Supplemental Information

TUTORIAL ON SCREENING AND ADVANCED METHODS

This material is designed for readers who are seeking supplemental information about the screening trade-offs presented in Rodday et al. Part 1 provides definitions of screening terminology and statistics used throughout the article; we have also provided the equations for calculating each of these statistics. Part 2 uses figures and examples to illustrate the trade-offs between sensitivity and specificity for any screening test. Part 3 demonstrates the effect of sensitivity, specificity, and prevalence on the number of true-positives, false-positives, PPV, and NPV. Part 4 describes in detail our methodology with respect to the HSROC used in Rodday et al. The use of HSROC was necessary because most studies in the meta-analysis reported sensitivity and specificity for multiple positivity thresholds. Finally, Part 5 discusses how we back-calculated the prevalence as defined by a non-ECG reference standard for LQTS.

PART 1: SCREENING DEFINITIONS

Screening tools in general can be described by different statistics that, taken together, permit appraisal of their utility in identifying certain disorders or diseases and can inform decision-making at the policy level.[52] For example, understanding the number of false-positives when detecting 1 case allows for a better understanding of the downstream implications for screening programs. We define commonly used screening test statistics below and in Supplemental Table 4.

Sensitivity = true-positive rate = probability of having a positive test among patients with disease = p(T+|D+) = TP/(TP + FN)

Specificity = true-negative rate = probability of having a negative test among patients without disease = p(T−|D−) = TN/(TN + FP)

False-positive rate = probability of having a positive test among patients without disease = 1 − specificity = FP/(FP+TN)

False-negative rate = probability of having a negative test among patients with disease = 1 − sensitivity = FN/(FN+TP)

Positive predictive value (PPV) = probability of having disease among patients with a positive test = p (D+|T+) = (sensitivity*prevalence)/[(sensitivity*prevalence)+(1-specificity)*(1− prevalence)]

Negative predictive value (NPV) = probability of not having disease among patients with a negative test = p(D−|T−) = [specificity*(1-prevalence)]/[specificity*(1-prevalence)+(1-sensitivity)*prevalence]

False-alarm rate = probability of not having disease in patients with a positive test = p(D−|T+) = 1 − PPV

False-reassurance rate = probability of having disease in patients with a negative test = p(D+|T−) = 1 − NPV

Number needed to screen to detect 1 case = 1/(sensitivity*prevalence)

Number of false-positives when detecting 1 case = (1− prevalence)*(1− specificity)/ (sensitivity*− prevalence)

Number of false-negatives per 100 000 screened = 100 000*prevalence*(1− sensitivity)

PART 2: TRADE-OFFS BETWEEN SENSITIVITY AND SPECIFICITY

With any screening program, there are trade-offs between sensitivity and specificity as demonstrated by receiver operating characteristic (ROC) curves (Supplemental Fig 4), which plot sensitivity on the y-axis as a function of 1 minus specificity on the x-axis. In the absence of a perfect test, increases in sensitivity will result in decreases in specificity, and vice versa. For example, at point A in Supplemental Fig 4, sensitivity is 0.4 and specificity is 0.99 (labeled as 1-0.01 in Supplemental Fig 4), but when sensitivity increases to 0.95 at point B, the specificity falls to 0.80 (1-0.2).

In general, when sensitivity increases, the number of true-positives increases, but this comes at the cost of more false-positives. Alternatively, when specificity increases, the number of true-negatives increases, but this, in turn, leads to more false-negatives. Supplemental Tables 5 and 6 demonstrate how true-positives, false-positives, false-negatives, and true-negatives change when sensitivity and specificity change for a theoretical sample of 100. Two illustrative points on a hypothetical ROC curve are used: (1) sensitivity = 90%, specificity = 56% and (2) sensitivity = 40%, specificity = 89%.

When designing a screening test, consideration must be given as to whether sensitivity, specificity, or both will be prioritized. This decision affects the number of true-positives, false-positives, true-negatives, and false-negatives. Different screening programs require different sensitivity and specificity trade-offs. For example, when screening for HIV in blood banks—a situation where one does not want to risk contaminating the blood supply—one would want to minimize false-negatives, which corresponds to selecting a high sensitivity and accepting a lower specificity. As a consequence, there would likely be many false-positives. The counterexample is former President Reagan's proposal to screen couples applying for marriage licenses for HIV. The prevalence of HIV in this population is known to be low, and false-positive results would result in stigmatization and perhaps the ending of the relationship. In this case, false-positives would want to be avoided, so specificity should be favored over sensitivity.[51]

PART 3: THE ROLE OF PREVALENCE IN SCREENING

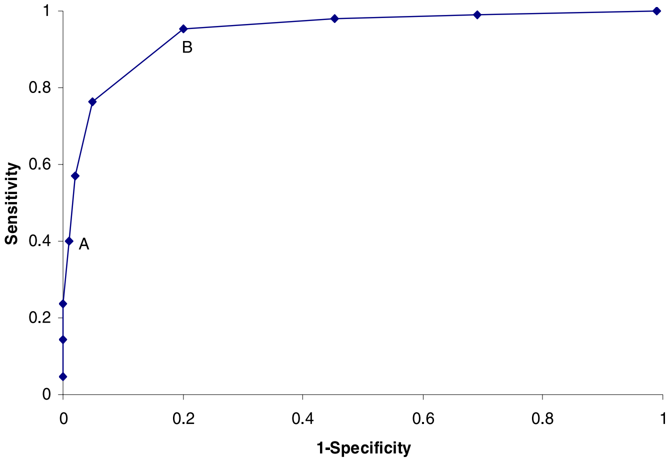
Prevalence estimates can further complicate the trade-off between sensitivity and specificity. In general, screening programs are most effective if preclinical prevalence is sufficiently high in the screened population.[11] There are instances, however, where screening is recommended for conditions with low prevalence, such as SCD.

The following figures and explanations focus on screening in the case of low prevalence conditions (0%–5%). We focus on the case where the sum of sensitivity and specificity is maximized ("maximal accuracy") and the case where specificity is maximized ("maximal specificity"). A point on the curve where sensitivity was maximized was not selected because the corresponding specificity was very low.

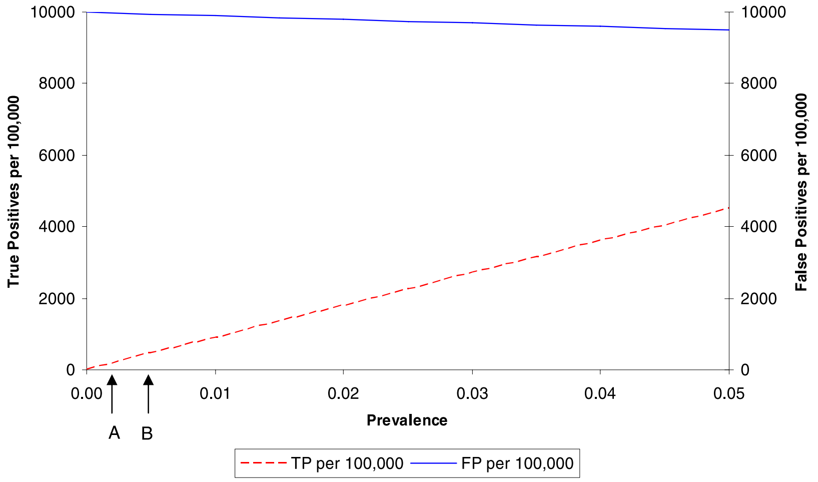
Supplemental Fig 5 demonstrates the number of true-positives and false-positives per 100 000 screened when sensitivity and specificity are both set as equal to 90% and prevalence ranges from 0% to 5%. As can be seen, as prevalence increases to 5%, the number of true-positives increases, but the false-positives remain high. Supplemental Fig 5 also provides a graphical display of the phenotypical prevalence estimates of SCD among children from Rodday et al. (point A) compared with often-cited prevalence estimates of SCD in the literature (point B).[13,19,20]

Supplemental Fig 6 next examines the impact on the PPV and NPV when sensitivity and specificity are both equal to 90% and prevalence ranges from 0% to 5%. Negative predictive value remains high and relatively stable as prevalence increases. This indicates that most of the people who test negative will not have the disorder. Contrastingly, PPV increases slightly as prevalence increases, but stays below 30%, which indicates that many of those who test positive will not actually have the disorder.

Next, we examine the situation in which specificity is maximized. Supplemental Fig 7 demonstrates the number of true-positives and false-positives per 100 000 screened when sensitivity is 50% and specificity is 99% and prevalence ranges from 0% to 5%. As compared



SUPPLEMENTAL FIGURE 4
Trade-offs between sensitivity and specificity.



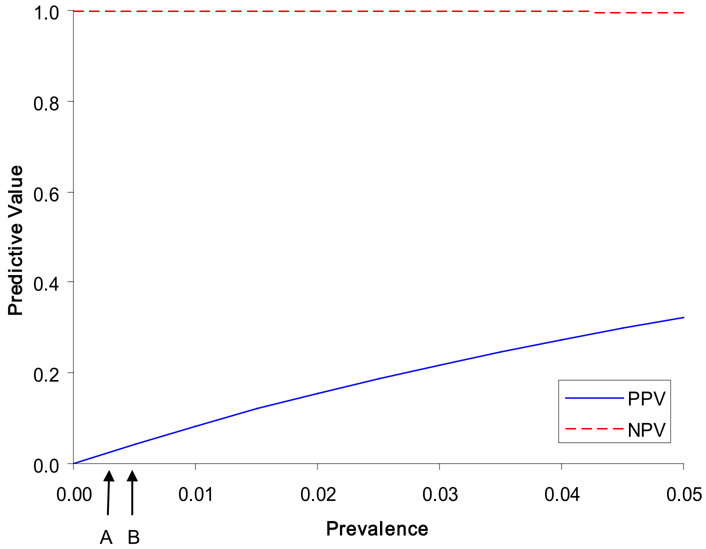
SUPPLEMENTAL FIGURE 5
Role of prevalence in number of true-positives (TP) and false-positives (FP) per 100 000 screened, where sensitivity = 90% and specificity = 90%. A = 188 per 100 000; prevalence estimate from Rodday et al. B = 450 per 100 000; prevalence estimate from often-cited references.[13,19,20]

with Supplemental Fig 5, where sensitivity and specificity were both 90%, the false-positive rate per 100 000 is much lower and the true-positive rate increases at a slower rate.

Supplemental Fig 8 demonstrates the PPV and NPV when sensitivity is 50% and specificity is 99% and prevalence ranges from 0% to 5%. Compared with Supplemental Fig 6, where sensitivity and specificity were both maximized at 90%, the PPV increases more quickly as prevalence increases, which indicates that among those who test positive, more will actually have the disease. Negative predictive value continues to remain high in this situation.

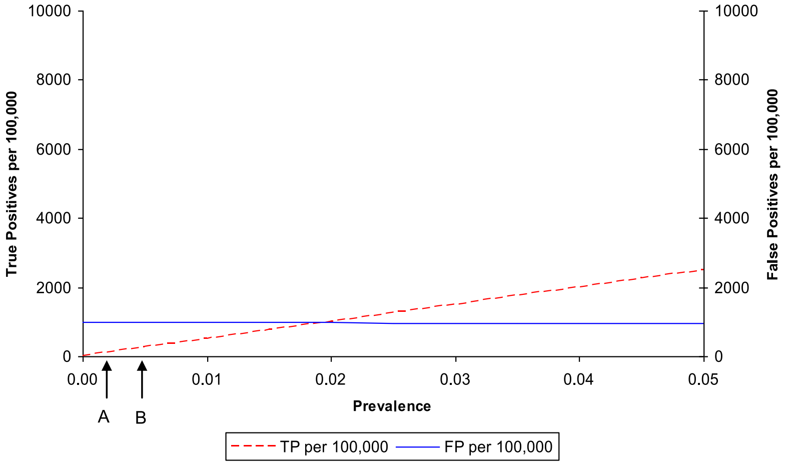
Two prevalence estimates are included on each figure: point A represents the phenotypic prevalence estimate from Rodday et al (188 per 100 000) and point B represents often-cited prevalence estimates for these disorders (450 per 100 000).[13,19,20] Although there is debate about which of these numbers represents a more accurate prevalence estimate of the disorders that cause SCD in children, we can see that from a policy perspective that the differences in the number of true-positives, false-positives, PPV, and NPV between these 2 estimates are not very large.

Together, these figures and examples demonstrate an important concept when considering screening programs for low prevalence disorders: increasing specificity helps to decrease the number of false-positives and increase the PPV, while having little effect on the NPV.



SUPPLEMENTAL FIGURE 6

Role of prevalence in PPV and NPV, where sensitivity = 90% and specificity = 90%. A = 188 per 100 000; prevalence estimate from Rodday et al. B = 450 per 100 000; prevalence estimate from often-cited references.[13,19,20]



SUPPLEMENTAL FIGURE 7

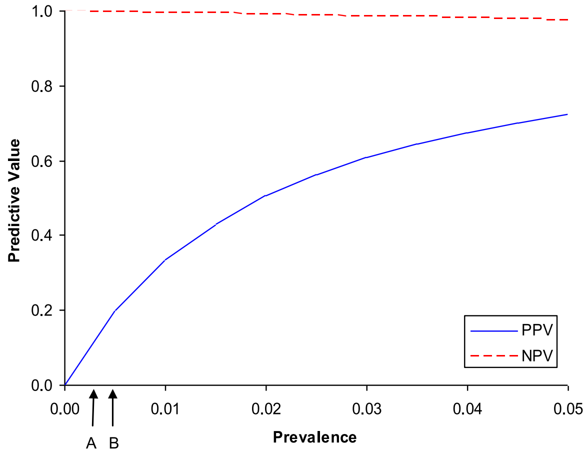
Role of prevalence in number of true-positives (TP) and false-positives (FP) per 100 000 screened, where sensitivity = 50% and specificity = 99%. A = 188 per 100 000; prevalence estimate from Rodday et al. B = 450 per 100 000; prevalence estimate from often-cited references.[13,19,20]

PART 4: APPLICATION OF HIERARCHICAL SUMMARY RECEIVER OPERATING CHARACTERISTIC IN RODDAY ET AL

Sensitivity and specificity are correlated across studies that use different thresholds for test positivity. When the one increases, we expect the other to decrease (ie, the correlation is negative). In such cases, the most appropriate summary of diagnostic performance is a Summary ROC curve (SROC). An SROC curve resembles the ROC curve of individual studies, and describes the tradeoff between average sensitivity and average specificity across the examined studies.

The Dukic and Gatsonis method that we used in our analysis was based on the Rutter and Gatsonis description of a hierarchical SROC (HSROC) model, essentially, a hierarchical (2-level) logistic regression. The first level (study level) models within-study variability. The second level models between-study variability and describes the relationship between the average sensitivity and specificity in terms of positivity threshold and accuracy parameters (ie, parameters that describe the shape of the curve). The resulting regression line is a hierarchical SROC, or HSROC line.[16,17]

The Rutter-Gatsonis model was developed for studies that report results at a single threshold (ie, a single 2 × 2 table), however. In our case, most



SUPPLEMENTAL FIGURE 8

Role of prevalence in PPV and NPV, where sensitivity = 50% and specificity = 99%. A = 188 per 100 000; prevalence estimate from Rodday et al. B = 450 per 100 000; prevalence estimate from often-cited references.[13,19,20]

SUPPLEMENTAL TABLE 1 Screening Characteristics

		Disease (D)	
		Present	Absent
Test Result (T)	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

SUPPLEMENTAL TABLE 2 Higher Sensitivity (90%), Lower Specificity (56%)

	Disease (D)			
	Present	Absent		Total
Test Result (T)	Positive	TP=9	FP=40	49
	Negative	FN=1	TN=50	51
	Total	10	90	100

FN, false-negative; FP, false-positive; TN, true-negative; TP, true-positive.

SUPPLEMENTAL TABLE 3 Higher Specificity (89%), Lower Sensitivity (40%)

	Disease (D)			
	Present	Absent		Total
TestResult (T)	Positive	TP=4	FP=10	14
	Negative	FN=6	TN=80	86
	Total	10	90	100

FN, false-negative; FP, false-positive; TN, true-negative; TP, true-positive.

studies reported sensitivity and specificity for multiple positivity thresholds (ie, there were multiple 2 × 2 tables with 1 per positivity criterion threshold). Therefore, the sensitivity and specificity pairs corresponding to different thresholds in the same study were not independent. To account for the complexity of data reported at multiple thresholds, we used an extension of the typical HSROC model, as described by Dukic and Gatsonis.[18] The key difference with the previous model is that it uses an ordinal hierarchical logistic regression approach (whereas logistic regressions handle 2 categories, ordinal logistic regressions handle multiple and ordered categories). We specified the model for multiple thresholds in the Bayesian framework, and fit it by using Gibbs sampling with Markov Chain Monte Carlo in OpenBugs 3.03 via R and the BRugs library. The model code is available from the authors on request.

PART 5: BACK-CALCULATING THE PREVALENCE AS DEFINED BY A NON-ECG REFERENCE STANDARD

We can estimate the general population prevalence of LQTS as defined by genetic testing from the proportion of screening ECGs that are suggestive of LQTS, provided that we have some knowledge of the sensitivity and specificity of ECG testing. The idea is straightforward: the sensitivity of ECG can provide information on the number of true-positive ECGs, and the specificity of ECG can provide information on the number of false-positive ECGs. The sum of true- and false-positive is the total number of positive screening ECGs. One can thus set up a system of equations and solve for the unobserved (latent, unknown) prevalence of LQTS as defined by genetic testing. To properly account for the uncertainty in our knowledge of the sensitivity and specificity of ECG, we should not treat them as fixed numbers, but as distributions. This intuitive description can be formalized as a calculation that combines the data (number of positive ECGs in various studies) and knowledge external to the data ("prior" knowledge, such as the distribution of the sensitivity and specificity of ECG and a "prior" [best guess] on the distribution of the unknown prevalence of LQTS as defined by genetic testing) to obtain a "posterior" distribution on the prevalence of LQTS as defined by genetic testing based on Bayes theorem.

Priors

We used normal priors on the logit-transformed sensitivity and specificity of ECG. We obtained the SDs of the logit transformed sensitivity and specificity of ECG from the results of univariate metaanalyses of sensitivity and specificity. The means of the logit-transformed sensitivity and specificity were selected based on the HSROC curves. We assumed that in a screening setting, one would select ECG criteria that would have a specificity of 95%; we selected the sensitivity that corresponded to this value from the meta-analytic HSROC curve (82%).

We explored 3 different priors for the latent prevalence of LQTS by genetic testing: the first was a noninformative prior (a very vague prior that assigns almost equal probability to all possible prevalence values). The other 2 priors were informative and were based on a study of genetic testing for LQTS in newborns by Schwartz et al.[13] Both informative priors assumed that the mean prevalence of LQTS by genetic testing in children and young adults was the same as that in the above-mentioned study in newborns (1 in 2534 children or young adults),[13] but differed in the precision with which this unobserved prevalence is known. The second choice for a prior distribution assumed that the prevalence is known with the precision conferred by a study of ~2534 children or young adults; and the third choice assumed that the prevalence is known with the precision conferred by a study of ~43 c080 sample size (equal to the sample size of the Schwartz et al study of newborns[13]).

Model

We made an extension of the model by Joseph et al[22] to more than 1 study. The code is available from the authors on request. A summary of the model follows.

$$X_i \sim \text{Binomial}(p_i, n_i),$$

where X_i is the number of positive ECG findings in a given study, p_i is the probability of a positive ECG in a given study, and n_i is the size of the study. p_i is calculated as follows:

$$\begin{aligned} p_i = {} & \text{sensitivity of ECG}^{*}\text{prevalence}_i \\ & + (1 - \text{specificity of ECG})^{*} \\ & \times (1 - \text{prevalence}_i), \end{aligned}$$

where

$$\begin{aligned} & \text{logit}(\text{sensitivity of ECG}) \sim \text{Normal} \\ & \quad (\text{logit}[\text{sensitivity of ECG from} \\ & \quad \text{meta-analysis}], \\ & \quad \text{variance}(\text{logit}[\text{sensitivity of ECG from} \\ & \quad \text{meta-analysis}])) \\ & \text{logit}(\text{specificity of ECG}) \sim \text{Normal} \\ & \quad (\text{logit}[\text{specificity of ECG from} \\ & \quad \text{meta-analysis}], \\ & \quad \text{variance}(\text{logit}[\text{sensitivity of ECG from} \\ & \quad \text{meta-analysis}])) \\ & \text{logit}(\text{prevalence}_i) \sim \text{Normal} \\ & \quad (\text{logit}(\overline{prevalence}), \tau^2) \end{aligned}$$

and

$$\tau \sim \text{uniform}(0, 2)$$

For the distribution of logit$(\overline{prevalence})$, we used different parameters depending on whether we wanted a noninformative prior or a prior based on Schwartz et al.[13]

Noninformative prior:

$$\text{logit}(\overline{prevalence}) \sim \text{Normal}(0, 0.01)$$

Prior from full Schwartz et al[13] data:

$$\begin{aligned} & \text{logit}(\overline{prevalence}) \sim \text{Normal}(\text{logit} \\ & \quad [17/43\,080], \text{logit}[1/17 + 1/ \\ & \quad (43\,080 - 17)]) \end{aligned}$$

Prior based on prevalence from Schwartz et al[13] data:

$$\begin{aligned} & \text{logit}(\overline{prevalence}) \sim \text{Normal}(\text{logit} \\ & \quad [1/2534], \text{logit}[1/1 + 1/ \\ & \quad (2534 - 1)]) \end{aligned}$$