may have lower reporting rates than those cited in the literature. Dividing the number of past SAs of each type for which a SAK is generated by the reporting probability for that type, and then averaging over all offenders (§2 of the SI and Table 2), we compute that the mean total number of SAs previously committed by an offender with at least one SAK with recoverable DNA in the backlog is 14.36 divided by the unknown proportion of reported SAs that generate a SAK.

Note that the 26.22 SAs potentially averted in the cost-benefit analysis can be thought of as the number of SAs during the forward recurrence time of an offender's offending career, whereas 14.36 (divided by the unknown proportion of reported SAs that generate a SAK) past SAs roughly corresponds to the number of SAs committed during the backward recurrence time of an offender's offending career; these recurrence times are with respect to the random time that the backlog is tested, not the random time of the original SA. For an exponential offending career, the mean backward recurrence time equals the mean forward recurrence time (Chapter 5 in Ref. (23)). However, by construction, the 26.22 estimate assumes that the offender is not incarcerated (because the lifetime is until the first conviction or desistance), whereas the 14.36 estimate does not account for possible incarcerations, which may partially explain the difference between the estimates. Nonetheless, it is reassuring that these two very different approaches give estimates that are comparable to each other.

As noted earlier, we estimate that the expected number of hits if Detroit tested its entire backlog is 2621.8. Even for a large city, the additional workload for investigation, victim notification and advocation, and prosecution would be extremely challenging for a system that may already be working at full capacity (6).

Finally, it is instructive to compare our SAK results to another technology used to solve violent crimes, ballistic imaging. A comparison of the results in Fig. 3 and the ballistic imaging results in Fig. 2 of Ref. (32) reveals two differences. First, the no-priority policy in Ref. (32) is much more convex (normalized AUC is 0.383 using data from Stockton, CA) than the no-priority policy in Fig. 3. Second, under full processing, the SAK hit probability is ≈2.5-fold higher than the ballistic imaging hit probability. There are two primary reasons for these two discrepancies. In ballistic imaging, gun crimes only match to one another, whereas here SAs can match to other SAs or to nonsexual crimes, with the latter having a linear (rather than approximately quadratic, as explained in Ref. (32)) effect. In addition, the model in Ref. (32) does not consider the processing of a one-time backlog, but rather a system that processes all of its evidence over a long period of time.

Although the 0.037 probability that a hit resulted in a conviction in New York City's forklift project may seem low, the corresponding probability for ballistic imaging appears to be even lower (33). Both of these technologies are intended to be used by criminal justice practitioners (e.g., investigators and prosecutors). It is important for researchers to understand the factors underlying the low utilization of hits generated by these technologies and develop ways to make them more useful (see (34) for an example in ballistic imaging).

A Six-step Argument for Testing the Entire Backlog

Our recommendation of testing the entire backlog is based on a six-step argument, which we discuss in turn: (i) in the absence of a sorting cost, prioritizing stranger SAKs offers only a marginal improvement in performance over not using victim–offender relationship data in the prioritization; (ii) the inclusion of a sorting cost reduces this marginal improvement by ≈20%; (iii) the Los Angeles recommendation (7) reinforces observed historical biases in the LAPD and LASD (35); (iv) a cost-benefit analysis of testing the entire backlog (relative to testing none of it) produces quite favorable results; (v) a summary of the probative value of stranger vs. nonstranger SAK hits; and (vi) a cost-benefit analysis of testing nonstranger SAKs (relative to not testing them), which incorporates the differential probative value of stranger SAK hits and nonstranger SAK hits. Steps (i) and (ii) strengthen the seminal results in Ref. (3), which uses hits as the primary performance metric. Step (iii) points out the pitfalls in focusing on downstream performance metrics, such as arrests and convictions. Steps (i)-(iii) make a tentative argument against prioritizing stranger SAKs over nonstranger SAKs, but an argument remains to be made that it is beneficial to test any SAKs, which is important given the limited testing of SAKs in some cities. The cost-benefit analysis in step (iv) shows that the forklift approach is quite cost-effective, even under conservative assumptions, including a restriction of benefits to downstream convictions. Step (v) provides a strong counterargument to testing nonstranger SAKs, and step (vi) uses information from step (v) to provide a lower bound on the benefits from testing nonstranger SAKs.

Step (i)—Because we use their data, it is not surprising that our overall conclusions in steps (i) and (ii) of our argument are

in accord with those in Ref. (3). Nonetheless, our analysis is able to build upon and refine the statistical results in Ref. (3) in several ways. We begin by reviewing the results in Ref. (3). Motivated by the fact that sorting and prioritizing SAKs is timely and costly, Campbell R et al. (3) use a medium effect cutoff (36) in defining (prior to their analysis) equivalency in outcomes between stranger and nonstranger SAKs as having the entire 90% CI for the odds ratio within [0.4,2.50]. They apply this definition to three odds ratios, related to the proportion of SAKs tested that yield a DNA profile that could be entered into CODIS, the proportion of CODIS entries that generate at least one hit, and the proportion of CODIS hits that are associated with a serial sexual offender. The three odds ratios are 1.73, 1.41 and 2.29 with 95% CIs [1.33,2.26], [0.94,2.10] and [1.24,4.25], respectively (Fig. 2 of Ref. (3)). Hence, they conclude that the DNA recovery probability and the CODIS hit probability conditioned on DNA recovery are equivalent between stranger and nonstranger SAKs.

Although the DNA recovery probability and the conditional hit probability (i.e., conditioned on DNA recovery) are worthwhile to estimate in isolation, because the cost of testing the DNA is much larger than the cost of entering the DNA into CODIS, and because a hit requires a successful DNA recovery, these two processes should be considered jointly when making a recommendation about prioritizing between stranger and nonstranger SAKs (6). Consequently, rather than using a statistical definition of equivalence, we explicitly generate curves of the expected number of hits vs. the proportion of backlog processed. In addition to the curves in Fig. 3, we use the normalized AUC as our primary performance metric, and find that the difference in the normalized AUC between the two curves is 0.045 (although a sensitivity analysis gives an estimate of 0.031), which can be loosely interpreted as the stranger-priority policy achieving 4.5% more hits than the no-priority policy on average, if the proportion of the backlog that is processed has a uniform distribution on [0,1]. Note that these curves allow us to extrapolate to any capacity level, and in particular to predict the expected number of hits that would be achieved under a forklift approach.

Step (ii)—Rather than cite the time-consuming nature of sorting the SAKs as a justification for a definition of equivalence (3), we quantify the extra time required to sort the SAKs into stranger vs. nonstranger SAKs and then recompute the normalized AUC for both policies. The difference in the normalized AUC between the two policies decreases by 24.4%, from 0.045 to 0.034. We also note that for cities with backlogs that date back one or two decades, at least some of the records are likely to be in paper form; otherwise, the extra sorting cost would be lower than estimated here. Overall, the argument against prioritizing stranger SAKs appears to be stronger in step (i) than in step (ii).

Step (iii)—Step (iii) of our argument concerns the Los Angeles analysis (7), which calls for focusing only on stranger SAKs. Their recommendation appears to be based on the fact that in Los Angeles, stranger SAKs led to higher rates of arrest, charging and conviction. As noted in Ref. (3), these researchers consider not just the utility of SAK analysis (as measured by hit probability), but the actual utilization of hits in the downstream processes. However, an in-depth collaborative study of the Los Angeles Police Department's and Los Angeles County Sheriff's Department's handling of SAs during 2005–2009 (years that are included in the analysis in Ref. (7)) reveals some troubling observations (35): (a) Some detectives never made arrests in nonstranger cases and instead forwarded the case to the district attorney's office for a pre-arrest filing evaluation; (b) deputy district attorneys based their charging decisions on the probability of conviction, which in turn reflected jurors' preconceived notions of what constitutes rape; (c) when the district attorney declined to file charges, the detective inappropriately used exceptional clearance of the case; (d) prosecutors varied in the extent to which they emphasized DNA as relevant in nonstranger cases; (e) although not required by law, Los Angeles did not file SA charges unless they found corroborating evidence, implying that they did not file charges in "she said/he said" nonstranger SAs (37); and (f) victims found that most detectives were skeptical of the victims' claim (we also note that the recommendations in Ref. (7) emphatically state that the prioritization policy should not be dictated by community pressure from victim groups). This behavior is not specific to Los Angeles and has been documented in many other locations in the U.S.A. (e.g., (4,31)).

At least with the benefit of hindsight (i.e., in light of Ref. (35)), it appears that the recommendation in Ref. (7) of focusing on stranger SAs suffers from circular reasoning: The historical biases on arrests, charges, and convictions of nonstranger assaults listed above led to fewer arrests, charges and convictions from testing nonstranger SAKs. Consequently, they recommend against testing all nonstranger SAKs due to their relatively poor performance on these metrics, even though these historical biases can be mitigated through training and a change in culture. Hence, the recommendation in Ref. (7), if followed, would perpetuate these historical biases. In addition, Los Angeles based their recommendation on the myopic value of successfully prosecuting any given case, and more generally many in law enforcement view SAKs purely as a prosecutorial tool (e.g., SAKs are not submitted for testing because there is no suspect in the case (2)). However, as pointed out by others, hits can have a variety of benefits, even if the victim does not want to participate or is coerced not to (3–6): They can provide a promising investigative lead, corroborate a victim's allegations and a suspect's identity even in the absence of a CODIS hit, produce additional convictions (Federal Law of Evidence 404(b) allows the use of SAK results that are past the statute of limitations if the suspect is on trial for a later SA (38)), discover serial offenders, prevent wrongful prosecution, and solve previous crimes (e.g., a nonstranger SAK hit to the CODIS forensic file).

Step (iv)—Step (iv) of our argument is the cost-benefit analysis performed earlier, which suggests that the forklift approach is very cost-effective (spending $1641 to test a SAK averts $133,484 in future SA-related costs on average). This cost-benefit analysis explicitly assumes that the utilization of hits—in particular, the probability that a hit leads to a conviction—is the same as in New York City's forklift project, and hence incorporates many criticisms that have been aimed at the testing of SAK backlogs; for example, 21.0% of hits led to offenders who were already arrested, 59.6% were beyond the statutes of limitations, and 10.4% of hits were associated with victims who were either missing, unwilling to move forward with the case, or deemed unreliable (11).

Although our cost-benefit analysis is conservative (i.e., overestimates the costs and underestimates the benefits), there is one hidden assumption that may not be conservative: We assume that the detection probability and the activity rate of an offender are independent, whereas there is some evidence that the most

active offenders are also the most difficult to detect (20). On the other hand, in addition to the conservative assumptions mentioned in the description of our analysis, our cost-benefit analysis ignores: (a) the cost of any nonsexual crimes committed before the first conviction or desistance, (b) the deterrent effect of offenders who become aware of a hit to their DNA but who fall short of conviction (it has been shown that offenders placed into CODIS reduce their crime rate (26)), (c) the fact that testing more SAKs will populate the CODIS database, which increases the number of hits to future crimes and may incentivize more SA victims to report their crimes; (d) the retribution and reduced trauma for the victim if they are informed that their assailant is incarcerated; and (e) the exoneration of falsely accused people.

Note that the benefit of testing the backlog decreases with the age of the backlog in two ways: the expected number of SAs potentially averted decreases exponentially with the delay in testing, albeit at the slow rate of $\gamma = 1/28$ per year (equation (23) in the SI), and the likelihood of being beyond the statute of limitations increases (and hence the conviction probability decreases).

As we were revising this paper for publication, we became aware of a recent unpublished study that assesses the costs and benefits of testing the backlog of SAKs in Cuyahoga County, Ohio (39). Their overall conclusion agrees with ours (i.e., testing the backlog is cost-effective), and here we point out differences and similarities between our two analyses, focusing on the three main elements: the testing cost, the probability that a test leads to a conviction, and the cost of averted SAs.

We assume a primary testing cost of \$980, a CODIS entry cost of \$167, a posthit processing cost of \$1334, justice costs of \$8503 and offender productivity costs of \$4610; Singer et al. assume \$949.29 for primary testing, \$882.80 for investigation, and \$372.60 for victim advocacy. While we include more downstream costs than Singer et al., this difference plays a minor role in the calculations because some of these costs are incurred only if there is a conviction, which occurs with a small probability in our analysis.

Cuyahoga County has a higher DNA recovery rate (0.592 vs. 0.491) and a higher hit probability given CODIS entry (0.670 vs. 0.640) than Detroit (Fig. 1 of Singer et al.). Interestingly, particularly given that most of these kits were approximately 20 years old, Cuyahoga County had a projected conviction probability that is much higher than NYC's (11): Their overall probability that a hit leads to a conviction is 0.218, which is 20-fold higher than our probability of 0.012.

The cost per averted SA in Singer et al. (based on McCollister et al. (40)) is approximately half of our estimate (\$203,768 vs. \$435,319). However, the biggest difference between the two analyses is in the expected number of future SAs averted. While we estimate 26.22 SAs averted (and provide an alternative approach that gives the same order of magnitude), Singer et al. use 0.25 based on an estimate that 25% of offenders in the backlog will commit a future reported SA. Taken together, our benefit-to-cost ratio is \$133,484/\$1641 = 81.34 compared to \$48,293,016/\$9,569,008 = 5.05 in Singer et al. We suspect that the true benefit-to-cost ratio is higher than both of these estimates: We are too conservative on the conviction probability and Singer et al. is too conservative on the mean number of averted SAs.

Step (v)—Despite the arguments favoring the forklift approach in steps (i)-(iv), there is a potentially compelling counterargument: Stranger SAKs have higher probative value than nonstranger SAKs (6,12). Recall that CODIS contains offender hits (i.e., identity in CODIS is known) and forensic hits (i.e., identity in CODIS is unknown). Hits in the Detroit data are classified in one of three ways: an offender hit, a forensic hit, and an offender & forensic hit. Because the identity of the forensic hit is already known in the third category (i.e., it has been linked to the corresponding offender hit), we count offender & forensic hits as offender hits in Table 3. Although these data differ from the numbers in figs. 4.13 and 4.14 in Ref. (9) because we include all four testing groups and those figures include only testing groups 1 and 2, the results are qualitatively similar: Recalling that cold hits are generated when stranger SAKs match to the offender file and when nonstranger SAKs match to the forensic file, it is clear that there is a much higher likelihood of cold hits from stranger SAKs, which may increase their probative value (6,12). Nonetheless, a nonstranger SAK hit to the offender file can reveal important investigative and prosecutorial information about the offender that would not have been obtained without testing the SAK.

TABLE 3—*A breakdown of hits from the four testing groups in Ref. (9).*

	Offender Hit	Forensic Hit
Stranger SAK	198 cold hits	13 linked crimes
Nonstranger SAK	196 new offender information	12 cold hits

Step (vi)—Our argument for testing all nonstranger SAKs is thus far not airtight. In addition to performing a cost-benefit analysis of testing all SAKs (i.e., comparing the strategy of no SAK testing vs. the strategy of testing all SAKs, as in step (iv)), it is desirable to perform a cost-benefit analysis of testing all nonstranger SAKs (i.e., comparing the strategy of stranger SAK testing vs. the strategy of testing all SAKs), which we do now.

The calculation of the hit probability derived earlier can be performed separately for nonstranger SAKs and stranger SAKs (§2 of the SI). If all SAKs in the backlog are tested, the estimates in Table 1 predict that the probability that a tested SAK generates a hit is 0.373 for stranger SAKs and 0.269 for nonstranger SAKs. Replacing the overall hit probability 0.316 by the nonstranger hit probability 0.269 and replacing the overall DNA recovery probability 0.491 by the nonstranger DNA recovery probability $d_{ns} = 0.463$, we calculate the mean testing cost of a nonstranger SAK to be 980 + 0.463(167) + 0.269(1334) + 0.269(0.037)13,113 = \$1547. To make a case against testing nonstranger SAKs, the number of SAs averted would need to be much smaller via shorter offender careers, lower offending intensity, higher probability of detection, and/or a lower probability that a hit leads to a conviction. Given that many offenders perform a mix of stranger and nonstranger SAs (Fig. 1), it seems unlikely that the first three of these four factors can be much smaller for nonstranger SAKs than for stranger SAKs (indeed, if anything, results in Ref. (22) suggest the opposite). Therefore, we conservatively assume that these first three factors do not change from the cost-benefit analysis in step (iv), which leaves the fourth factor, where—as stated in step (iii)—the data are contaminated by the historical biases, but nonetheless—as stated in step (iv)—stranger SAK hits may have much higher probative value than nonstranger SAK hits.

The argument now appears to boil down to quantifying the value of a nonstranger SAK hit to the offender file, which

comprises 93.8% of all nonstranger SAK hits (Table 3). It is not clear how to quantify this value without incorporating the biases mentioned in step (iii), and we believe that research along these lines would be valuable. Nonetheless, the cost-benefit analysis in step (iv) suggests that there is considerable slack with regards to the benefit of testing nonstranger SAKs in the backlog. So rather than trying to estimate the value of a nonstranger SAK hit to the offender file, we simply—and very conservatively—assume that they provide no benefit; that is, we assume that only cold hits can lead to convictions.

To compute the probability that a hit is cold, we first compute the proportion of hits that are nonstranger hits if we test the entire backlog, which is $\frac{0.551(0.269)}{0.551(0.269)+0.449(0.373)} = 0.470$. Because 94.2% of stranger hits and 6.2% of nonstranger hits are cold hits (Table 3), it follows that the probability that a hit is cold if we test the entire backlog is 0.470(0.062)+0.530(0.942)=0.528. Recall that in the cost-benefit analysis in step (iv) (i.e., assuming we test the entire backlog), the probability that a tested SAK generates a conviction is the product of the hit probability 0.316 and the probability that a hit generates a conviction, 0.037. To compute the probability that a tested nonstranger SAK generates a conviction, we replace 0.316 by 0.269, and—by assuming that the probability that a cold hit leads to a conviction is the same regardless of whether it is a stranger or nonstranger SAK, we replace 0.037 by (0.062/0.528)0.037. These substitutions yield a probability of 0.001 that a tested nonstranger SAK generates a conviction. Multiplying this probability times 26.22 SAs gives 0.028 averted SAs on average, for a mean averted cost of 0.028 (435,419) = $12,192. In summary, an investment of $1518 to test a nonstranger SAK generates an expected savings of 8.0 times that amount. Alternatively, the mean cost to avert a SA by testing a nonstranger SAK is $1547/0.028 = $54,250, which is within the $26,300–$62,600 range of the cost to avert a serious crime by adding police officers (26).

Finally, we emphasize that the assumptions made in step (vi) are extremely conservative. Indeed, a sample of 203 cases (not including 42 cases that had been indicted but the defendant was capias) from Cuyahoga County's backlog shows that the probability of conviction given a hit is 58/106 = 0.547 for stranger SAs and 38/95 = 0.400 in nonstranger SAs (table 1 of Lovell R et al. (41)). This modest differential, if used instead of assuming that only cold hits could possibly lead to a conviction, would barely put a dent in the cost-effectiveness analysis from step (iv).

In summary, after assuming that there is no benefit from averting nonsexual crimes, no benefit from generating nonstranger SAK hits to the offender file in CODIS and after making a slew of other conservative assumptions, the cost per averted serious crime for testing a nonstranger SAK is comparable to that of adding police officers. We believe, on balance, that our six-step argument goes a considerable way toward providing a convincing argument for testing the entire backlog of SAKs.

Prioritization

As mentioned earlier, the focus of our analysis is on what SAKs to test, not on the order in which they are tested. However, if—as we recommend—all SAKs in the backlog are to be tested, then the next natural question is how and if to prioritize SAKs within the backlog; this is particularly relevant if—due to limited resources—a multiyear effort is required. Steps (i)-(ii) of our argument suggest only a modest improvement in the hit probability, after accounting for the sorting cost. Incorporating weapon and delay information increases the improvement slightly, but perhaps not enough to warrant sorting and prioritization if the records are in paper form. For municipalities that store this information electronically, given the increase in hit probability in Fig. S7 and the differential in probative value implied by Table 3, it seems appropriate to both test the entire backlog and prioritize stranger SAKs, SAKs with weapon use and SAKs with a short examination delay. Another variable that has a high probative value is the statute of limitations. Although beyond the scope of this study (a careful analysis would require not only data on statute of limitations, but also an estimate of the processing rate of SAKs, which would vary across municipalities), prioritizing SAKs that are approaching the statute of limitations may increase the number of convictions.

Data Limitations

More information about CODIS (e.g., the offense types for all hits in the offender and forensic files in CODIS) may have allowed us to assess the impact that legislative changes to CODIS (e.g., expanding or contracting the classes of crimes that are included in CODIS) would have on SAK hit probabilities (although an analysis of this type for all CODIS crimes has been performed (26)). A higher volume of data would allow for a more definitive conclusion regarding the nature of the beta distribution, which dictates the specialization of offenders between stranger and nonstranger SAs. Finally, although it is difficult to assess how an analysis of Detroit data generalizes to other cities, it is reassuring to note that all four NIJ-funded cities have qualitatively similar hit probabilities for stranger vs. nonstranger SAKs (table 13 in (10)) and the predicted hit probability from processing the entire Detroit backlog (0.316) is smaller than the actual hit probability from processing the entire backlog in NYC (1300/3700 = 0.351 in (11)) and Cuyahoga County (0.396 in Singer et al. (39)).

Conclusion

What to do about the immense backlog of untested SAKs in the U.S.A. has been referred to as a national dilemma (5). We use a mathematical model to quantify the impact of prioritizing stranger SAKs over nonstranger SAKs using data from an important recent study in Detroit (3), revisit the Los Angeles study (7), and perform two cost-benefit analyses: one to assess the policy of testing the entire backlog relative to not testing any of it, and one that assumes lower probative value of nonstranger SAK hits to assess the marginal impact of testing nonstranger SAKs. The picture that emerges is that testing SAKs in a backlog can be viewed as a $\approx$ $1600 lottery with a small probability of a huge $\approx$$11.4 M payoff (the payoff is $\approx$$17.3 M if the SAK is tested at the time of the SA because more future SAs can be averted), which is the cost associated with averting 26.22 SAs, where the winning probability is the proportion of SAK tests that turn into a conviction. If the probability of a payoff is greater than 1.4×10^{-4}, then this is an advantageous gamble on average. We conservatively estimate that the probability of a payoff is $\approx 10^{-2}$ when considering all SAKs in the backlog and is $\approx 10^{-3}$ for just the nonstranger SAKs in the backlog. Given the conservative assumptions involved, these analyses lead us to recommend that all SAKs in the backlog should be tested. However, if this forklift approach requires a multiyear effort, it is not unreasonable to prioritize stranger SAKs, SAKs associated with weapon use, and short examination delays (and SAKs within the stature of limitations and SAKs without an identified suspect,

even though we did not explicitly study these two factors). Although we stand by this recommendation, perhaps our most useful contribution is to help frame the debate in a more rigorous manner and offer a starting point for supporting calculations.

Acknowledgments

The authors would like to thank Rebecca Campbell for helpful conversations regarding the Detroit data in Ref. (9). Co-author Can Wang is a PhD student in electrical engineering and is supported by the Graduate School of Business, Stanford University, via tuition and stipend.

References

1. Lovrich NP, Pratt TC, Gaffney MJ, Johnson CJ, Asplen CH, Hurst LH, et al. National DNA study report, final report. Washington, DC: U.S. Department of Justice, 2004.
2. Strom KJ, Hickman MJ. Unanalyzed evidence in law-enforcement agencies: a national examination of forensic processing in police departments. Criminol Public Policy 2010;9:381–404.
3. Campbell R, Pierce SJ, Sharma DB, Feeney H, Fehler-Cabral G. Should rape kit testing be prioritized by victim-offender relationship? Empirical comparison of forensic testing outcomes for stranger and nonstranger sexual assaults. Criminol Public Policy 2016;15:555–83.
4. Campbell R. What really happened? A validation study of rape survivors' help-seeking experiences with the legal and medical systems. Violence Vict 2005;20:55–68.
5. Spohn C. Untested sexual assault kits: a national dilemma. Criminol Public Policy 2016;15:551–4.
6. Wells W. Some considerations when making decisions about prioritizing sexual assault kits for forensic testing. Criminol Public Policy 2016;15:585–92.
7. Peterson J, Johnson D, Herz D, Graziano L, Oehler T. Sexual assault kit backlog study. Washington, DC: National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, 2012.
8. Ritter N. New Orleans sexual assault evidence project: results and recommendations. NIJ Journal 2013;272:13–8.
9. Campbell R, Fehler-Cabral G, Pierce SJ, Sharma DB, Bybee D, Shaw J, et al. The Detroit sexual assault kit (SAK) action research project (ARP), final report. Washington, DC: National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, 2015.
10. Wells W, Campbell B, Franklin C. Unsubmitted sexual assault kits in Houston, TX: case characteristics, forensic testing results, and the investigation of CODIS hits, final report. Washington, DC: National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, 2016.
11. Bashford M. How New York City tackled its backlog. Webinar presented for the National Center for Victims of Crime, March 2013; http://victimsofcrime.org/docs/DNA%20Trainings/brashford-slides.pdf?sfvrsn=0 (accessed December 4, 2017).
12. Strom KJ, Hickman MJ. Untested sexual assault kits: searching for an empirical foundation to guide forensic case processing decisions. Criminol Public Policy 2016;15:593–601.
13. Campbell R, Fehler-Cabral G, Harder and Company Community Research. The Detroit Sexual Assault Kit Action Research Project; 1980-2009 (ICPSR 35632), 2017; http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/35632?searchSource=revise&q =35632 (accessed November 29, 2017).
14. Lloyd SP. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28:129–37.
15. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, FL: Chapman and Hall/CRC Press, 1993.
16. National Institute of Justice. Creating a plan to test a large number of sexual assault kits. Washington, DC: U.S. Department of Justice Report NCJ249234, 2016.
17. The National Center for Victims of Crime. Sexual assault kit testing: what victims need to know, 2017; http://victimsofcrime.org/docs/default-source/dna-resource-center-documents/dna-sak-victim-brofinal.pdf?sfvrsn=2 (accessed April 25, 2017).
18. Roman JK, Reid SE, Chalfin AJ, Knight CR. The DNA field experiment: a randomized trial of the cost-effectiveness of using DNA to solve property crimes. J Exp Criminol 2009;5:345–69.
19. DeLisi M, Kosloski A, Sween M, Hachmeister E, Moore M, Drury A. Murder by numbers: monetary costs imposed by a sample of homicide offenders. J Forens Psychiatry Psychol 2010;21:501–13.
20. Lussier P, Bouchard M, Beauregard E. Patterns of criminal achievement in sexual offending: unravelling the "successful" sex offender. J Crim Justice 2011;39:433–44.
21. Loughran TA, Paternoster R, Thomas KJ. Incentivizing responses to self-report questions in perceptual deterrence studies: an investigation of the validity of deterrence theory using Bayesian truth serum. J Quant Criminol 2014;30:677–707.
22. Lussier P. Criminal career of sex offenders. In: Weisburd D, Bruinsma G, editors. Encyclopedia of criminology and criminal justice. New York, NY: Springer-Verlag, 2014;787–800.
23. Karlin S, Taylor HM. A first course in stochastic processes, 2nd edn. New York, NY: Academic Press, 1975.
24. Davis RC, Jensen C, Kitchens KE. Cold-case investigations: an analysis of current practices and factors associated with successful outcomes. Document No. 237558. Washington, DC: U.S. Department of Justice, 2012.
25. Arrow K, Solow R, Portnoy PR, Leamer EE, Radner R, Schuman H. Report on the NOAA panel on contingent valuation. Fed Reg 1993;15 (58):4601–14.
26. Doleac JL. The effects of DNA databases on crime. Am Econ J: Appl Econ 2017;9:165–201.
27. Lisak D, Miller PM. Repeat rape and multiple offending among undetected rapists. Violence Vict 2002;17:73–84.
28. Lovell R, Luminais M, Flannery DJ, Overman L, Huang D, Walker T, et al. Offending patterns for serial sex offenders identified via the DNA testing of previously unsubmitted sexual assault kits. J Crim Justice 2017;52:68–78.
29. Iritani KM. Sexual assault: Information on training, funding, and the availability of forensic examiners. Washington, DC: Government Accountability Office, 2016; Report GAO-16-334.
30. Wire SD. Two California congresswomen are behind the new Sexual Assault Survivors Bill of Rights. LA Times, September 7, 2016; http://www.latimes.com/politics/la-pol-ca-sexual-assault-congress-20160907-snap-story.html (accessed November 29, 2017).
31. Walfield SM. When a cleared rape is not cleared: a multilevel study of arrest and exceptional clearance. J Interpers Violence 2016;31:1767–92.
32. Wang C, Beggs-Cassin M, Wein LM. Optimizing ballistic imaging operations. J Forensic Sci 2017;62:1188–96.
33. King W, Wells W, Katz C, Maguire E, Frank J. Opening the black box of NIBIN: a descriptive process and outcome evaluation of the use of NIBIN and its effects on criminal investigations. Washington, DC: National Institute of Justice, 2013.
34. King WR, Campbell BA, Matusiak MC, Katz CM. Forensic evidence and criminal investigations: the impact of ballistics information on the investigation of violent crime in nine cities. J Forensic Sci 2017;62:874–80.
35. Spohn C, Tellis K. Policing and prosecuting sexual assault in Los Angeles City and County: a collaborative study in partnership with Los Angeles Police Department, the Los Angeles County Sheriff's Department, and the Los Angeles County District Attorney's Office. Washington, DC: National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, 2012.
36. Rosenthal JA. Qualitative descriptors of strength of association and effect size. J Soc Serv Res 1996;21:37–59.
37. Spohn C, Tellis K. Policing and prosecuting sexual assault: Inside the criminal justice system. Boulder, CO: Lynne Rienner, 2014.
38. Ritter N. Untested evidence in sexual assault cases: using research to guide policy and practice. Sex Assault Rep 2013;16:33–43.
39. Singer M, Lovell R, Flannery D. Cost savings and cost effectiveness of the Cuyahoga County sexual assault kit task force. Cleveland, OH: Begun Center for Violence Prevention Research and Education, Case Western Reserve University, 2016; http://begun.case.edu/wp-content/uploads/2016/06/Cost-Savings-and-Cost-Effectiveness-Brief-1.pdf (accessed November 6, 2017).
40. McCollister KE, French MT, Fang H. The cost of crime to society: new crime-specific estimates for policy and program evaluation. Drug Alcohol Depend 2010;108:98–109.
41. Lovell R, Flannery D, Overman L, Walker T. What happened with the sexual assault reports? Then vs. now. Cleveland, OH: Begun Center for Violence Prevention Research and Education, Case Western Reserve University, 2016; http://begun.case.edu/wp-content/uploads/2016/06/Then-and-Now-Brief.pdf (accessed November 6, 2017).

Additional information and reprint requests:
Lawrence M. Wein, Ph.D.
Graduate School of Business
Stanford University
655 Knnight Way
Stanford, CA 94305
USA
E-mail: lwein@stanford.edu

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix SI Supplemental Information containing details of the analysis.

Figure S1 Histogram of the number of backlogged SAKs created in each year.

Figure S2 Classification of the 1595 SAKs for the eight-class model.

Figure S3 Cluster analysis of the bootstrapped parameters of the beta distribution. (a) The 1000 bootstrapped samples (α, β) clustered into five colored groups, with the maximum likelihood estimate denoted by a black star. The proportion of points in each group (from left to right) is 0.463 (cyan), 0.353 (orange), 0.135 (magenta), 0.040 (red) and 0.009 (blue). (b) The five beta PDFs corresponding to the centers of the five groups from part (a).

Figure S4 The scatter plot of the 1000 bootstrapped values of (a) (d_s, d_{ns}), (b) (p_s, p_{ns}), (c) (b_s, b_{ns}).

Figure S5 The bootstrapped histogram for the normalized AUC of the stranger-priority policy minus the normalized AUC of the no-priority policy.

Figure S6 The expected number of hits vs. the cost after adding the sorting cost of $30 per stranger SAK to the testing cost of $1000/SAK.

Figure S7 For the eight-class model, the expected number of hits vs. the proportion of the backlog processed.

Figure S8 The mean number of additional crimes in CODIS for an offender (i.e., $\lambda(1 - b)$) vs. the offender's specialization parameter s, and (b) the mean number of additional SAKs in the backlog for an offender (i.e., λb) vs. the offender's specialization parameter s.

Table S1 Definition of the eight classes.

Table S2 Class-dependent parameter estimates for the eight-class model.

Table S3 Among offenders who are associated with only one tested SAK in the backlog, the number of offenders of each class who have (i.e., $x_i - z_i \geq 0$) and do not have (i.e., $x_i - z_i = -1$), a DNA Profile in CODIS.

Table S4 Class affiliation data for offenders with multiple tested SAKs in the backlog.

Table S5 Maximum likelihood estimates for the eight-class model.

Table S6 The priority policy in the eight-class model, where rank 1 is the top priority and rank 8 is the bottom priority.

Supporting Information

We estimate the model parameters in §1, undertake a performance analysis in §2, calculate the expected number of SAs potentially averted in §3, and incorporate weapon and delay information in §4. Figs. S3-S8 are discussed in the main text.

1 Parameter Estimation

In this section, we describe the estimation procedure for the various model parameters.

1.1 The DNA Recovery Probabilities, the Composition of the Backlog, and the Testing Probabilities

Of the 1595 tested SAKs in the four testing groups, 127 have missing victim-offender relationship information. Of the remaining 1468 SAKs, 641 are stranger SAKs and 827 are nonstranger SAKs. DNA was recovered from 340 of the 641 stranger SAKs, and 383 of the 827 nonstranger SAKs. Thus, we estimate the DNA recovery probabilities by

$$\hat{d}_s = \frac{340}{641} = 0.530,$$

$$\hat{d}_{ns} = \frac{383}{827} = 0.463.$$

Of the 400 randomly selected SAKs in the pilot project in [1], 247 have data on both the victim-offender relationship and the statute-of-limitations status. Of these 247 SAKs, 111 are stranger SAKs and 136 are nonstranger SAKs. Consequently, we estimate that a SAK in the backlog is a nonstranger SAK with probability $\hat{q}_{ns} = \frac{136}{136+111} = 0.551$ and is a stranger SAK with probability $\hat{q}_s = \frac{111}{136+111} = 0.449$; these values differ slightly from the corresponding values in [1] because they restrict themselves to the 88 samples within the 400 SAKs that have victim-offender relationships and are within the statute of limitations. Because we are allowing the testing of kits in the backlog that are beyond the statute of limitations, we do not restrict to these 88 samples when estimating q_s and q_{ns}.

Of the 11,219 SAKs in the backlog, we consider the sample of $N = 11,219-2512-400 = 8307$ SAKs, where 2512 were originally submitted for testing and 400 were randomly sampled for a pilot project. Because we do not know whether a DNA profile can be recovered from a SAK before it is tested, within this sample of size N, we estimate that there are approximately $\hat{q}_{ns}\hat{d}_{ns}N$ nonstranger SAKs with recoverable DNA profiles and $\hat{q}_s\hat{d}_sN$ stranger SAKs with recoverable DNA profiles. As noted above, within the tested backlog, DNA was recovered from 340 stranger SAKs and 383 nonstranger SAKs. Consequently, we estimate the testing probabilities to be

$$\hat{\theta}_s = \frac{340}{\hat{q}_s\hat{d}_sN} = \frac{641}{(0.449)8307} = 0.172 \tag{1}$$

and

$$\hat{\theta}_{ns} = \frac{383}{\hat{q}_{ns}\hat{d}_{ns}N} = \frac{827}{(0.551)8307} = 0.181. \tag{2}$$

That is, we estimate that 17.2% of stranger SAKs and 18.1% of nonstranger SAKs in the Detroit backlog were tested for the presence of foreign DNA in [2].

1.2 Maximum Likelihood Estimation

We now jointly estimate the remaining eight parameters (α, β, r, p_s, p_{ns}, b_s and b_{ns}) via maximum likelihood estimation.

We begin by describing the data within the framework of the probabilistic model introduced above. Recall that there are 340+383=723 SAKs in the four testing groups in [1] that have a known victim-offender relationship and a recovered DNA profile. These 723 SAKs are affiliated with $a = 692$ unique offenders. For offender $i = 1, \ldots, a$, we let T_i^s and T_i^{ns} be the number of stranger and nonstranger SAKs (among the 723 SAKs) that the offender is associated with. We assume $T_1^s, ..., T_a^s \overset{iid}{\sim} T_s$ and $T_1^{ns}, ..., T_a^{ns} \overset{iid}{\sim} T_{ns}$, where T_s and T_{ns} are random variables defined during the model formulation. Moreover, we assume that $T_1^s, ..., T_a^s$ and $T_1^{ns}, ..., T_a^{ns}$ have corresponding random variables $X_1, ..., X_a$ and $Z_1, ..., Z_a$, which are defined during the model formulation. In our data, we observe the

values of $T_1^s, ..., T_a^s$, $T_1^{ns}, ..., T_a^{ns}$ and $I_{\{X_1 - Z_1 \geq 0\}}, ..., I_{\{X_a - Z_a \geq 0\}}$, which are denoted by $t_1^s, ..., t_a^s$, $t_1^{ns}, ..., t_a^{ns}$ and $I_{\{x_1 - z_1 \geq 0\}}, ..., I_{\{x_a - z_a \geq 0\}}$, where $I_{\{x\}}$ is the indicator function of the event x. Define the sets

$$\mathcal{H} = \{i \in \mathbb{N} | 1 \leq i \leq a, x_i - z_i \geq 0\},$$

and

$$\mathcal{H}^c = \{i \in \mathbb{N} | 1 \leq i \leq a, x_i - z_i = -1\},$$

where $\mathcal{H}$ is the set of offenders with existing CODIS profiles and $\mathcal{H}^c$ is the set of offenders without existing CODIS profiles.

Because we can only observe the offenders with $T_i^s + T_i^{ns} > 0$, we want to maximize the conditional likelihood function

$$\begin{aligned} &\prod_{i \in \mathcal{H}} \mathbb{P}(T_i^s = t_i^s, T_i^{ns} = t_i^{ns}, x_i - z_i \geq 0 | T_i^s + T_i^{ns} > 0) \\ &\times \prod_{i \in \mathcal{H}^c} \mathbb{P}(T_i^s = t_i^s, T_i^{ns} = t_i^{ns}, x_i - z_i = -1 | T_i^s + T_i^{ns} > 0). \end{aligned} \tag{3}$$

Define the quantities $n_{ij0} = \#\{k : t_k^s = i, t_k^{ns} = j, x_k - z_k = -1\}$ and $n_{ij1} = \#\{k : t_k^s = i, t_k^{ns} = j, x_k - y_k \geq 0\}$. For example, n_{010} is the number of offenders associated with only one nonstranger SAK and do not have an existing CODIS profile, and n_{101} is the number of offenders associated with only one stranger SAK and who have an existing CODIS profile. Ignoring the offenders related to SAKs with missing victim-offender relationships in the Detroit data (124 offenders associated with only one SAK with recovered DNA, and two offenders related to two SAKs with recovered DNA), we have $n_{100} = 129$, $n_{010} = 175$, $n_{101} = 172$, $n_{011} = 189$, $n_{200} = 1$, $n_{020} = 3$, $n_{110} = 2$, $n_{201} = 11$, $n_{021} = 2$, $n_{111} = 5$, $n_{301} = 2$ and $n_{121} = 1$; otherwise, $n_{ijk} = 0$.

With these data and taking logs, the log-likelihood function corresponding to equation (3) can be expressed as

$$\sum_{i=0}^{3} \sum_{j=0}^{3} n_{ij0} \log(\mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1 | T_s + T_{ns} > 0))$$

$$
\begin{aligned}
& + \sum_{i=0}^{3}\sum_{j=0}^{3} n_{ij1} \log(\mathbb{P}(T_s = i, T_{ns} = j, X - Z \geq 0 | T_s + T_{ns} > 0)) \\
= \quad & \sum_{i=0}^{3}\sum_{j=0}^{3} n_{ij0} \log(\mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1)) && (4) \\
& + \sum_{i=0}^{3}\sum_{j=0}^{3} n_{ij1} \log(\mathbb{P}(T_s = i, T_{ns} = j, X - Z \geq 0)) && (5) \\
& - \sum_{i=0}^{3}\sum_{j=0}^{3} (n_{ij0} + n_{ij1}) \log(\mathbb{P}(T_s + T_{ns} > 0)), && (6)
\end{aligned}
$$

where the last equation holds because n_{ij1} and n_{ij0} both equal zero for $i+j = 0$ and $i+j > 3$.

In the following three subsections, we derive the probabilities appearing in expressions (4)-(6), which are the probability of not hitting CODIS, the probability of hitting CODIS, and the probability of being observed in the backlog, respectively. After substituting these probabilities and $\hat{\theta}_s$, $\hat{\theta}_{ns}$ from (1)-(2) into (4)-(6), we denote the log-likelihood function by $l(\alpha, \beta, r, p_s, p_{ns}, b_s, b_{ns})$ and derive the maximum likelihood estimates by solving

$$
\begin{aligned}
\max_{\alpha,\beta,r,p_s,p_{ns},b_s,b_{ns}} \quad & l(\alpha, \beta, r, p_s, p_{ns}, b_s, b_{ns}), \\
\text{subject to} \quad & \alpha \geq 0, \\
& \beta \geq 0, \\
& r \geq 0, \\
& 0 \leq p_s \leq 1, \\
& 0 \leq p_{ns} \leq 1, \\
& 0 \leq b_s \leq 1, \\
& 0 \leq b_{ns} \leq 1.
\end{aligned}
$$

1.3 The Probability of not Hitting CODIS

Let $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt$ be the beta function. Conditioning on s, we can

express the probability in (4) as

$$\mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1) = \int_0^1 \mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1|s) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds. \tag{7}$$

The conditional probability in (7) can be written as

$$\begin{aligned}
&\mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1|s) \\
&= \mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1, Z \geq i + j, X \geq i + j - 1|s), \\
&= \sum_{n=i+j-1}^{\infty} \mathbb{P}(X = n|s)\mathbb{P}(Z - 1 = n|X = n, s)\mathbb{P}(T_s = i, T_{ns} = j|Z = 1 + n, s).
\end{aligned} \tag{8}$$

The last of the three conditional probabilities in (8) is given by

$$\begin{aligned}
&\mathbb{P}(T_s = i, T_{ns} = j|Z = 1 + n, s), \\
&= \mathbb{P}(T_s = i, T_{ns} = j, Z_s \geq i, Z_{ns} \geq j|Z = 1 + n, s), \\
&= \sum_{l=i}^{1+n-j} \mathbb{P}(Z_s = l|Z = 1 + n, s)\mathbb{P}(T_s = i, T_{ns} = j|Z_s = l, Z = 1 + n, s), \\
&= \sum_{l=i}^{1+n-j} \binom{1+n}{l} s^l (1-s)^{1+n-l} \binom{l}{i} \theta_s^i (1-\theta_s)^{l-i} \binom{1+n-l}{j} \theta_{ns}^j (1-\theta_{ns})^{1+n-l-j}, \\
&= \sum_{l=i}^{1+n-j} \frac{(1+n-i-j)!}{(l-i)!(1+n-l-j)!} \left(\frac{s(1-\theta_s)}{(1-s)(1-\theta_{ns})}\right)^{l-i} \\
&\quad \times \frac{(1+n)!}{(1+n-i-j)!i!j!} s^i (1-s)^{1+n-i} \theta_s^i \theta_{ns}^j (1-\theta_{ns})^{1+n-i-j}, \\
&= \left(1 + \frac{s(1-\theta_s)}{(1-s)(1-\theta_{ns})}\right)^{1+n-i-j} \frac{(1+n)!}{(1+n-i-j)!i!j!} s^i (1-s)^{1+n-i} \theta_s^i \theta_{ns}^j (1-\theta_{ns})^{1+n-i-j}, \\
&= \frac{(1+n)!}{(1+n-i-j)!} ((1-s)(1-\theta_{ns}) + s(1-\theta_s))^{1+n-i-j} \frac{s^i (1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}.
\end{aligned} \tag{9}$$

Let $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\, dx$ be the gamma function. Substituting (9) into (8) and defining

$$p = sp_s + (1-s)p_{ns},$$

$$b = sb_s + (1-s)b_{ns}$$

and

$$\Theta = (1-s)(1-\theta_{ns}) + s(1-\theta_s), \tag{10}$$

we have

$$
\begin{aligned}
&\mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1|s) \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)n!}(1-p)^r p^n b^n \frac{(1+n)!}{(1+n-i-j)!}\Theta^{1+n-i-j}\frac{s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!} \quad \text{by} \quad (9)-(10), \\
&= \frac{(1-p)^r}{\Gamma(r)} \sum_{n=i+j-1}^{\infty} \Gamma(n+r)\frac{(1+n)(pb\Theta)^{1+n-i-j}}{(1+n-i-j)!}\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \frac{(1-p)^r}{\Gamma(r)} \sum_{n=i+j-1}^{\infty} \Gamma(n+r)\frac{(1+n-i-j)(pb\Theta)^{1+n-i-j}+(i+j)(pb\Theta)^{1+n-i-j}}{(1+n-i-j)!}\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!} \\
&= \frac{(1-p)^r}{\Gamma(r)} \left(\sum_{n=i+j}^{\infty} \Gamma(n+r)\frac{(pb\Theta)^{1+n-i-j}}{(n-i-j)!} + \sum_{n=i+j-1}^{\infty} \Gamma(n+r)\frac{(i+j)(pb\Theta)^{1+n-i-j}}{(1+n-i-j)!}\right)\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \frac{(1-p)^r}{\Gamma(r)} \left(\sum_{k=0}^{\infty} \Gamma(k+r+i+j)\frac{(pb\Theta)^{1+k}}{k!} + \sum_{k=0}^{\infty} \Gamma(k+r+i+j-1)\frac{(i+j)(pb\Theta)^{k}}{k!}\right)\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!} \\
&= \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j)pb\Theta}{(1-pb\Theta)^{r+i+j}} + \frac{\Gamma(r+i+j-1)(i+j)}{(1-pb\Theta)^{r+i+j-1}}\right)\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta+i+j)}{(1-pb\Theta)^{r+i+j}}\right)\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}. \qquad (11
\end{aligned}
$$

Finally, substituting (11) into (7) gives

$$
\begin{aligned}
&\mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1) \\
&= \int_0^1 \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta+i+j)}{(1-pb\Theta)^{r+i+j}}\right)\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}\frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds. \qquad (12)
\end{aligned}
$$

1.4 The Probability of Hitting CODIS

Conditioning on s, we can write the probability in (5) as

$$
\begin{aligned}
&\mathbb{P}(T_s = i, T_{ns} = j, X - Z \geq 0) \\
&= \int_0^1 \mathbb{P}(T_s = i, T_{ns} = j, X - Z \geq 0|s)\frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds, \\
&= \int_0^1 [\mathbb{P}(T_s = i, T_{ns} = j|s) - \mathbb{P}(T_s = i, T_{ns} = j, X - Z = -1|s)]\frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds \qquad (13)
\end{aligned}
$$

The second conditional probability in (13) is given in (11) and the first conditional probability in (13) is derived, after defining

$$B = \frac{b}{1-b}((1-s)(1-\theta_{ns}) + s(1-\theta_s)) = \frac{b}{1-b}\Theta, \qquad (14)$$

as follows:

$$
\begin{aligned}
&\mathbb{P}(T_s = i, T_{ns} = j|s) \\
&= \mathbb{P}(T_s = i, T_{ns} = j, Z \geq i+j, X \geq i+j-1|s) \quad \text{because} \quad Z \geq T_s + T_{ns} \quad \text{and} \quad X \geq Z-1, \\
&= \sum_{n=i+j-1}^{\infty} \mathbb{P}(X = n|s) \sum_{k=i+j-1}^{n} \mathbb{P}(Z-1 = k|X = n, s)\mathbb{P}(T_s = i, T_{ns} = j|Z = 1+k, s), \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)n!}(1-p)^r p^n \sum_{k=i+j-1}^{n} \binom{n}{k} b^k (1-b)^{n-k} \mathbb{P}(T_s = i, T_{ns} = j|Z = 1+k, s), \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)n!}(1-p)^r p^n \sum_{k=i+j-1}^{n} \binom{n}{k} b^k (1-b)^{n-k} \\
&\times \frac{(1+k)!}{(1+k-i-j)!}((1-s)(1-\theta_{ns}) + s(1-\theta_s))^{1+k-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!} \quad \text{by} \quad (9), \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r p^n \sum_{k=i+j-1}^{n} \frac{(1+k)}{(n-k)!(1+k-i-j)!} \left(\frac{b}{1-b}((1-s)(1-\theta_{ns}) + s(1-\theta_s))\right)^k \\
&\quad \times (1-b)^n((1-s)(1-\theta_{ns}) + s(1-\theta_s))^{1-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r p^n \sum_{k=i+j-1}^{n} \frac{(1+k)}{(n-k)!(1+k-i-j)!} B^k \\
&\quad \times (1-b)^n \Theta^{1-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \quad \text{by (10) and (14)}, \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r p^n \frac{d}{dB} \sum_{k=i+j-1}^{n} \frac{B^{k+1}}{(n-k)!(1+k-i-j)!}(1-b)^n \Theta^{1-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r p^n \frac{d}{dB} \sum_{k=i+j-1}^{n} \frac{(n+1-i-j)!B^{k+1-i-j}}{(n-k)!(1+k-i-j)!} \frac{B^{i+j}}{(n+1-i-j)!} \\
&\quad \times (1-b)^n \Theta^{1-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r p^n \frac{(n+1-i-j)(B+1)^{n-i-j}B^{i+j} + (i+j)(B+1)^{n+1-i-j}B^{i+j-1}}{(n+1-i-j)!} \\
&\quad \times (1-b)^n \Theta^{1-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \\
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r p^n \frac{(n+1-i-j)(B+1)^{n-i-j}B^{i+j} + (i+j)(B+1)^{n+1-i-j}B^{i+j-1}}{(n+1-i-j)!} \\
&\quad \times (1-b)^n \Big[\frac{B(1-b)}{b}\Big]^{1-i-j} \frac{s^i(1-s)^j \theta_s^i \theta_{ns}^j}{i!j!} \quad \text{by} \quad (14),
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{n=i+j-1}^{\infty} \frac{\Gamma(n+r)}{\Gamma(r)}(1-p)^r \frac{p^n(B+1)^n(1-b)^n}{(n+1-i-j)!}\left((n+1-i-j)\frac{B}{B+1}+(i+j)\right) \\
&\quad \times \left(\frac{b}{(B+1)(1-b)}\right)^{i+j-1} \frac{s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \left[\left(\sum_{n=i+j}^{\infty}\Gamma(n+r)\frac{p^n(B+1)^n(1-b)^n}{(n-i-j)!}\frac{B}{B+1}\right)+\left(\sum_{n=i+j-1}^{\infty}\Gamma(n+r)\frac{p^n(B+1)^n(1-b)^n}{(n+1-i-j)!}(i+j)\right)\right] \\
&\quad \times \frac{(1-p)^r}{\Gamma(r)}\left(\frac{b}{(B+1)(1-b)}\right)^{i+j-1}\frac{s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \left[\left(\sum_{k=0}^{\infty}\Gamma(k+r+i+j)\frac{(p(1-b)(B+1))^{k+1}}{k!}\frac{B}{B+1}\right)\right. \\
&\quad \left.+\left(\sum_{k=0}^{\infty}\Gamma(k+r+i+j-1)\frac{(p(1-b)(B+1))^k}{k!}(i+j)\right)\right] \\
&\quad \times (p(1-b)(B+1))^{i+j-1}\frac{(1-p)^r}{\Gamma(r)}\left(\frac{b}{(B+1)(1-b)}\right)^{i+j-1}\frac{s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \left[\left(\frac{\Gamma(r+i+j)}{(1-p(1-b)(B+1))^{r+i+j}}p(1-b)B\right)+\left(\frac{\Gamma(r+i+j-1)}{(1-p(1-b)(B+1))^{r+i+j-1}}(i+j)\right)\right] \\
&\quad \times (p(1-b)(B+1))^{i+j-1}\frac{(1-p)^r}{\Gamma(r)}\left(\frac{b}{(B+1)(1-b)}\right)^{i+j-1}\frac{s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \frac{\Gamma(r+i+j-1)}{(1-p(1-b)(B+1))^{r+i+j}}((r-1)p(1-b)B+(i+j)(1-p(1-b))) \\
&\quad \times (p(1-b)(B+1))^{i+j-1}\frac{(1-p)^r}{\Gamma(r)}\left(\frac{b}{(B+1)(1-b)}\right)^{i+j-1}\frac{s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}, \\
&= \frac{(1-p)^r}{\Gamma(r)}\left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta+(i+j)(1-p(1-b)))}{(1-pb\Theta-p(1-b))^{r+i+j}}\right)\frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}.
\end{aligned} \tag{15}
$$

Finally, substituting (11) and (15) into (13) yields

$$
\begin{aligned}
&\mathbb{P}(T_s=i,T_{ns}=j,X-Z\geq 0) \\
&= \int_0^1 \frac{(1-p)^r}{\Gamma(r)}\left[\left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta+(i+j)(1-p(1-b)))}{(1-pb\Theta-p(1-b))^{r+i+j}}\right)\right. \\
&\quad \left.-\left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta+i+j)}{(1-pb\Theta)^{r+i+j}}\right)\right] \\
&\quad \frac{(pb)^{i+j-1}s^i(1-s)^j\theta_s^i\theta_{ns}^j}{i!j!}\frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\,ds.
\end{aligned} \tag{16}
$$

1.5 The Probability of Being Observed in the Backlog

The probability in expression (6) is given by

$$\begin{aligned}
\mathbb{P}(T_s + T_{ns} > 0) &= \int_0^1 \mathbb{P}(T_s + T_{ns} > 0|s) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds, \\
&= \int_0^1 (1 - \mathbb{P}(T_s = 0, T_{ns} = 0|s)) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds, \\
&= \int_0^1 ((1 - \frac{(1-p)^r)\Theta}{(1-pb\Theta - p(1-b))^r} \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds. \qquad (17)
\end{aligned}$$

To avoid numerical instabilities, we use integration by parts to compute the integrals in (12), (16) and (17). For the sake of brevity, this procedure is explained in §4.7 in the context of the more general model considered in §4. Substituting equations (12), (16) and (17) into expressions (4)-(6), respectively, gives the log-likelihood.

1.6 The Parametric Bootstrap

To construct confidence intervals, we use the parametric bootstrap [3] in three stages: bootstrap the pair (d_s, d_{ns}), then the pair (q_s, q_{ns}) and (θ_s, θ_{ns}), and then jointly bootstrap the remaining seven parameters.

To bootstrap the DNA recovery probabilities d_s and d_{ns}, we generate a random variable $N_{nm} \sim Bin(1595, 1468/1595)$, which is the number of SAKs with non-missing victim-offender relationship. Then we generate $N_s \sim Bin(N_{nm}, 641/1468)$ as the number of stranger cases and $N_{ns} = N_{nm} - N_s$ as the number of non-stranger cases. Then we generate $d_s \sim Bin(N_s, \hat{d}_s)/N_s$ and $d_{ns} \sim Bin(N_{ns}, \hat{d}_{ns})/N_{ns}$.

Similarly, to bootstrap the proportion of backlog that are stranger or nonstranger SAKs q_s and q_{ns}, and the testing probabilities θ_s and θ_{ns}, we generate $K_{nm} \sim Bin(250, 247/250)$ as the number of SAKs with a non-missing victim-offender relationship in the 250 samples from the pilot project in [1]. Then we generate $K_{ns} \sim Bin(K_{nm}, 136/247)$ as the number of non-stranger SAKs with a non-missing relationship and set $q_{ns} = K_{ns}/K_{nm}$ and $q_s = 1 - p_{ns}$. Then we set $\theta_s = \frac{N_s}{8307 \times q_s}$ and $\theta_{ns} = \frac{N_{ns}}{8307 \times q_{ns}}$.

We bootstrap the other parameters by generating 692 observable offenders (i.e., $t_s + t_{ns} > 0$) using our model with the maximum likelihood estimates. Each offender i is characterized by $(t_i^s, t_i^{ns}, I_{\{x_i - z_i \geq 0\}})$. Then we calculate the maximum likelihood estimates based on the samples and the bootstrap parameters θ_s and θ_{ns}.

This entire process is repeated 1000 times to generate 1000 sets of parameter values, which allow us to calculate 95% confidence intervals for the 10 parameters (i.e., constructed from the 0.025^{th} and 0.975^{th} fractiles of the empirical distributions of the 1000 points [4] to perform sensitivity analysis of our main results.

1.7 Additional Estimates

In this subsection, we compute several additional quantities from the parameter estimates in Table 1 of the main text. For an offender with at least one SAK in the backlog with recoverable DNA, let us define three associated quantities: CS is the mean number of SAs in CODIS, CNS is the mean number of nonsexual crimes in CODIS, and BS is the mean number of SAKs with recoverable DNA in the backlog. We compute these three unknown quantities as follows. The definitions of r, p_s, p_{ns}, b_s and b_{ns} imply that

$$\frac{\text{BS} - 1}{\text{CS} + \text{CNS} + \text{BS} - 1} = \int_0^1 [sb_s + (1-s)b_{ns}] \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)}\, ds = 0.161, \tag{18}$$

and

$$\text{CS+CNS+BS} = 1 + \int_0^1 \left[\frac{r(1 - 1/(s/(1-p_s) + (1-s)/(1-p_{ns})))}{1/(s/(1-p_s) + (1-s)/(1-p_{ns}))} \right] \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)}\, ds = 2.304. \tag{19}$$

In addition, among the SAKs in the tested backlog that generate hits in CODIS with known qualifying offenses, we find 44 offenders whose associated qualifying offense in CODIS is a SA, and 318 offenders whose associated qualifying offense in CODIS is a nonsexual crime. This yields a third equation,

$$\frac{\text{CS}}{\text{CS} + \text{CNS}} = \frac{44}{44 + 318} = 0.122. \tag{20}$$

Solving (18)-(20) for the three unknowns gives CS=0.085, CNS=0.613 and BS=1.134.

With these values in hand, we can infer that the probability that an additional crime with recovered DNA is a SA is

$$\frac{\text{CS} + \text{BS} - 1}{\text{CS} + \text{CNS} + \text{BS} - 1} = 0.264, \tag{21}$$

and the probability that an additional SA with recovered DNA is in the backlog is

$$\frac{\text{BS} - 1}{\text{CS} + \text{BS} - 1} = 0.612. \tag{22}$$

Moreover, for an offender with at least one SAK in the backlog with recoverable DNA, the mean number of past SAs for which a SAK is generated is

$$(\text{CS} + \text{BS}) \int_0^1 \left(\frac{s}{d_s} + \frac{1-s}{d_{ns}} \right) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)} \, ds = 2.740. \tag{23}$$

Finally, assuming that all reported SAs result in a DNA kit, and using the SA reporting probabilities of 0.209 for stranger SAs and 0.180 for nonstranger SAs [5], we compute the mean number of past SAs for an offender with at least one SAK in the backlog with recoverable DNA to be

$$(\text{CS} + \text{BS}) \int_0^1 \left(\frac{s}{0.209 d_s} + \frac{1-s}{0.180 d_{ns}} \right) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)} \, ds = 14.359. \tag{24}$$

As a rough estimate of how these quantities vary with an offender's specialization level, we consider the two extremes (i.e., $s = 0$ and $s = 1$) and replace the right side of (18) by b_s and b_{ns}, the right side of (19) by $1 + \frac{rp_s}{1-p_s}$ and $1 + \frac{rp_{ns}}{1-p_{ns}}$, and the right side of (20) by the same quantity, but when restricted to offenders who performed only stranger SAs and only nonstranger SAs, which yields 23/156=0.147 and 16/170=0.094. For $s = 1$, the quantities in (21)-(24) reduce to

$$\frac{\text{CS} + \text{BS} - 1}{\text{CS} + \text{CNS} + \text{BS} - 1} = 0.398,$$
$$\frac{\text{BS} - 1}{\text{CS} + \text{BS} - 1} = 0.738,$$
$$\frac{\text{CS} + \text{BS}}{d_s} = 3.264,$$

$$\frac{\text{CS} + \text{BS}}{0.209 d_s} = 15.618.$$

For $s = 0$, the quantities in (21)-(24) reduce to

$$\frac{\text{CS} + \text{BS} - 1}{\text{CS} + \text{CNS} + \text{BS} - 1} = 0.145,$$

$$\frac{\text{BS} - 1}{\text{CS} + \text{BS} - 1} = 0.388,$$

$$\frac{\text{CS} + \text{BS}}{d_{ns}} = 2.437,$$

$$\frac{\text{CS} + \text{BS}}{0.180 d_{ns}} = 13.540.$$

2 Performance Analysis

As noted in the main text, we assume that a SAK generates a hit if its offender has at least one match in CODIS or we tested at least two SAKs associated with the offender. In this section, we compute the expected number of hits for a generic policy (x_s, x_{ns}), which specifies the proportions of stranger and nonstranger SAKs in the backlog that are tested. At the end of this section, we derive (x_s, x_{ns}) values for the stranger-priority policy and the no-priority policy. Mathematically, (x_s, x_{ns}) takes the place of (θ_s, θ_{ns}) in the MLE calculations in §1; i.e., we replace the actual proportions of the backlog tested in [1] by generic values.

Let T_s and T_{ns} be defined as in §1, but for the generic (x_s, x_{ns}) policy. Each of the $T_s + T_{ns}$ SAKs tested under the (x_s, x_{ns}) policy belongs to a certain offender. Because our calculations in §1 apply only to SAKs that generate recoverable DNA, the probability that a tested SAK generates a hit conditioned on it having recoverable DNA, which we denote by $H(x_s, x_{ns})$, is

$$\begin{aligned} H(x_s, x_{ns}) &= \frac{\mathbb{E}[(T_s + T_{ns}) \mathbb{1}_{\{T_s + T_{ns} > 1 \text{ or } X - Z \geq 0\}} | T_s + T_{ns} > 0]}{\mathbb{E}[T_s + T_{ns} | T_s + T_{ns} > 0]}, \\ &= \frac{\mathbb{E}[T_s + T_{ns} | T_s + T_{ns} > 0] - \mathbb{P}(T_s + T_{ns} = 1, X - Z = -1 | T_s + T_{ns} > 0)}{\mathbb{E}[T_s + T_{ns} | T_s + T_{ns} > 0]}, \\ &= \frac{\mathbb{E}[T_s + T_{ns}] - \mathbb{P}(T_s + T_{ns} = 1, X - Z = -1)}{\mathbb{E}[T_s + T_{ns}]}, \end{aligned}$$

$$= \quad 1 - \frac{\mathbb{P}(T_s = 0, T_{ns} = 1, X - Z = -1) + \mathbb{P}(T_s = 1, T_{ns} = 0, X - Z = -1)}{\mathbb{E}[T_s + T_{ns}]} \tag{25}$$

where

$$\begin{aligned}
\mathbb{E}[T_s + T_{ns}] &= \sum_{k=1}^{\infty} k \sum_{i=0}^{k} \mathbb{P}(T_s = i, T_{ns} = k - i), \\
&= \sum_{k=1}^{\infty} k \sum_{i=0}^{k} \int_0^1 \mathbb{P}(T_s = i, T_{ns} = k - i|s) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)}\, ds, \\
&= \int_0^1 \sum_{k=1}^{\infty} k \sum_{i=0}^{k} \mathbb{P}(T_s = i, T_{ns} = k - i|s) \frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha, \beta)}\, ds. \qquad (26)
\end{aligned}$$

By equation (15), we have

$$\begin{aligned}
&\sum_{k=1}^{\infty} k \sum_{i=0}^{k} \mathbb{P}(T_s = i, T_{ns} = k - i|s) \\
&= \sum_{k=1}^{\infty} k \sum_{i=0}^{k} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+k-1)((r-1)pb\Theta + k(1-p(1-b)))}{(1 - pb\Theta - p(1-b))^{r+k}} \right) \frac{(pb)^{k-1} s^i (1-s)^{k-i} \theta_s^i \theta_{ns}^{k-i}}{i!(k-i)!}, \\
&= \sum_{k=1}^{\infty} k \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+k-1)((r-1)pb\Theta + k(1-p(1-b)))}{(1 - pb\Theta - p(1-b))^{r+k}} \right) \frac{(pb)^{k-1} (s\theta_s + (1-s)\theta_{ns})^k}{k!}, \\
&= \sum_{k=1}^{\infty} k \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+k-1)((r-1)pb\Theta + k(1-p(1-b)))}{(1 - pb\Theta - p(1-b))^{r+k}} \right) \frac{(pb)^{k-1} (1-\Theta)^k}{k!} \quad \text{by} \quad (10), \\
&= \frac{(1-p)^r}{\Gamma(r)(1 - pb\Theta - p(1-b))^r} \left[(r-1)\Theta \sum_{k=1}^{\infty} \frac{\Gamma(r+k-1)}{(k-1)!} \left(\frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^k \right. \\
&\quad \left. + \frac{1 - p(1-b)}{pb} \sum_{k=1}^{\infty} k \frac{\Gamma(r+k-1)}{(k-1)!} \left(\frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^k \right], \\
&= \frac{(1-p)^r}{\Gamma(r)(1 - pb\Theta - p(1-b))^r} \left[(r-1)\Theta \sum_{k=0}^{\infty} \frac{\Gamma(r+k)}{k!} \left(\frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^{k+1} \right. \\
&\quad \left. + \frac{1 - p(1-b)}{pb} \sum_{k=0}^{\infty} (k+1) \frac{\Gamma(r+k)}{k!} \left(\frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^{k+1} \right], \\
&= \frac{(1-p)^r pb(1-\Theta)}{\Gamma(r)(1 - pb\Theta - p(1-b))^{r+1}} \left[\left((r-1)\Theta + \frac{1 - p(1-b)}{pb} \right) \sum_{k=0}^{\infty} \frac{\Gamma(r+k)}{k!} \left(\frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^k \right. \\
&\quad \left. + \frac{1 - p(1-b)}{pb} \sum_{k=0}^{\infty} k \frac{\Gamma(r+k)}{k!} \left(\frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^k \right], \\
&= \frac{(1-p)^r pb(1-\Theta)}{\Gamma(r)(1 - pb\Theta - p(1-b))^{r+1}} \left[\left((r-1)\Theta + \frac{1 - p(1-b)}{pb} \right) \frac{\Gamma(r)}{\left(1 - \frac{pb(1-\Theta)}{1 - pb\Theta - p(1-b)} \right)^r} \right.
\end{aligned}$$

$$+\frac{1-p(1-b)}{pb}\frac{\Gamma(r)}{\left(1-\frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}\right)^r}\frac{r\frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}}{1-\frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}}\Bigg],$$

$$=\frac{pb(1-\Theta)}{(1-pb\Theta-p(1-b))}\left[(r-1)\Theta+\frac{1-p(1-b)}{pb}+\frac{r(1-p(1-b))(1-\Theta)}{1-p}\right]. \quad (27)$$

Substituting equation (27) into (26) gives us

$$\mathbb{E}[T_s+T_{ns}]$$

$$=\int_0^1\frac{pb(1-\Theta)}{(1-pb\Theta-p(1-b))}\left[(r-1)\Theta+\frac{1-p(1-b)}{pb}+\frac{r(1-p(1-b))(1-\Theta)}{1-p}\right]\frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha,\beta)}\,ds. \quad (28)$$

Equation (12) implies that

$$\mathbb{P}(T_s=0,T_{ns}=1,X-Z=-1) = \int_0^1\frac{(1-p)^r((r-1)pb\Theta+1)}{(1-pb\Theta)^{r+1}}x_{ns}\frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha,\beta)}\,ds \quad (29)$$

$$\mathbb{P}(T_s=1,T_{ns}=0,X-Z=-1) = \int_0^1\frac{(1-p)^r((r-1)pb\Theta+1)}{(1-pb\Theta)^{r+1}}x_{s}\frac{s^{\alpha}(1-s)^{\beta-1}}{B(\alpha,\beta)}\,ds. \quad (30)$$

Substituting equations (28)-(30) into (25) gives the hit probability conditioned on DNA recovery, $H(x_s,x_{ns})$.

When we test a proportion x_s of all stranger SAKs and a proportion x_{ns} of all non-stranger SAKs in the backlog, the number of SAKs with recoverable DNA is roughly $N\hat{q}_s\hat{d}_s x_s + N\hat{q}_{ns}\hat{d}_{ns}x_{ns}$, where $N,\hat{q}_s,\hat{q}_{ns},\hat{d}_s$ and $\hat{d}_{ns}$ were defined earlier in the estimation of the DNA recovery probabilities and testing probabilities. Hence, the expected number of hits under the (x_s,x_{ns}) policy is

$$(N\hat{q}_s\hat{d}_s x_s + N\hat{q}_{ns}\hat{d}_{ns}x_{ns})H(x_s,x_{ns}). \quad (31)$$

Let $\rho\in[0,1]$ be the proportion of the backlog that is tested; i.e., it is the horizontal axis in Fig. 2 in the main text. Then $x_s=x_{ns}=\rho$ for the no-priority policy, and $x_s=\min\{\rho/\hat{q}_s,1\}$ and $x_{ns}=\max\{(\rho-\hat{q}_s)/\hat{q}_{ns},0\}$ for the stranger-priority policy. Substituting these expressions into (31) generates the tradeoff curves in Fig. 3 in the main text.

We consider the hit probability of stranger SAKs and nonstranger SAKs separately in the last step of the six-step argument for testing the entire backlog in the main text. Here

we calculate the two hit probabilities. For the generic (x_s, x_{ns}) policy, the probability that a tested stranger SAK generates a hit conditioned on it having recoverable DNA, which we denote by $H_s(x_s, x_{ns})$, is

$$\begin{aligned} H_s(x_s, x_{ns}) &= \frac{\mathbb{E}[T_s \mathbb{1}_{\{T_s+T_{ns}>1 \text{ or } X-Z\geq 0\}} | T_s > 0]}{\mathbb{E}[T_s | T_s > 0]}, \\ &= \frac{\mathbb{E}[T_s \mathbb{1}_{\{T_s+T_{ns}>1 \text{ or } X-Z\geq 0\}}]}{\mathbb{E}[T_s]}, \\ &= \frac{\mathbb{E}[T_s] - \mathbb{P}(T_s = 1, T_{ns} = 0, X - Z = -1)}{\mathbb{E}[T_s]}, \\ &= 1 - \frac{\mathbb{P}(T_s = 1, T_{ns} = 0, X - Z = -1)}{\mathbb{E}[T_s]}, \end{aligned} \tag{32}$$

where

$$\begin{aligned} \mathbb{E}[T_s] &= \sum_{i=1}^{\infty} i \sum_{j=0}^{\infty} \mathbb{P}(T_s = i, T_{ns} = j), \\ &= \sum_{i=1}^{\infty} i \sum_{j=0}^{\infty} \int_0^1 \mathbb{P}(T_s = i, T_{ns} = j | s) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha, \beta)} \, ds, \\ &= \int_0^1 \sum_{i=1}^{\infty} i \sum_{j=0}^{\infty} \mathbb{P}(T_s = i, T_{ns} = j | s) \frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha, \beta)} \, ds. \end{aligned} \tag{33}$$

By equation (15), we have

$$\begin{aligned} &\sum_{i=1}^{\infty} i \sum_{j=0}^{\infty} \mathbb{P}(T_s = i, T_{ns} = j | s) \\ &= \sum_{i=1}^{\infty} i \sum_{j=0}^{\infty} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta + (i+j)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+i+j}} \right) \frac{(pb)^{i+j-1} s^i (1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \\ &= \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j-1)((r-1)pb\Theta + (i+j)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+i+j}} \right) \frac{(pb)^{i+j-1} s^i (1-s)^j \theta_s^i \theta_{ns}^j}{(i-1)!j!}, \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j)((r-1)pb\Theta + (i+j+1)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+i+j+1}} \right) \frac{(pb)^{i+j} s^{i+1} (1-s)^j \theta_s^{i+1} \theta_{ns}^j}{i!j!}, \\ &= s\theta_s \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+i+j)((r-1)pb\Theta + (i+j+1)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+i+j+1}} \right) \frac{(pb)^{i+j} s^i (1-s)^j \theta_s^i \theta_{ns}^j}{i!j!}, \\ &= s\theta_s \sum_{k=0}^{\infty} \sum_{i=0}^{k} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+k)((r-1)pb\Theta + (k+1)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+k+1}} \right) \frac{(pb)^k s^i (1-s)^{k-i} \theta_s^i \theta_{ns}^{k-i}}{i!(k-i)!}, \\ &= s\theta_s \sum_{k=0}^{\infty} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+k)((r-1)pb\Theta + (k+1)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+k+1}} \right) \frac{(pb)^k (s\theta_s + (1-s)\theta_{ns})^k}{k!}, \end{aligned}$$

$$= s\theta_s \sum_{k=0}^{\infty} \frac{(1-p)^r}{\Gamma(r)} \left(\frac{\Gamma(r+k)((r-1)pb\Theta + (k+1)(1-p(1-b)))}{(1-pb\Theta - p(1-b))^{r+k+1}} \right) \frac{(pb)^k(1-\Theta)^k}{k!} \text{ by (10)},$$

$$= s\theta_s \frac{(1-p)^r}{\Gamma(r)(1-pb\Theta-p(1-b))^{r+1}} \left[((r-1)pb\Theta + 1 - p(1-b)) \sum_{k=0}^{\infty} \frac{\Gamma(r+k)}{k!} \left(\frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)} \right)^k \right.$$
$$\left. + (1-p(1-b)) \sum_{k=0}^{\infty} k \frac{\Gamma(r+k)}{k!} \left(\frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)} \right)^k \right],$$

$$= s\theta_s \frac{(1-p)^r}{\Gamma(r)(1-pb\Theta-p(1-b))^{r+1}} \left[((r-1)pb\Theta + 1 - p(1-b)) \frac{\Gamma(r)}{\left(1 - \frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}\right)^r} \right.$$
$$\left. + (1-p(1-b)) \frac{\Gamma(r)}{\left(1 - \frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}\right)^r} \frac{r\frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}}{1 - \frac{pb(1-\Theta)}{1-pb\Theta-p(1-b)}} \right],$$

$$= \frac{pbs\theta_s}{(1-pb\Theta-p(1-b))} \left[(r-1)\Theta + \frac{1-p(1-b)}{pb} + \frac{r(1-p(1-b))(1-\Theta)}{1-p} \right]. \tag{34}$$

Substituting equation (34) into (33) gives us

$$\mathbb{E}[T_s]$$
$$= \int_0^1 \frac{s\theta_s pb}{(1-pb\Theta-p(1-b))} \left[(r-1)\Theta + \frac{1-p(1-b)}{pb} + \frac{r(1-p(1-b))(1-\Theta)}{1-p} \right] \frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha,\beta)}\, ds. \tag{35}$$

Substituting equations (30) and (35) into (32) gives the hit probability of stranger SAKs conditioned on DNA recovery, $H_s(x_s, x_{ns})$.

Similarly, the hit probability of nonstranger SAKs conditioned on DNA recovery is

$$\begin{aligned} H_{ns}(x_s, x_{ns}) &= \frac{\mathbb{E}[T_{ns}\mathbb{1}_{\{T_s+T_{ns}>1 \text{ or } X-Z\geq 0\}}|T_{ns}>0]}{\mathbb{E}[T_{ns}|T_{ns}>0]}, \\ &= \frac{\mathbb{E}[T_{ns}\mathbb{1}_{\{T_s+T_{ns}>1 \text{ or } X-Z\geq 0\}}]}{\mathbb{E}[T_{ns}]}, \\ &= \frac{\mathbb{E}[T_{ns}] - \mathbb{P}(T_s=0, T_{ns}=1, X-Z=-1)}{\mathbb{E}[T_{ns}]}, \\ &= 1 - \frac{\mathbb{P}(T_s=0, T_{ns}=1, X-Z=-1)}{\mathbb{E}[T_{ns}]}, \end{aligned} \tag{36}$$

where

$$\mathbb{E}[T_{ns}]$$

$$= \mathbb{E}[T_s + T_{ns}] - \mathbb{E}[T_s]$$
$$= \int_0^1 \frac{pb(1-s)\theta_{ns}}{(1-pb\Theta - p(1-b))}\left[(r-1)\Theta + \frac{1-p(1-b)}{pb} + \frac{r(1-p(1-b))(1-\Theta)}{1-p}\right]\frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha,\beta)}\, ds. \tag{37}$$

3 The Expected Number of SAs Potentially Averted

The calculation of the expected number of SAs potentially averted involves three random variables. The random variable L, which is the time until an offender desists, is exponential with parameter γ and PDF $g(l) = \gamma e^{-\gamma l}$. The random variable C, which is the time until an offender's first conviction, is exponential with parameter δ and PDF $h(c) = \delta e^{-\delta c}$. The random variable A, which is the age of a SAK in the backlog at the time of testing, is derived empirically. For notational convenience, we assume it has PDF $f(a)$, but we actually use a probability mass function in our calculations, which is derived by letting $A = 2015 - \tau$, where τ is the year in which the SAK was created, and using Fig. S1, which displays the histogram of the number of SAKs in the Detroit backlog that were created in each year.

The goal of this section is to derive the expected number of SAs potentially averted, which is given by

$$\mathbb{E}[\max\{0, \min\{L, C\} - A\} | C > A] = \frac{\mathbb{E}[\max\{0, \min\{L, C\} - A\} I_{C>A}]}{\mathbb{P}(C > A)}, \tag{38}$$

where $I_{\{x\}}$ is the indicator function of the event x.

The denominator in (38) is

$$\begin{aligned} \mathbb{P}(C > A) &= \int_0^\infty \mathbb{P}(C > A | A = a) f(a)\, da, \\ &= \int_0^\infty e^{-\delta a} f(a)\, da. \end{aligned} \tag{39}$$

The numerator in (38) is

$$\mathbb{E}[\max\{0, \min\{L, C\} - A\} I_{C>A}]$$

$$
\begin{aligned}
&= \int_{a=0}^{\infty}\int_{l=0}^{\infty}\int_{c=0}^{\infty} \max\{0,\min\{l,c\}-a\} I_{c>a} f(a)g(l)h(c)\ da\ dl\ dc,\\
&= \int_{a=0}^{\infty}\int_{l=0}^{\infty}\int_{c=a}^{\infty} \max\{0,\min\{l,c\}-a\} f(a)g(l)h(c)\ da\ dl\ dc,\\
&= \int_{a=0}^{\infty}\int_{l=a}^{\infty}\int_{c=a}^{\infty} (\min\{l,c\}-a) f(a)g(l)h(c)\ da\ dl\ dc,\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty}\int_{c=a}^{\infty} \min\{l,c\} g(l)h(c)\ dl\ dc - a\int_{l=a}^{\infty} g(l)\ dl \int_{c=a}^{\infty} h(c)\ dc \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty}\int_{c=a}^{\infty} \min\{l,c\} g(l)h(c)\ dl\ dc - ae^{-\gamma a}e^{-\delta a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty} g(l)\ dl \left(\int_{c=a}^{l} ch(c)\ dc + \int_{c=l}^{\infty} lh(c)\ dc \right) - ae^{-(\gamma+\delta)a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty} g(l)\ dl \left(\int_{c=a}^{l} c\delta e^{-\delta c}\ dc + le^{-\delta l} \right) - ae^{-(\gamma+\delta)a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty} g(l)\ dl \left(-\int_{c=a}^{l} c\ de^{-\delta c} + le^{-\delta l} \right) - ae^{-(\gamma+\delta)a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty} g(l)\ dl \left(-ce^{-\delta c}|_{c=a}^{l} + \int_{c=a}^{l} e^{-\delta c}\ dc + le^{-\delta l} \right) - ae^{-(\gamma+\delta)a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty} g(l)\ dl \left(-le^{-\delta l} + ae^{-\delta a} - \frac{e^{-\delta l}-e^{-\delta a}}{\delta} + le^{-\delta l} \right) - ae^{-(\gamma+\delta)a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(\int_{l=a}^{\infty} g(l)\ dl \left(ae^{-\delta a} + \frac{e^{-\delta a}}{\delta} \right) - \int_{l=a}^{\infty} \gamma e^{-\gamma l}\frac{e^{-\delta l}}{\delta}\ dl - ae^{-(\gamma+\delta)a} \right),\\
&= \int_{a=0}^{\infty} f(a)\ da \left(ae^{-(\gamma+\delta)a} + \frac{e^{-(\gamma+\delta)a}}{\delta} - \frac{\gamma}{\delta(\gamma+\delta)}e^{-(\gamma+\delta)a} - ae^{-(\gamma+\delta)a} \right),\\
&= \frac{1}{\gamma+\delta}\int_{0}^{\infty} e^{-(\gamma+\delta)a} f(a)\ da. \qquad (40)
\end{aligned}
$$

Substituting (39)-(40) into (38) yields

$$
\mathbb{E}[\max\{0,\min\{L,C\}-A\}|C>A] = \frac{\int_0^\infty e^{-(\gamma+\delta)a} f(a)\ da}{(\gamma+\delta)\int_0^\infty e^{-\delta a} f(a)\ da}. \qquad (41)
$$

Note that if we delay backlog testing by t years, so that the backlog age is the random variable $\tilde{A}$ with values $\tilde{a} = a + t$, then

$$
\begin{aligned}
\mathbb{E}[\max\{0,\min\{L,C\}-\tilde{A}\}|C>\tilde{A}] &= \frac{\int_t^\infty e^{-(\gamma+\delta)\tilde{a}} f(\tilde{a}-t)\ d\tilde{a}}{(\gamma+\delta)\int_t^\infty e^{-\delta\tilde{a}} f(\tilde{a}-t)\ d\tilde{a}},\\
&= \frac{\int_t^\infty e^{-(\gamma+\delta)(\tilde{a}-t)} e^{-(\gamma+\delta)t} f(\tilde{a}-t)\ d(\tilde{a}-t)}{(\gamma+\delta)\int_t^\infty e^{-\delta(\tilde{a}-t)} e^{-\delta t} f(\tilde{a}-t)\ d(\tilde{a}-t)},\\
&= \frac{e^{-(\gamma+\delta)t}}{e^{-\delta t}} \frac{\int_t^\infty e^{-(\gamma+\delta)(\tilde{a}-t)} f(\tilde{a}-t)\ d(\tilde{a}-t)}{(\gamma+\delta)\int_t^\infty e^{-\delta(\tilde{a}-t)} f(\tilde{a}-t)\ d(\tilde{a}-t)},
\end{aligned}
$$

$$
\begin{aligned}
&= e^{-\gamma t} \frac{\int_0^\infty e^{-(\gamma+\delta)a} f(a)\ da}{(\gamma+\delta)\int_0^\infty e^{-\delta a} f(a)\ da}, \\
&= e^{-\gamma t} \mathbb{E}[\max\{0, \min\{L, C\} - A\} | C > A]. \qquad (42)
\end{aligned}
$$

That is, the expected number of SAs potentially averted decreases exponentially with the testing delay at rate γ.

Finally, substituting $\gamma^{-1} = 28$ years, $\delta^{-1} = 7$ years, and the empirical probability mass function $\mathbb{P}(A = a)$ for the PDF $f(a)$ into (41) allows us to calculate

$$
\mathbb{E}[\max\{0, \min\{L, C\} - A\} | C > A] = \frac{\sum_{a=0}^\infty e^{-(\gamma+\delta)a} \mathbb{P}(A = a)}{(\gamma+\delta)\sum_{a=0}^\infty e^{-\delta a} \mathbb{P}(A = a)} = 3.693 \text{ years.}
$$

4 Incorporating Weapon and Delay Information

In this section, we generalize the analysis to incorporate information about weapon use and the time delay between the assault and the victim's exam. For simplicity, we use the Poisson distribution instead of the negative binomial distribution to model the number of additional crimes by an offender.

4.1 The Model

We group the SAKs into eight classes according to Table S1, and assume that each SAK belongs to class i with probability q_i for $i = 1, \ldots, 8$. We assume that the DNA recovery probability for SAKs in class i is d_i, and define $\boldsymbol{q} = [q_1, \ldots, q_8]^T$ and $\boldsymbol{d} = [d_1, \ldots, d_8]^T$. As in our original model, we assume that each defender has a random specialization level S taken from a beta distribution with parameters (α, β), the number of additional crimes with recoverable DNA related to an offender in either CODIS or the backlog is X, which is a Poisson random variable with mean $s\lambda_s + (1-s)\lambda_{ns}$ after conditioning on $S = s$, and the total number of SAKs with recoverable DNA in the backlog for an offender is Z, which is equal to one plus a binomial random variable with parameters x and $sb_s + (1-s)b_{ns}$, after conditioning on $X = x$ and $S = s$. In a similar manner, we introduce the parameter pairs

(w_s, w_{ns}) and (τ_s, τ_{ns}), and after conditioning on s, assume that the probability of weapon use is

$$w = w_s + (1 - s)w_{ns},$$

and the probability that the delay is ≤ 1 day is

$$\tau = s\tau_s + (1 - s)\tau_{ns}.$$

We define $\boldsymbol{Z} = [Z_1, \ldots, Z_8]^T$, where Z_i equals the number of SAKs with recoverable DNA of class i in the backlog related to an offender; hence, $Z = \sum_{i=1}^{8} Z_i = \mathbf{1}^T \boldsymbol{Z}$, where $\mathbf{1}$ is an eight-dimensional column vector of ones. We assume that $\mathbf{Z}$ conditioned on $Z = z$ and $S = s$ is multinomial with parameters z and $\boldsymbol{m}$, where

$$\boldsymbol{m} = \begin{bmatrix} sw\tau \\ sw(1-\tau) \\ s(1-w)\tau \\ s(1-w)(1-\tau) \\ (1-s)w\tau \\ (1-s)w(1-\tau) \\ (1-s)(1-w)\tau \\ (1-s)(1-w)(1-\tau) \end{bmatrix}.$$

Finally, conditioned on $Z_i = z_i$ and $S = s$, we assume that the number of SAKs with recoverable DNA of class i in the backlog that are tested is T_i, which is binomial with parameters z_i and θ_i.

4.2 The Class Probabilities, the DNA Recovery Probabilities and the Testing Probabilities

For ease of reference, we summarize the information available for the 1595 SAKs in testing groups 1, 2, 3 and 4 in Fig. S2. We ignore the 356 SAKs in Fig. S2 that have missing data for any of the three criteria. Among the remaining 1239 SAKs, the number of SAKs in each of the eight classes is $(n_1, \ldots, n_8) = (224, 23, 277, 35, 117, 12, 473, 78)$. Based on the 250 samples from the 400 Project, where 111 are stranger SAKs, 136 are non-stranger SAKs

and three have missing relationship data, we require that $\sum_{i=1}^{4} \hat{q}_i = \frac{111}{136+111} = 0.449$ and $\sum_{i=5}^{8} \hat{q}_i = \frac{136}{136+111} = 0.551$. Then we estimate $\hat{q}_i = 0.449 \times \frac{n_i}{\sum_{j=1}^{4} n_j}$ for $i = 1, 2, 3, 4$ and $\hat{q}_i = 0.551 \times \frac{n_i}{\sum_{j=5}^{8} n_j}$ for $i = 5, 6, 7, 8$, which yields the estimates in Table S2.

To estimate the class-dependent DNA recovery probabilities, we count the number of SAKs with recoverable DNA profiles in each class i, which gives $(m_1, \ldots, m_8) = (124, 10, 144, 11, 57, 6, 234, 16)$. We estimate the DNA recovery probabilities via $\hat{d}_i = m_i/n_i$, which yields the estimates in Table S2.

From page 150 of the Detroit report [1], we know that when they drew samples from the backlogs, they didn't want to draw from the 400 kits tested in the 400 Project and the SAKs with police department crime laboratory ID numbers. Thus we set the number of backlogs they sample from to be 8707-400=8307.

Of the $N = 8307$ SAKs, there are approximately $\hat{q}_i \hat{d}_i N$ SAKs in group i with recoverable DNA profiles (ignoring the fact that some SAKs are missing at least one of the three attributes). We estimate

$$\hat{\theta}_i = \frac{m_i}{\hat{q}_i \hat{d}_i N} = \frac{n_i}{\hat{q}_i N} = \begin{cases} \frac{\sum_{j=1}^{4} n_j}{0.449N} & i = 1, 2, 3, 4; \\ \frac{\sum_{j=5}^{8} n_j}{0.551N} & i = 5, 6, 7, 8, \end{cases} \tag{43}$$

which are given in Table S2. The first four groups and the last four groups in (43) have the same testing probabilities because we have no extra information about the proportion of SAKs that involve weapon use or have a short (i.e., ≤ 1 day) delay in the population.

4.3 Maximum Likelihood Estimation

In this subsection, we use the Detroit Data to calculate maximum likelihood estimates for α, β, λ_s, λ_{ns}, b_s, b_{ns}, w_s, w_{ns}, τ_s and τ_{ns}.

The $\sum_{i=1}^{8} m_i = 602$ SAKs with known criteria and recovered DNA are affiliated with $a = 576$ unique offenders, and the number of tested SAKs with recoverable DNA in the backlog belonging to the eight classes for each of these offenders is denoted by $\boldsymbol{T}_1, ..., \boldsymbol{T}_a$.

We assume $\boldsymbol{T}_1, ..., \boldsymbol{T}_a \overset{iid}{\sim} \boldsymbol{T}$ and have corresponding $X_1, ..., X_a$, $Z_1, ..., Z_a$, $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_a$. In our data, we observe the values of $\boldsymbol{T}_1, ..., \boldsymbol{T}_a$ and $I_{X_1-Z_1\geq 0}, \ldots, I_{X_a-Z_a\geq 0}$, which are denoted by $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_a$ and $I_{x_1-z_1\geq 0}, \ldots, I_{x_a-z_a\geq 0}$. Define the sets

$$\mathcal{H} = \{i \in \mathbb{N} | 1 \leq i \leq a, x_i - z_i \geq 0\}$$

and

$$\mathcal{H}^c = \{i \in \mathbb{N} | 1 \leq i \leq a, x_i - z_i = -1\},$$

which include the offenders with and without, respectively, an existing DNA profile in CODIS.

Because we can only observe the offenders with $\mathbf{1}^T\boldsymbol{T_i} > 0$, we want to maximize the conditional likelihood function

$$\prod_{i\in\mathcal{H}} \mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_i, x_i - z_i \geq 0 | \mathbf{1}^T\boldsymbol{T}_i > 0) \times \prod_{i\in\mathcal{H}^c} \mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_i, x_i - z_i = -1 | \mathbf{1}^T\boldsymbol{T}_i > 0). \qquad (44)$$

Data for offenders associated with only one tested SAK with recoverable DNA in the backlog appears in Table S3. Denote the first row in Table S3 (i.e., $x_i - z_i \geq 0$) by an eight-dimensional vector $\boldsymbol{m_{11}}$ and the second row ($x_i - z_i = -1$) by $\boldsymbol{m_{10}}$. We have $\#\{i : \boldsymbol{T}_i = \boldsymbol{e_j}, x_i - z_i \geq 0\} = \boldsymbol{e_j}^T\boldsymbol{m_{11}}$ and $\#\{i : \boldsymbol{T}_i = \boldsymbol{e_j}, x_i - z_i = -1\} = \boldsymbol{e_j}^T\boldsymbol{m_{10}}$, where $\boldsymbol{e_j}$ is an eight-dimensional column vector with the j^{th} entry being 1 and the other entries being 0.

Data for offenders associated with two or three tested backlogs with known criteria and recoverable DNA appear in Table S4. We define $n_{i0} = \#\{k : \boldsymbol{t}_k = \boldsymbol{e_i}, x_k - z_k = -1\}$ and $n_{i1} = \#\{k : \boldsymbol{t}_k = \boldsymbol{e_i}, x_k - z_k \geq 0\}$. For example, n_{10} is the number of offenders associated with only one tested Detroit SAK in class 1 that does not have an existing CODIS profile, and n_{11} is the number of offenders associated with only one tested Detroit SAK in class 1 that has an existing CODIS profile. We have $n_{i0} = \boldsymbol{e_i}^T\boldsymbol{m_{10}}$ and $n_{i1} = \boldsymbol{e_i}^T\boldsymbol{m_{11}}$ (Table S3). Similarly, we define $n_{ij0} = \#\{k : \boldsymbol{t}_k = \boldsymbol{e_i} + \boldsymbol{e_j}, x_k - z_k = -1\}$, $n_{ij1} = \#\{k : \boldsymbol{t}_k = \boldsymbol{e_i} + \boldsymbol{e_j}, x_k - z_k \geq 0\}$, $n_{ijl0} = \#\{k : \boldsymbol{t}_k = \boldsymbol{e_i} + \boldsymbol{e_j} + \boldsymbol{e_l}, x_k - z_k = -1\}$, $n_{ijl1} = \#\{k : \boldsymbol{t}_k = \boldsymbol{e_i} + \boldsymbol{e_j} + \boldsymbol{e_l}, x_k - z_k \geq 0\}$. From Tables S3 and S4, we know that $n_{131} = 1, n_{331} = 1, n_{111} = 2, n_{141} = 1, n_{371} = 3, n_{571} =$

$2, n_{370} = 1, n_{150} = 1, n_{550} = 1, n_{3341} = 1$ and all other $n_{ij0}, n_{ij1}, n_{ijl0}, n_{ijl1}$ $(1 \leq i \leq j \leq l \leq 8)$ not listed above all equal 0.

With these data, we can express the log-likelihood function corresponding to (44) as

$$\begin{aligned}
&\sum_{i=1}^{8} n_{i1} \log(\mathbb{P}(\boldsymbol{T} = \boldsymbol{e_i}, X - Z \geq 0)) + n_{i0} \log(\mathbb{P}(\boldsymbol{T} = \boldsymbol{e_i}, X - Z = -1)) \\
&+ \sum_{i=1}^{8} \sum_{j=i}^{8} n_{ij1} \log(\mathbb{P}(\boldsymbol{T} = \boldsymbol{e_i} + \boldsymbol{e_j}, X - Z \geq 0)) + n_{ij0} \log(\mathbb{P}(\boldsymbol{T} = \boldsymbol{e_i} + \boldsymbol{e_j}, X - Z = -1)) \\
&+ \sum_{i=1}^{8} \sum_{j=i}^{8} \sum_{l=j}^{8} n_{ijl1} \log(\mathbb{P}(\boldsymbol{T} = \boldsymbol{e_i} + \boldsymbol{e_j} + \boldsymbol{e_l}, X - Z \geq 0)) + n_{ijl0} \log(\mathbb{P}(\boldsymbol{T} = \boldsymbol{e_i} + \boldsymbol{e_j} + \boldsymbol{e_l}, X - Z = -1)) \\
&- \left(\sum_{i=1}^{8} (n_{i1} + n_{i0}) + \sum_{i=1}^{8} \sum_{j=i}^{8} (n_{ij1} + n_{ij0}) + \sum_{i=1}^{8} \sum_{j=i}^{8} \sum_{l=j}^{8} (n_{ijl1} + n_{ijl0}) \right) \log(\mathbb{P}(\mathbf{1}^T \boldsymbol{T} > 0)). \qquad (45)
\end{aligned}$$

The three probabilities in (45) are computed in the next three subsections. After substituting these expressions and the estimates $\hat{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_8)$ from Table S2, we can express the log-likelihood as a function of α, β, λ_s, λ_{ns}, b_s, b_{ns}, w_s, w_{ns}, τ_s and τ_{ns}, which we denote by $l(\alpha, \beta, \lambda_s, \lambda_{ns}, b_s, b_{ns}, w_s, w_{ns}, \tau_s, \tau_{ns})$, and derive the maximum likelihood estimates by solving

$$\begin{aligned}
\max_{\alpha,\beta,\lambda_s,\lambda_{ns},b_s,b_{ns},w_s,w_{ns},\tau_s,\tau_{ns}} \quad & l(\alpha, \beta, \lambda_s, \lambda_{ns}, b_s, b_{ns}, w_s, w_{ns}, \tau_s, \tau_{ns}), \\
\text{subject to} \quad & \alpha \geq 0, \\
& \beta \geq 0, \\
& \lambda_s \geq 0, \\
& \lambda_{ns} \geq 0, \\
& 0 \leq b_s \leq 1, \\
& 0 \leq b_{ns} \leq 1, \\
& 0 \leq w_s \leq 1, \\
& 0 \leq w_{ns} \leq 1, \\
& 0 \leq \tau_s \leq 1,
\end{aligned}$$

$$0 \leq \tau_{ns} \leq 1.$$

The resulting estimates appear in Table S5.

4.4 The Probability of Not Hitting CODIS in Expression (45)

The probability of not hitting CODIS, which appears in (45), is

$$\mathbb{P}(\boldsymbol{T} = \boldsymbol{t}, X - Z = -1) = \int_0^1 \mathbb{P}(\boldsymbol{T} = \boldsymbol{t}, X - Z = -1|s) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds. \tag{46}$$

Before computing the conditional probability on the right side of (46), we define

$$\Gamma = \sum_{i=1}^{8} m_i(1-\theta_i) = (\mathbf{1} - \boldsymbol{\theta})^T \boldsymbol{m},$$

which generalizes the definition in (10) to the eight-class model, and

$$\eta = \prod_{i=1}^{8} \frac{m_i^{t_i}\theta_i^{t_i}}{t_i!},$$

and prove the following lemma.

Lemma 1: Assuming $0 \leq t_i \leq z_i, i = 1, \ldots, 7$ and $0 \leq t_8 \leq 1 + n - \sum_{i=1}^{7} z_i$, we have for $l = 1, \ldots, 7$,

$$\begin{aligned}
&\sum_{z_l=t_l}^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l+1}^{8} t_i} \cdots \sum_{z_7=t_7}^{1+n-\sum_{j=1}^{6} z_j - t_8} \frac{m_l^{z_l} \cdots m_7^{z_7} m_8^{1+n-\sum_{k=1}^{7} z_k}}{z_l! \cdots z_7! (1+n-\sum_{k=1}^{7} z_k)!} \frac{z_l!}{(z_l - t_l)!} (1-\theta_l)^{z_l - t_l} \\
&\quad \cdots \frac{z_7!}{(z_7 - t_7)!} (1-\theta_7)^{z_7 - t_7} \frac{(1+n-\sum_{k=1}^{7} z_k)!}{(1+n-\sum_{k=1}^{7} z_k - t_8)!} (1-\theta_8)^{1+n-\sum_{k=1}^{7} z_k - t_8} \\
&= \frac{1}{(1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l}^{8} t_i)!} \left(\sum_{k=l}^{8} m_k(1-\theta_k) \right)^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l}^{8} t_i} \prod_{k=l}^{8} m_k^{t_k}.
\end{aligned} \tag{47}$$

Proof: We first show (47) for the case where $l = 7$. If we let $N = 1 + n - \sum_{j=1}^{6} z_j$, then

$$\begin{aligned}
&\sum_{z_7=t_7}^{N-t_8} \frac{m_7^{z_7} m_8^{N-z_7}}{z_7!(N-z_7)!} \frac{z_7!}{(z_7-t_7)!} (1-\theta_7)^{z_7-t_7} \frac{(N-z_7)!}{(N-z_7-t_8)!} (1-\theta_8)^{N-z_7-t_8} \\
&= \sum_{z_7=t_7}^{N-t_8} \frac{1}{(z_7-t_7)!(N-z_7-t_8)!} \left(\frac{m_7(1-\theta_7)}{m_8(1-\theta_8)} \right)^{z_7-t_7} m_7^{t_7} m_8^{N-t_7} (1-\theta_8)^{N-t_7-t_8},
\end{aligned}$$

$$= \frac{1}{(N-t_7-t_8)!}\left(\frac{m_7(1-\theta_7)}{m_8(1-\theta_8)}+1\right)^{N-t_7-t_8} m_7^{t_7} m_8^{N-t_7}(1-\theta_8)^{N-t_7-t_8},$$
$$= \frac{1}{(N-t_7-t_8)!}\left(m_7(1-\theta_7)+m_8(1-\theta_8)\right)^{N-t_7-t_8} m_7^{t_7} m_8^{t_8},$$
$$= \frac{1}{(1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l}^{8} t_i)!}\left(\sum_{k=l}^{8} m_k(1-\theta_k)\right)^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l}^{8} t_i} \prod_{k=l}^{8} m_k^{t_k}.$$

Suppose (47) holds for $l+1$. We now show that it holds for l:

$$\sum_{z_l=t_l}^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l+1}^{8} t_i} \cdots \sum_{z_7=t_7}^{1+n-\sum_{j=1}^{6} z_j - t_8} \frac{m_l^{z_l}\cdots m_7^{z_7} m_8^{1+n-\sum_{k=1}^{7} z_k}}{z_l!\cdots z_7!(1+n-\sum_{k=1}^{7} z_k)!}\frac{z_l!}{(z_l-t_l)!}(1-\theta_l)^{z_l-t_l}$$
$$\cdots \frac{z_7!}{(z_7-t_7)!}(1-\theta_7)^{z_7-t_7}\frac{(1+n-\sum_{k=1}^{7} z_k)!}{(1+n-\sum_{k=1}^{7} z_k - t_8)!}(1-\theta_8)^{1+n-\sum_{k=1}^{7} z_k - t_8}$$
$$= \sum_{z_l=t_l}^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l+1}^{8} t_i} \frac{m_l^{z_l}}{z_l!}\frac{z_l!}{(z_l-t_l)!}(1-\theta_l)^{z_l-t_l}\frac{1}{(1+n-\sum_{j=1}^{l} z_j - \sum_{i=l+1}^{8} t_i)!}$$
$$\left(\sum_{k=l+1}^{8} m_k(1-\theta_k)\right)^{1+n-\sum_{j=1}^{l} z_j - \sum_{i=l+1}^{8} t_i} \prod_{k=l+1}^{8} m_k^{t_k},$$
$$= \sum_{z_l=t_l}^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l+1}^{8} t_i} \frac{(m_l(1-\theta_l))^{z_l-t_l}\left(\sum_{k=l+1}^{8} m_k(1-\theta_k)\right)^{1+n-\sum_{j=1}^{l} z_j - \sum_{i=l+1}^{8} t_i}}{(z_l-t_l)!(1+n-\sum_{j=1}^{l} z_j - \sum_{i=l+1}^{8} t_i)!}\prod_{k=l}^{8} m_k^{t_k},$$
$$= \frac{1}{(1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l}^{8} t_i)!}\left(\sum_{k=l}^{8} m_k(1-\theta_k)\right)^{1+n-\sum_{j=1}^{l-1} z_j - \sum_{i=l}^{8} t_i} \prod_{k=l}^{8} m_k^{t_k}. \quad \square$$

Returning to the conditional probability in (46), we have

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z=-1|s)\\
&= \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z=-1, Z\geq \mathbf{1}^T\boldsymbol{t}, X\geq \mathbf{1}^T\boldsymbol{t}-1|s),\\
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \mathbb{P}(X=n|s)\mathbb{P}(Z-1=n|X=n,s)\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|Z=1+n,s),\\
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} b^n \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|Z=1+n,s). \qquad (48)
\end{aligned}$$

The conditional probability on the right side of (48) can be expressed as

$$\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|Z=1+n,s)$$
$$= \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, Z_i\geq t_i\ \forall i|Z=1+n,s),$$

$$
\begin{aligned}
&= \sum_{z_1=t_1}^{1+n-\sum_{i=2}^{8} t_i} \sum_{z_2=t_2}^{1+n-z_1-\sum_{i=3}^{8} t_i} \cdots \sum_{z_7=t_7}^{1+n-\sum_{j=1}^{6} z_j - t_8} \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, \boldsymbol{Z}=\boldsymbol{z} | Z = 1+n, s), \\
&= \sum_{z_1=t_1}^{1+n-\sum_{i=2}^{8} t_i} \sum_{z_2=t_2}^{1+n-z_1-\sum_{i=3}^{8} t_i} \cdots \sum_{z_7=t_7}^{1+n-\sum_{j=1}^{6} z_j - t_8} \frac{(1+n)!}{z_1! \cdots z_7! (1+n-\sum_{k=1}^{7} z_k)!} m_1^{z_1} \cdots m_7^{z_7} m_8^{1+n-\sum_{k=1}^{7} z_k} \\
&\quad \times \binom{z_1}{t_1} \theta_1^{t_1} (1-\theta_1)^{z_1-t_1} \cdots \binom{z_7}{t_7} \theta_7^{t_7} (1-\theta_1)^{z_7-t_7} \binom{1+n-\sum_{k=1}^{7} z_k}{t_8} \theta_8^{t_8} (1-\theta_8)^{1+n-\sum_{k=1}^{7} z_k - t_8}, \\
&= (1+n)! \prod_{i=1}^{8} \frac{\theta_i^{t_i}}{t_i!} \sum_{z_1=t_1}^{1+n-\sum_{i=2}^{8} t_i} \sum_{z_2=t_2}^{1+n-z_1-\sum_{i=3}^{8} t_i} \cdots \sum_{z_7=t_7}^{1+n-\sum_{j=1}^{6} z_j - t_8} \frac{m_1^{z_1} \cdots m_7^{z_7} m_8^{1+n-\sum_{k=1}^{7} z_k}}{z_1! \cdots z_7! (1+n-\sum_{k=1}^{7} z_k)!} \\
&\quad \times \frac{z_1!}{(z_1-t_1)!} (1-\theta_1)^{z_1-t_1} \cdots \frac{z_7!}{(z_7-t_7)!} (1-\theta_7)^{z_7-t_7} \frac{(1+n-\sum_{k=1}^{7} z_k)!}{(1+n-\sum_{k=1}^{7} z_k - t_8)!} (1-\theta_8)^{1+n-\sum_{k=1}^{7} z_k - t_8}, \\
&= (1+n)! \prod_{i=1}^{8} \frac{\theta_i^{t_i}}{t_i!} \frac{1}{(1+n-\sum_{i=1}^{8} t_i)!} \left(\sum_{k=1}^{8} m_k (1-\theta_k) \right)^{1+n-\sum_{i=1}^{8} t_i} \prod_{k=1}^{8} m_k^{t_k} \quad \text{by} \quad (47), \\
&= \frac{(1+n)!}{(1+n-\mathbf{1}^T \boldsymbol{t})!} \left((\mathbf{1}-\boldsymbol{\theta})^T \boldsymbol{m} \right)^{1+n-\mathbf{1}^T \boldsymbol{t}} \prod_{i=1}^{8} \frac{m_i^{t_i} \theta_i^{t_i}}{t_i!}. \qquad (49)
\end{aligned}
$$

Now we substitute (49) into (48) to obtain the final expression for the conditional probability of not hitting CODIS:

$$
\begin{aligned}
&\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X - Z = -1 | s) \\
&= \sum_{n=\mathbf{1}^T \boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} b^n \frac{(1+n)!}{(1+n-\mathbf{1}^T \boldsymbol{t})!} \left((\mathbf{1}-\boldsymbol{\theta})^T \boldsymbol{m} \right)^{1+n-\mathbf{1}^T \boldsymbol{t}} \prod_{i=1}^{8} \frac{m_i^{t_i} \theta_i^{t_i}}{t_i!}, \\
&= \sum_{n=\mathbf{1}^T \boldsymbol{t}-1}^{\infty} \frac{1+n}{(1+n-\mathbf{1}^T \boldsymbol{t})!} (\lambda b \Gamma)^{1+n-\mathbf{1}^T \boldsymbol{t}} (\lambda b)^{\mathbf{1}^T \boldsymbol{t}-1} e^{-\lambda} \eta, \\
&= \sum_{n=\mathbf{1}^T \boldsymbol{t}-1}^{\infty} \frac{1+n-\mathbf{1}^T \boldsymbol{t}+\mathbf{1}^T \boldsymbol{t}}{(1+n-\mathbf{1}^T \boldsymbol{t})!} (\lambda b \Gamma)^{n+1-\mathbf{1}^T \boldsymbol{t}} (\lambda b)^{\mathbf{1}^T \boldsymbol{t}-1} e^{-\lambda} \eta, \\
&= \left(\sum_{n=\mathbf{1}^T \boldsymbol{t}}^{\infty} \frac{(\lambda b \Gamma)^{n+1-\mathbf{1}^T \boldsymbol{t}}}{(n-\mathbf{1}^T \boldsymbol{t})!} + \mathbf{1}^T \boldsymbol{t} \sum_{n=\mathbf{1}^T \boldsymbol{t}-1}^{\infty} \frac{(\lambda b \Gamma)^{n+1-\mathbf{1}^T \boldsymbol{t}}}{(1+n-\mathbf{1}^T \boldsymbol{t})!} \right) (\lambda b)^{\mathbf{1}^T \boldsymbol{t}-1} e^{-\lambda} \eta, \\
&= e^{\Gamma \lambda b - \lambda} (\Gamma \lambda b + \mathbf{1}^T \boldsymbol{t}) (\lambda b)^{\mathbf{1}^T \boldsymbol{t}-1} \eta. \qquad (50)
\end{aligned}
$$

Finally substituting (50) into (46) yields

$$
\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X - Z = -1) = \int_0^1 e^{\Gamma \lambda b - \lambda} (\Gamma \lambda b + \mathbf{1}^T \boldsymbol{t}) (\lambda b)^{\mathbf{1}^T \boldsymbol{t}-1} \eta \frac{s^{\alpha-1} (1-s)^{\beta-1}}{B(\alpha, \beta)} \, ds. \qquad (51)
$$

4.5 The Probability of Hitting CODIS in Expression (45)

The probability of hitting CODIS, which appears in (45), is

$$\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z\geq 0)=\int_0^1 \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z\geq 0|s)\frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\,ds, \tag{52}$$

where

$$\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z\geq 0|s)=\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|s)-\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z=-1|s). \tag{53}$$

Before calculating the first conditional probability on the right side of (53), we prove the following lemma.

Lemma 2: When $n\geq \mathbf{1}^T\boldsymbol{t}-1$,

$$\sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n}\frac{(1+k)}{(n-k)!(1+k-\mathbf{1}^T\boldsymbol{t})!}x^k=\frac{(n+1-\mathbf{1}^T\boldsymbol{t})(x+1)^{n-\mathbf{1}^T\boldsymbol{t}}x^{\mathbf{1}^T\boldsymbol{t}}+(\mathbf{1}^T\boldsymbol{t})(x+1)^{n+1-\mathbf{1}^T\boldsymbol{t}}x^{\mathbf{1}^T\boldsymbol{t}-1}}{(n+1-\mathbf{1}^T\boldsymbol{t})!}. \tag{54}$$

Proof:

$$\begin{aligned}
\sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n}\frac{(1+k)}{(n-k)!(1+k-\mathbf{1}^T\boldsymbol{t})!}x^k &= \frac{d}{dx}\sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n}\frac{x^{k+1}}{(n-k)!(1+k-\mathbf{1}^T\boldsymbol{t})!},\\
&= \frac{d}{dx}\sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n}\frac{(n+1-\mathbf{1}^T\boldsymbol{t})!x^{k+1-\mathbf{1}^T\boldsymbol{t}}}{(n-k)!(1+k-\mathbf{1}^T\boldsymbol{t})!}\frac{x^{\mathbf{1}^T\boldsymbol{t}}}{(n+1-\mathbf{1}^T\boldsymbol{t})!},\\
&= \frac{d}{dx}\frac{(x+1)^{n+1-\mathbf{1}^T\boldsymbol{t}}x^{\mathbf{1}^T\boldsymbol{t}}}{(n+1-\mathbf{1}^T\boldsymbol{t})!},\\
&= \frac{(n+1-\mathbf{1}^T\boldsymbol{t})(x+1)^{n-\mathbf{1}^T\boldsymbol{t}}x^{\mathbf{1}^T\boldsymbol{t}}+(\mathbf{1}^T\boldsymbol{t})(x+1)^{n+1-\mathbf{1}^T\boldsymbol{t}}x^{\mathbf{1}^T\boldsymbol{t}-1}}{(n+1-\mathbf{1}^T\boldsymbol{t})!}.
\end{aligned}$$

□

The first conditional probability on the right side of (53) can be calculated as follows:

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|s)\\
&=\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, Z\geq \mathbf{1}^T\boldsymbol{t}, X\geq \mathbf{1}^T\boldsymbol{t}-1|s),\\
&=\sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty}\mathbb{P}(X=n|s)\sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n}\mathbb{P}(Z-1=k|X=n,s)\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|Z=1+k,s),\\
&=\sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty}\frac{\lambda^n e^{-\lambda}}{n!}\sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n}\binom{n}{k}b^k(1-b)^{n-k}\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|Z=1+k,s),
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n} \binom{n}{k} b^k (1-b)^{n-k} \frac{(1+k)!}{(1+k-\mathbf{1}^T\boldsymbol{t})!} \left((\mathbf{1}-\boldsymbol{\theta})^T\boldsymbol{m}\right)^{1+k-\mathbf{1}^T\boldsymbol{t}} \eta \quad \text{by} \quad (49), \\
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n} \frac{1+k}{(n-k)!(1+k-\mathbf{1}^T\boldsymbol{t})!} b^k (1-b)^{n-k} \Gamma^{1+k-\mathbf{1}^T\boldsymbol{t}} n! \eta, \\
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \sum_{k=\mathbf{1}^T\boldsymbol{t}-1}^{n} \frac{1+k}{(n-k)!(1+k-\mathbf{1}^T\boldsymbol{t})!} \left(\frac{b\Gamma}{(1-b)}\right)^k (1-b)^n \Gamma^{1-\mathbf{1}^T\boldsymbol{t}} n! \eta, \\
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \frac{(n+1-\mathbf{1}^T\boldsymbol{t}) \left(\frac{b\Gamma}{(1-b)}+1\right)^{n-\mathbf{1}^T\boldsymbol{t}} \left(\frac{b\Gamma}{(1-b)}\right)^{\mathbf{1}^T\boldsymbol{t}} + (\mathbf{1}^T\boldsymbol{t}) \left(\frac{b\Gamma}{(1-b)}+1\right)^{n+1-\mathbf{1}^T\boldsymbol{t}} \left(\frac{b\Gamma}{(1-b)}\right)^{\mathbf{1}^T\boldsymbol{t}-1}}{(n+1-\mathbf{1}^T\boldsymbol{t})!} \\
&\quad \times (1-b)^n \Gamma^{1-\mathbf{1}^T\boldsymbol{t}} n! \eta \quad \text{by} \quad (54), \\
&= \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \frac{n!(b\Gamma+1-b)^n}{(n+1-\mathbf{1}^T\boldsymbol{t})!} \left[(n+1-\mathbf{1}^T\boldsymbol{t}) \left(\frac{b\Gamma}{b\Gamma+1-b}\right)^{\mathbf{1}^T\boldsymbol{t}} + \mathbf{1}^T\boldsymbol{t} \left(\frac{b\Gamma}{b\Gamma+1-b}\right)^{\mathbf{1}^T\boldsymbol{t}-1}\right] \Gamma^{1-\mathbf{1}^T\boldsymbol{t}} \eta, \\
&= \left[\sum_{n=\mathbf{1}^T\boldsymbol{t}}^{\infty} \frac{\lambda^n (b\Gamma+1-b)^{n-1}}{(n-\mathbf{1}^T\boldsymbol{t})!} b\Gamma + \sum_{n=\mathbf{1}^T\boldsymbol{t}-1}^{\infty} \frac{\lambda^n (b\Gamma+1-b)^n}{(n+1-\mathbf{1}^T\boldsymbol{t})!} (\mathbf{1}^T\boldsymbol{t})\right] \left(\frac{b\Gamma}{b\Gamma+1-b}\right)^{\mathbf{1}^T\boldsymbol{t}-1} \Gamma^{1-\mathbf{1}^T\boldsymbol{t}} e^{-\lambda} \eta, \\
&= \left(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t}\right) e^{\lambda(b\Gamma+1-b)} (\lambda(b\Gamma+1-b))^{\mathbf{1}^T\boldsymbol{t}-1} \left(\frac{b\Gamma}{b\Gamma+1-b}\right)^{\mathbf{1}^T\boldsymbol{t}-1} \Gamma^{1-\mathbf{1}^T\boldsymbol{t}} e^{-\lambda} \eta, \\
&= \left(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t}\right) e^{\Gamma\lambda b-\lambda b} (\lambda b)^{\mathbf{1}^T\boldsymbol{t}-1} \eta. \qquad (55)
\end{aligned}
$$

Substituting (50) and (55) into (53) gives the conditional probability of hitting CODIS,

$$
\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z\geq 0|s) = (e^{\Gamma\lambda b-\lambda b} - e^{\Gamma\lambda b-\lambda}) \left(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t}\right) (\lambda b)^{\mathbf{1}^T\boldsymbol{t}-1} \eta. \qquad (56)
$$

Substituting (56) into (52) gives the probability of hitting CODIS,

$$
\mathbb{P}(\boldsymbol{T}=\boldsymbol{t}, X-Z\geq 0) = \int_0^1 (e^{\Gamma\lambda b-\lambda b} - e^{\Gamma\lambda b-\lambda}) \left(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t}\right) (\lambda b)^{\mathbf{1}^T\boldsymbol{t}-1} \eta \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds. \quad (57)
$$

4.6 The Probability of Being Observed in Expression (45)

The probability of being observed in (45) is given by

$$
\begin{aligned}
\mathbb{P}(\mathbf{1}^T\boldsymbol{T}>\mathbf{0}) &= \int_0^1 \mathbb{P}(\mathbf{1}^T\boldsymbol{T}>0|s) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds, \\
&= \int_0^1 (1-\mathbb{P}(\mathbf{1}^T\boldsymbol{T}=0|s)) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds, \\
&= \int_0^1 (1-\mathbb{P}(\boldsymbol{T}=\mathbf{0}|s)) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds,
\end{aligned}
$$

$$= \int_0^1 (1 - \Gamma e^{\Gamma\lambda b - \lambda b}) \frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds \quad \text{by (55).} \tag{58}$$

4.7 Integrating With Respect to the Beta Distribution

Because the beta density has singular points when $\alpha < 1, s \downarrow 0$ and $\beta < 1, s \uparrow 1$, we use integration by parts to avoid the numerical instability when computing equations (51), (57) and (58), which appear in the log-likelihood in (45).

If we define

$$f_1(s) = (e^{\Gamma\lambda b - \lambda b} - e^{\Gamma\lambda b - \lambda})(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t})(\lambda b)^{\mathbf{1}^T\boldsymbol{t}-1} \prod_{i=1}^{8} m_i^{t_i},$$

$$f_2(s) = e^{\Gamma\lambda b - \lambda}(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t})(\lambda b)^{\mathbf{1}^T\boldsymbol{t}-1} \prod_{i=1}^{8} m_i^{t_i},$$

$$f_3(s) = 1 - \Gamma e^{\Gamma\lambda b - \lambda b},$$

then equations (51), (57) and (58) can be expressed as

$$\mathbb{P}(\boldsymbol{T} = \boldsymbol{t}, X - Z \geq 0) \;\; = \frac{\prod_{i=1}^{8} \frac{\theta_i^{t_i}}{t_i!}}{B(\alpha,\beta)} \int_0^1 f_1(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds,$$

$$\mathbb{P}(\boldsymbol{T} = \boldsymbol{t}, X - Z = -1) \;\; = \frac{\prod_{i=1}^{8} \frac{\theta_i^{t_i}}{t_i!}}{B(\alpha,\beta)} \int_0^1 f_2(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds,$$

$$\mathbb{P}(\mathbf{1}^T\boldsymbol{T} > \mathbf{0}) = \frac{1}{B(\alpha,\beta)} \int_0^1 f_3(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds.$$

The function $f_3(s)$ does not include negative exponents of s or $1-s$ and when $\mathbf{1}^T\boldsymbol{t} \geq 0$, $f_1(s)$ and $f_2(s)$ also do not include negative exponents of s or $1-s$. Thus we can calculate $\int_0^1 f_i(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds$ $(\alpha > 0, \beta > 0)$ as follows to avoid the numerical instability:

$$\begin{aligned}
&\int_0^1 f_i(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds \\
&= \int_0^{0.5} f_i(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds + \int_{0.5}^1 f_i(s) s^{\alpha-1}(1-s)^{\beta-1}\, ds, \\
&= \frac{1}{\alpha} \int_0^{0.5} f_i(s)(1-s)^{\beta-1}\, d(s^\alpha) - \frac{1}{\beta} \int_{0.5}^1 f_i(s) s^{\alpha-1}\, d((1-s)^\beta), \\
&= \frac{1}{\alpha} \left(\frac{f_i(0.5)}{2^{\alpha+\beta-1}} - \int_0^{0.5} s^\alpha [f_i(s)(1-s)^{\beta-1}]'\, ds \right) + \frac{1}{\beta} \left(\frac{f_i(0.5)}{2^{\alpha+\beta-1}} + \int_{0.5}^1 (1-s)^\beta [f_i(s) s^{\alpha-1}]'\, ds \right).
\end{aligned}$$

This procedure is also used to compute the integrals in equations (12), (16) and (17).

4.8 Performance Analysis and Policy Optimization

We consider a generic policy $\boldsymbol{x} = (x_1, \ldots, x_8)$, where $\boldsymbol{x}$ substitutes for $\boldsymbol{\theta}$. Let $\boldsymbol{t}$ be defined as in §4.3, but for the generic policy $\boldsymbol{x}$. Each of the $\mathbf{1}^T\boldsymbol{t}$ SAKs tested under the $\boldsymbol{x}$ policy belongs to a certain offender. Because our earlier calculations apply only to SAKs that generate recoverable DNA, the probability that a tested SAK generates a hit conditioned on it having recoverable DNA, which we denote by $H(\boldsymbol{x})$, is

$$\begin{aligned} H(\boldsymbol{x}) &= \frac{\mathbb{E}[(\mathbf{1}^T\boldsymbol{T})\mathbb{1}_{\{\mathbf{1}^T\boldsymbol{T}>1 \text{ or } X-Z\geq 0\}}|\mathbf{1}^T\boldsymbol{T}>0]}{\mathbb{E}[\mathbf{1}^T\boldsymbol{T}|\mathbf{1}^T\boldsymbol{T}>0]}, \\ &= \frac{\mathbb{E}[\mathbf{1}^T\boldsymbol{T}|\mathbf{1}^T\boldsymbol{T}>0] - \mathbb{P}(\mathbf{1}^T\boldsymbol{T}=1, X-Z=-1|\mathbf{1}^T\boldsymbol{T}>0)}{\mathbb{E}[\mathbf{1}^T\boldsymbol{T}|\mathbf{1}^T\boldsymbol{T}>0]}, \\ &= \frac{\mathbb{E}[\mathbf{1}^T\boldsymbol{T}] - \mathbb{P}(\mathbf{1}^T\boldsymbol{T}=1, X-Z=-1)}{\mathbb{E}[\mathbf{1}^T\boldsymbol{T}]}, \\ &= 1 - \frac{\sum_{i=1}^{8}\mathbb{P}(\boldsymbol{T}=\boldsymbol{e_i}, X-Z=-1)}{\mathbb{E}[\mathbf{1}^T\boldsymbol{T}]}, \end{aligned} \tag{59}$$

where

$$\begin{aligned} \mathbb{E}[\mathbf{1}^T\boldsymbol{T}] &= \sum_{k=1}^{\infty} k \sum_{\mathbf{1}^T\boldsymbol{t}=k} \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}), \\ &= \sum_{k=1}^{\infty} k \sum_{\mathbf{1}^T\boldsymbol{t}=k} \int_0^1 \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|s)\frac{s^{\alpha-1}(1-s)^{\beta-1}}{B(\alpha,\beta)}\, ds, \\ &= \int_0^1 \sum_{k=1}^{\infty} k \sum_{\mathbf{1}^T\boldsymbol{t}=k} \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|s)\frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha,\beta)}\, ds. \end{aligned} \tag{60}$$

By equation (55), we have

$$\begin{aligned} &\sum_{k=1}^{\infty} k \sum_{\mathbf{1}^T\boldsymbol{t}=k} \mathbb{P}(\boldsymbol{T}=\boldsymbol{t}|s), \\ &= \sum_{k=1}^{\infty} k \sum_{\mathbf{1}^T\boldsymbol{t}=k} \left(\Gamma\lambda b + \mathbf{1}^T\boldsymbol{t}\right) e^{\Gamma\lambda b-\lambda b}(\lambda b)^{\mathbf{1}^T\boldsymbol{t}-1}\prod_{i=1}^{8}\frac{m_i^{t_i}\theta_i^{t_i}}{t_i!}, \\ &= \sum_{k=1}^{\infty} k \left(\Gamma\lambda b + k\right) e^{\Gamma\lambda b-\lambda b}(\lambda b)^{k-1} \sum_{\mathbf{1}^T\boldsymbol{t}=k}\prod_{i=1}^{8}\frac{m_i^{t_i}\theta_i^{t_i}}{t_i!}, \\ &= \sum_{k=1}^{\infty} k \left(\Gamma\lambda b + k\right) e^{\Gamma\lambda b-\lambda b}(\lambda b)^{k-1}\frac{\left(\sum_{i=1}^{8} m_i\theta_i\right)^k}{k!}, \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} (\Gamma\lambda b + k)\, e^{\Gamma\lambda b - \lambda b} (\lambda b)^{k-1} \frac{(1-\Gamma)^k}{(k-1)!}, \\
&= \sum_{k=0}^{\infty} (\Gamma\lambda b + k + 1)\, e^{\Gamma\lambda b - \lambda b} (\lambda b)^{k} \frac{(1-\Gamma)^{k+1}}{k!}, \\
&= e^{\Gamma\lambda b - \lambda b} (\Gamma\lambda b + 1)(1-\Gamma) \sum_{k=0}^{\infty} \frac{(\lambda b(1-\Gamma))^k}{k!} + e^{\Gamma\lambda b - \lambda b}(1-\Gamma) \sum_{k=0}^{\infty} k \frac{(\lambda b(1-\Gamma))^k}{k!}, \\
&= e^{\Gamma\lambda b - \lambda b} (\Gamma\lambda b + 1)(1-\Gamma) e^{\lambda b(1-\Gamma)} + e^{\Gamma\lambda b - \lambda b}(1-\Gamma) e^{\lambda b(1-\Gamma)} \lambda b (1-\Gamma), \\
&= (1+\lambda b)(1-\Gamma). \qquad (61)
\end{aligned}
$$

Substituting equation (61) into (60) gives us

$$
\mathbb{E}[\mathbf{1}^T \boldsymbol{T}] = \int_0^1 (1+\lambda b)(1-\Gamma) \frac{s^{\alpha-1}(1-s)^{\beta}}{B(\alpha,\beta)}\, ds. \qquad (62)
$$

Substituting equations (51) and (62) into (59) yields the hit probability.

Let ρ be the proportion of the backlog that is tested. Repeating the logic that leads to equation (31), we find that the optimal policy satisfies the optimization problem:

$$
\begin{aligned}
\max_{x_1,\cdots,x_8} \quad & N(\sum_{i=1}^{8} \hat{q}_i \hat{d}_i x_i) H(x_1, \ldots, x_8), && (63) \\
\text{subject to} \quad & \sum_{i=1}^{8} \hat{q}_i x_i \le \rho, && (64) \\
& 0 \le x_i \le 1,\ i = 1, \cdots, 8. && (65)
\end{aligned}
$$

The optimal solution to (63)-(65) is a priority policy, where the eight classes are ranked as in Table S6. Note that this optimal policy differs from an intuitive policy that ranks the classes according to $\hat{d}_i H(\boldsymbol{e_i})$; this policy yields the order (1,3,6,5,2,7,4,8) rather than the optimal order, (1,3,2,6,5,7,4,8), and is suboptimal because it only captures the value of each class separately without considering the interaction among classes. To calculate the expected number of hits for the no-priority policy, we substitute $x_i = \rho, i = 1, \ldots, 8$ into (63).

References

[1] Campbell R, Fehler-Cabral G, Pierce SJ, Sharma DB, Bybee D, Shaw J et al. The Detroit sexual assault kit (SAK) action research project (ARP), final report. National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, Washington, D.C., 2015.

[2] Campbell R, Pierce SJ, Sharma DB, Feeney H, Fehler-Cabral G. Should rape kit testing be prioritized by victim-offender relationship? Empirical comparison of forensic testing outcomes for stranger and nonstranger sexual assaults. Criminology & Public Policy 2016;15:555-583.

[3] Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, FL: Chapman and Hall/CRC Press, 1993.

[4] Mandel M, Betensky RA. Simultaneous confidence intervals based on the percentile bootstrap approach. Computational Statistics & Data Analysis 2008;52:2158-216.

[5] Walfield SM. When a cleared rape is not cleared: A multilevel study of arrest and exceptional clearance. Journal of Interpersonal Violence 2016;31:1767-1792.

Table S1. *Definition of the eight classes.*

Class	Victim-Offender Relationship	Weapon Use	Time Delay
1	stranger	yes	≤ 1 day
2	stranger	yes	≥ 2 days
3	stranger	no	≤ 1 day
4	stranger	no	≥ 2 days
5	nonstranger	yes	≤ 1 day
6	nonstranger	yes	≥ 2 days
7	nonstranger	no	≤ 1 day
8	nonstranger	no	≥ 2 days

Table S2. *Class-dependent parameter estimates for the eight-class model.*

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
$\hat{q}_i$	0.180	0.018	0.223	0.028	0.095	0.010	0.383	0.063
$\hat{d}_i$	0.554	0.435	0.520	0.314	0.487	0.500	0.495	0.205
$\hat{\theta}_i$	0.150	0.150	0.150	0.150	0.149	0.149	0.149	0.149

Table S3. *Among offenders who are associated with only one tested SAK in the backlog, the number of offenders of each class who have (i.e., $x_i - z_i \geq 0$), and do not have (i.e., $x_i - z_i = -1$), a DNA Profile in CODIS.*

	1	2	3	4	5	6	7	8
$x_i - z_i \geq 0$	73	5	63	5	32	5	114	7
$x_i - z_i = -1$	39	5	68	4	19	1	113	9

Table S4. *Class affiliation data for offenders with multiple tested SAKs in the backlog.*

$x_i - z_i$	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
≥ 1	1	0	1	0	0	0	0	0
≥ 1	0	0	2	0	0	0	0	0
≥ 1	2	0	0	0	0	0	0	0
≥ 1	1	0	0	1	0	0	0	0
≥ 1	0	0	1	0	0	0	1	0
≥ 1	2	0	0	0	0	0	0	0
≥ 1	0	0	1	0	0	0	1	0
≥ 1	0	0	0	0	1	0	1	0
≥ 1	0	0	1	0	0	0	1	0
≥ 1	0	0	0	0	1	0	1	0
$= -1$	0	0	1	0	0	0	1	0
$= -1$	1	0	0	0	1	0	0	0
$= -1$	0	0	0	0	2	0	0	0
≥ 1	0	0	2	1	0	0	0	0

Table S5. *Maximum likelihood estimates for the eight-class model.*

α	β	λ_s	λ_{ns}	b_s	b_{ns}	w_s	w_{ns}	τ_s	τ_{ns}
0.750	0.983	1.530	0.530	0.346	1.335×10^{-6}	0.688	1.134×10^{-7}	0.911	0.942

Table S6. *The priority policy in the eight-class model, where rank 1 is the top priority and rank 8 is the bottom priority.*

Ranking	Victim-Offender Relationship	Weapon Use	Time Delay
1	stranger	yes	≤ 1 day
2	stranger	no	≤ 1 day
3	stranger	yes	≥ 2 days
4	nonstranger	yes	≥ 2 days
5	nonstranger	yes	≤ 1 day
6	nonstranger	no	≤ 1 day
7	stranger	no	≥ 2 days
8	nonstranger	no	≥ 2 days

Figure S1

Histogram of the number of backlogged SAKs created in each year.

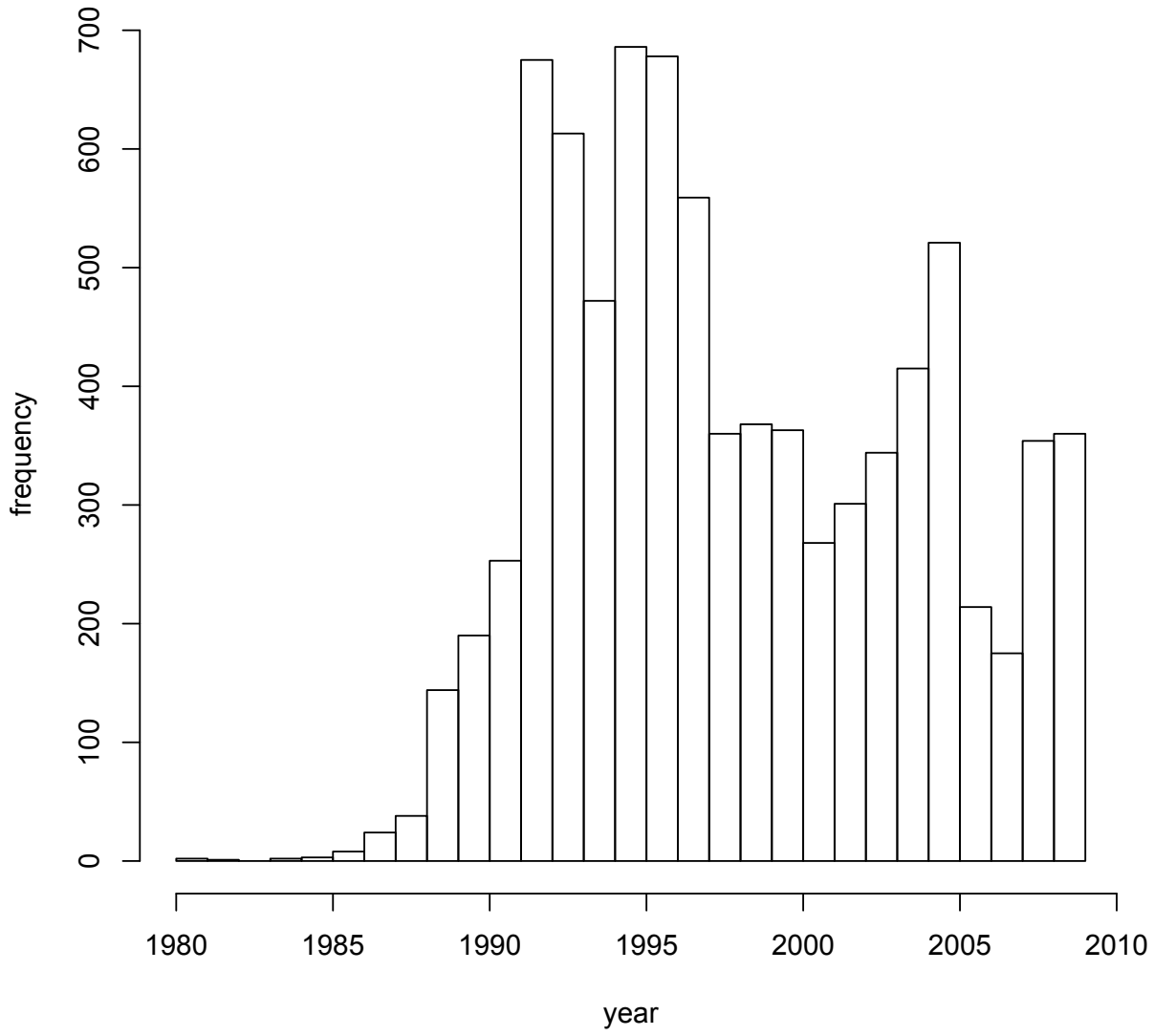


Figure S2

Classification of the 1595 SAKs for the eight-class model.

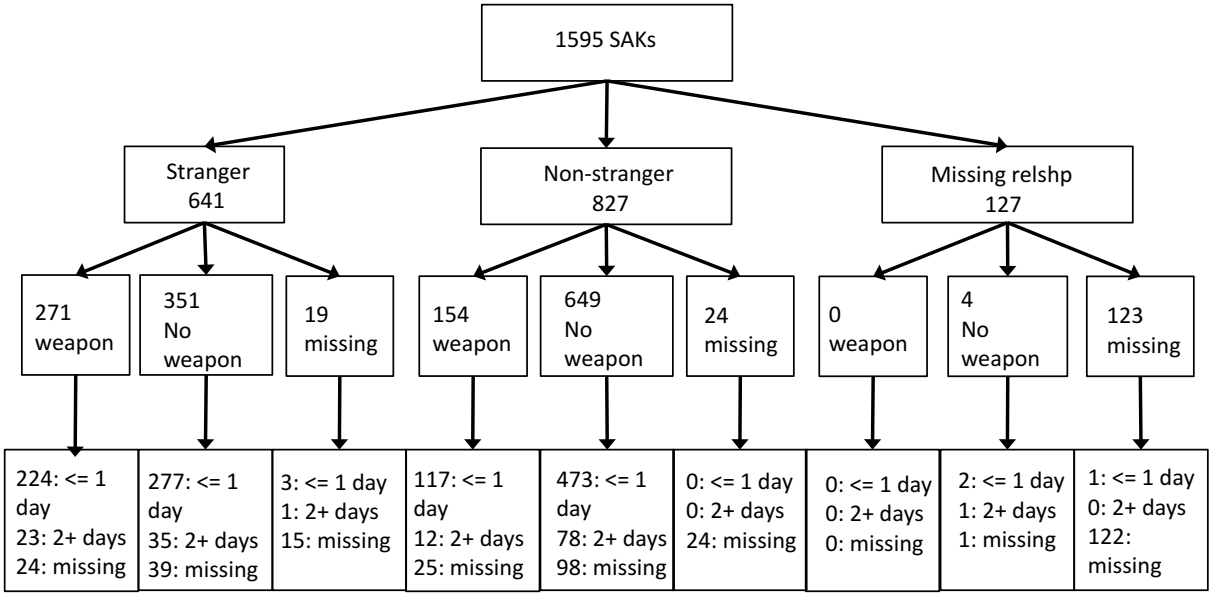
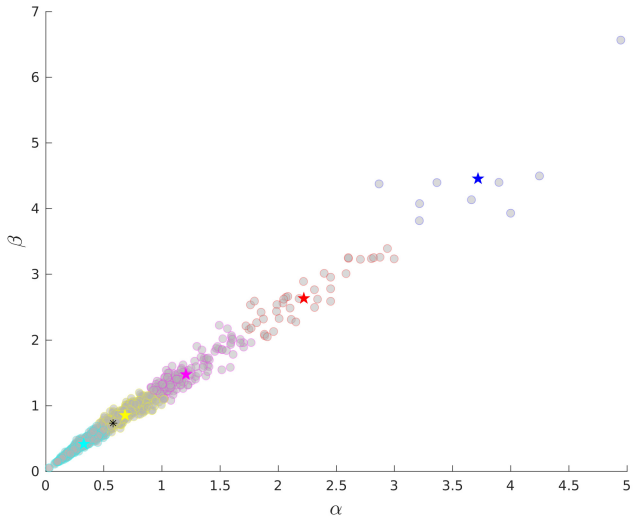
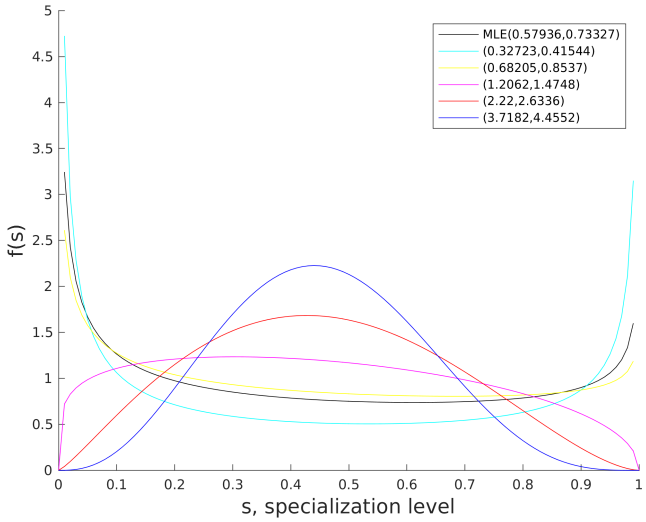


Figure S3

Cluster Analysis of the bootstrapped parameters of the beta distribution.



(a)

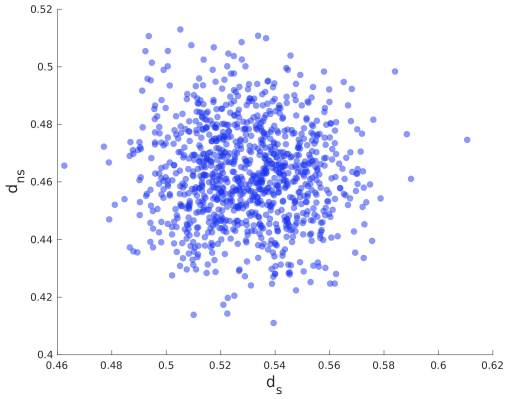


(b)

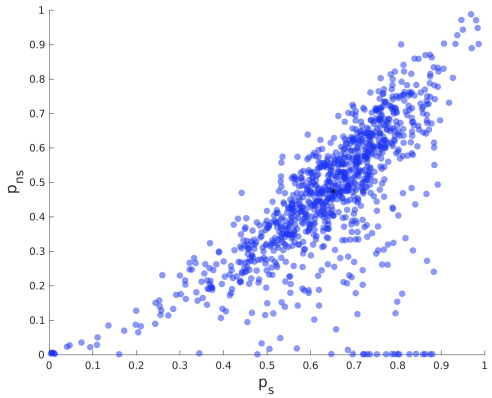
NOTE: **(a)** The 1000 bootstrapped samples (α, β) clustered into five colored groups, with the maximum likelihood estimate denoted by a black star. The proportion of points in each group (from left to right) is 0.463 (cyan), 0.353 (orange), 0.135 (magenta), 0.040 (red) and 0.009 (blue). **(b)** The five beta PDFs corresponding to the centers of the five groups from part (a).

Figure S4

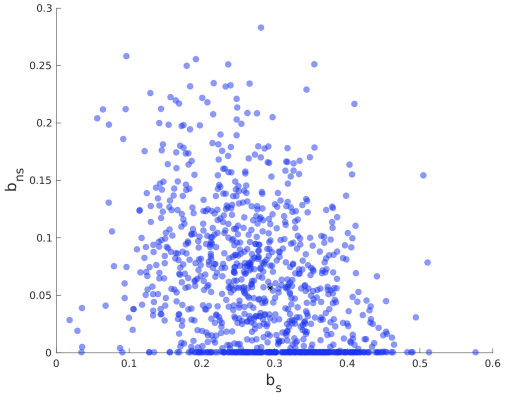
The scatter plot of the 1000 bootstrapped values of (a) $(\mathbf{d_s}, \mathbf{d_{ns}})$, (b) $(\mathbf{p_s}, \mathbf{p_{ns}})$ and (c) $(\mathbf{b_s}, \mathbf{b_{ns}})$.



(a)



(b)



(c)

NOTE: The number of points above the diagonal in the three plots are 9, 24 and 41, respectively.

Figure S5

The bootstrapped histogram for the normalized AUC of the stranger-priority policy minus the normalized AUC of the no-priority policy.

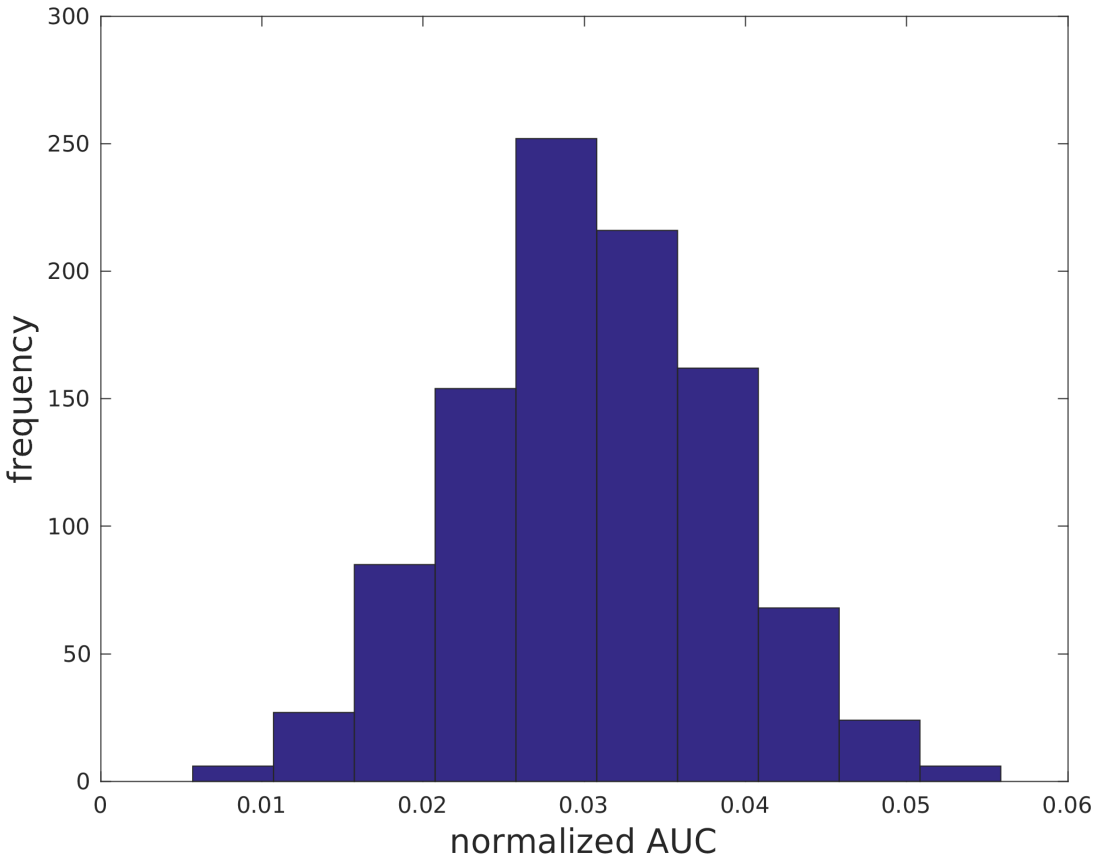
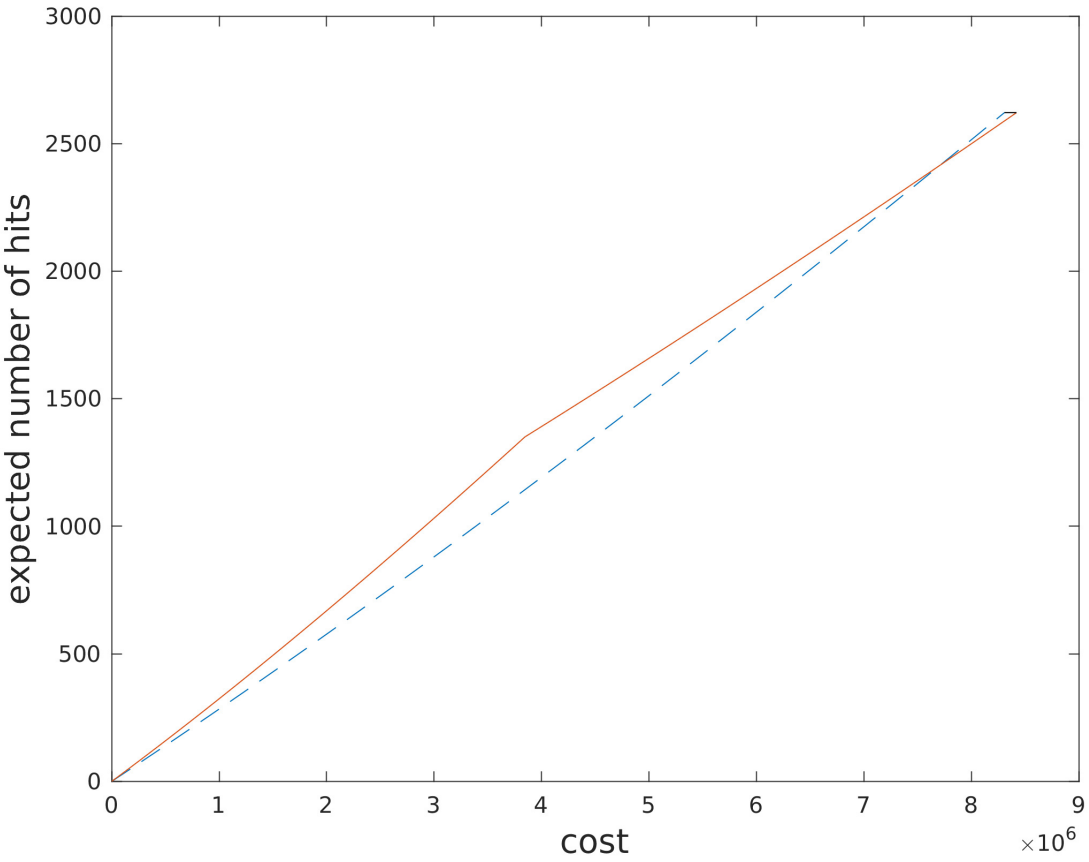


Figure S6

The expected number of hits vs. the cost after adding the sorting cost of $30 per stranger SAK to the testing cost of $1000/SAK.



NOTE: The cost to process the entire backlog is $8.307M for the no-priority policy, and is 8307[0.4494(1030)+0.5506(1000)]=$8.419M for the stranger-priority policy. The net normalized area between the two curves is 0.034.

Figure S7

For the eight-class model, the expected number of hits vs. the proportion of the backlog processed, for the optimal policy (–), which is described in Table S6, and the no-priority policy (- -)

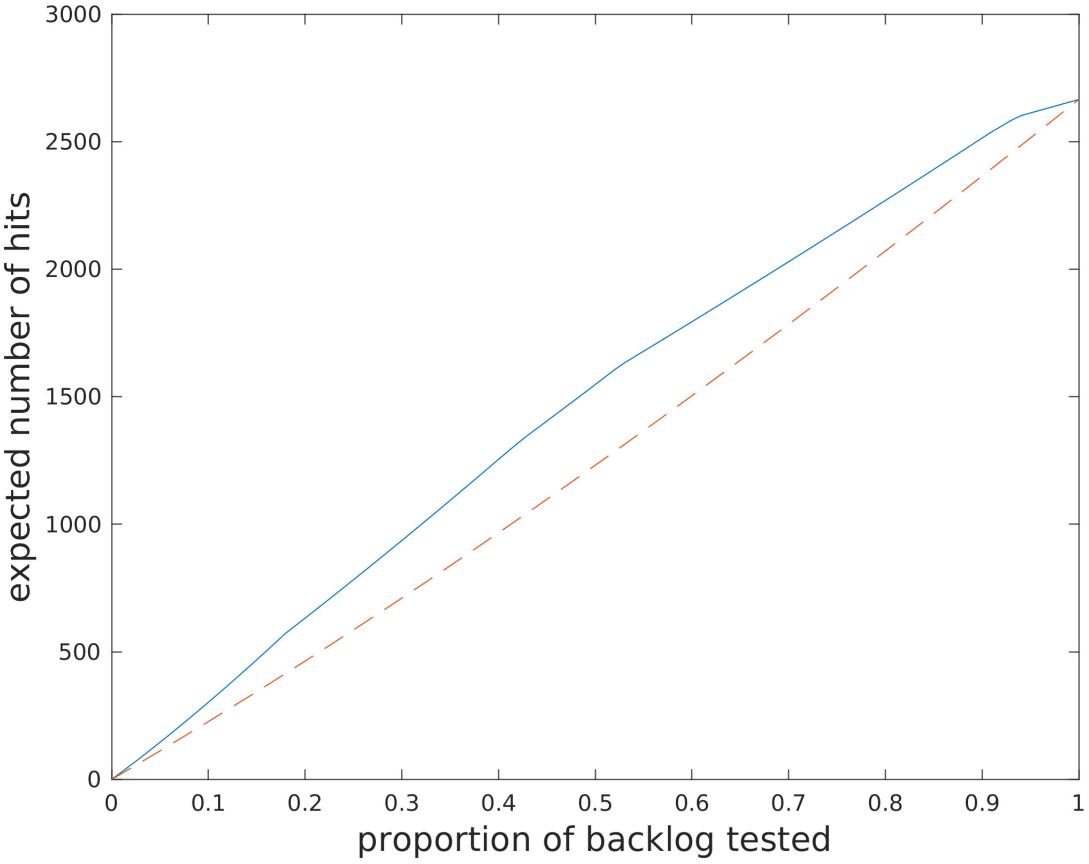
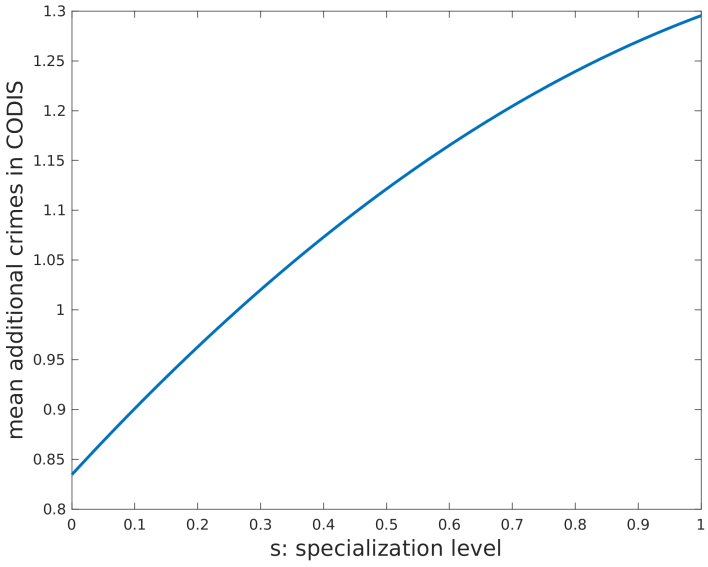
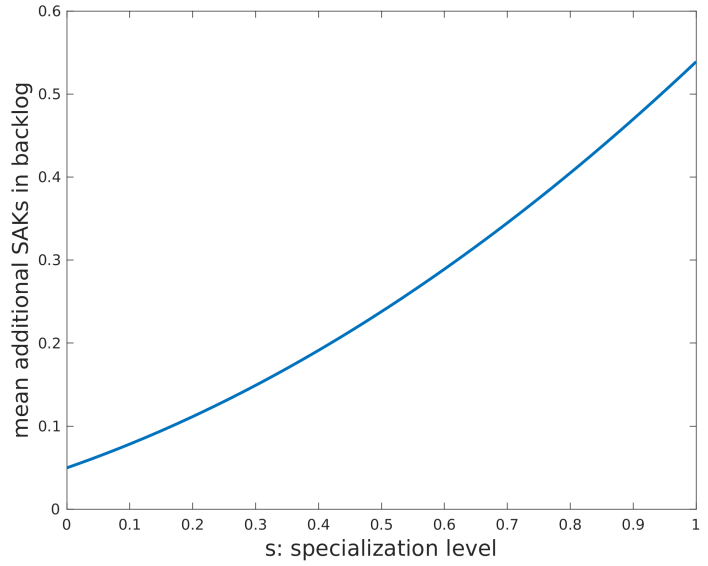


Figure S8

(a) The mean number of additional crimes in CODIS for an offender (i.e., $\lambda(1-\mathrm{b})$) vs. the offender's specialization parameter s, and (b) the mean number of additional SAKs in the backlog for an offender (i.e., $\lambda\mathrm{b}$) vs. the offender's specialization parameter s.





(a)