**Individual and Population Level Uncertainty Interact to Determine Performance of Outbreak Detection**

Callum R.K. Arnold1,2,[[1]](#footnote-20), Alex C. Kong3, Amy K. Winter4, William J. Moss3,5, Bryan N. Patenaude3, Matthew J. Ferrari1,2

1 Department of Biology, Pennsylvania State University, University Park, PA, USA 16802  
2 Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA, USA 16802  
3 Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA 21205  
4 Department of Epidemiology, College of Public Health, University of Georgia, Athens, GA, USA 30602  
5 Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA 21205

# Abstract

## Background

Infectious disease surveillance and outbreak detection systems often utilize diagnostic testing to validate case identification. The metrics of sensitivity, specificity, and positive predictive value are commonly discussed when evaluating the performance of diagnostic tests, and to a lesser degree, the performance of outbreak detection systems. However, the interaction of the two levels’ (the test and the alert system) metrics, is typically overlooked. Here, we describe how equivalent regions of detection accuracy can exist over a range of diagnostic test characteristics, examining the sensitivity to background noise structure and magnitude.

## Methods

We generated a stochastic SEIR model with importation to simulate true measles and non-measles sources of febrile rash (noise) daily incidence. We generated time series of febrile rash (i.e., measles clinical case definition) by summing the daily incidence of measles and either independent Poisson noise or non-measles dynamical noise (consistent with rubella virus). For each time series we assumed a fraction of all cases were seen at a healthcare clinic, and a subset of those were diagnostically confirmed using a test with sensitivity and specificity consistent with either a rapid diagnostic test (RDT) or an enzyme-linked immunosorbent assay (ELISA). From the resulting time series of test-positive cases, we define an outbreak alert as the exceedance of a threshold by the 7-day rolling average of observed (test positive) cases. For each threshold level, we calculated percentages of alerts that were aligned with an outbreak (analogous to the positive predictive value), the percentage of outbreaks detected (analogous to the sensitivity), and combined these two measures into an accuracy metric for outbreak detection. We selected the optimal threshold as the value that maximizes accuracy. We show how the optimal threshold and resulting accuracy depend on the diagnostic test, testing rate, and the type and magnitude of the non-measles noise.

## Results

The optimal threshold for each test increased monotonically as the percentage of clinic visits who were tested increased. With Poisson-only noise, similar outbreak detection accuracies could be achieved with imperfect RDT-like tests as with ELISA-like diagnostic tests (c. 93%), given moderately high testing rates. With larger delays (14 days) between the ELISA test administration and result date, RDTs could outperform the ELISA. Similar numbers of unavoidable cases and outbreak alert delays could be achieved between the test types. With dynamical noise, however, the accuracy of ELISA scenarios was far superior to those achieved with RDTs (c. 93% vs. 73%). For dynamical noise, RDT-based scenarios typically favored more sensitive alert threshold than ELISA-based scenarios (at a given testing rate), observed with lower numbers of unavoidable cases and detection delays.

## Conclusions

The performance of an outbreak detection system is highly sensitive to the structure and the magnitude of background noise. Under the assumption that the noise is relatively static over time, RDTs can perform as well as ELISA in a surveillance system. However, when the noise is temporally correlated, as from a separate SEIR process, imperfect tests cannot overcome their accuracy limitations through higher testing rates.

**Key words:** Rapid-Diagnostic Tests; ELISA; Infectious Disease Surveillance; Outbreak Detection.

# Background

At the heart of an outbreak detection system is a surveillance program built upon case detection, often utilizing individual diagnostic tests as necessary components of epidemiological investigations before the declaration of an outbreak [murrayInfectiousDiseaseSurveillance2017][zhouEliminationTropicalDisease2013][pahoIntegratedApproachCommunicable2000][worldhealthorganizationMeaslesOutbreakGuide2022][craggOutbreakResponse2018]. For diseases with non-specific symptoms, accurate measurement tools are often necessary to confidently and correctly ascribe changes in symptom prevalence within a population to a particular disease, and therefore detect outbreaks of specific pathogens. As a result, it has been commonplace for surveillance systems to be developed around high-accuracy tests, such as Polymerase Chain Reaction (PCR) tests and immunoglobulin (Ig) tests, when financially and logistically feasible [gastanaduyMeasles2019][commissionerCoronavirusCOVID19Update2020][grasslyComparisonMolecularTesting2020][ezhilanSARSCoVMERSCoVSARSCoV22021][worldhealthorganizationCholera2023][essentialprogrammeonimmunizationepiimmunizationClinicalSpecimensLaboratory2018]. Depending on the disease in question, either sensitivity (the ability to correctly detect a true positive individual) or specificity (the ability to correctly discount a true negative) will be prioritized, as they often are at odds with each other [westreichDiagnosticTestingScreening2019][shrefflerDiagnosticTestingAccuracy2024][parikhUnderstandingUsingSensitivity2008]. This balance is commonly defined within the Target Product Profile (TPP) of a test [worldhealthorganizationTargetProductProfiles], which is a set of minimum characteristics that should be met for production and widespread use, helping to guide research and development. For example, in the wake of the 2013 Ebola outbreak in Guinea a TPP was developed that listed the minimum acceptable sensitivity of 95% and specificity of 99% [chuaCaseImprovedDiagnostic2015]. Recognizing that Ebola is not the major cause of fever and other non-specific symptoms in the region, it is arguably more important to prioritize the specificity of the disease, although the authors note that the severity of the infection requires a high level of sensitivity as the consequences of a missed case are dire at an individual and population level [chuaCaseImprovedDiagnostic2015].

Much like the accuracy of an individual test, outbreak detection systems face the same issue regarding the prioritization of sensitive or specific alerts [germanSensitivityPredictiveValue2000][worldhealthorganizationOperationalThresholds2014][lewisTimelyDetectionMeningococcal2001]. For many disease systems, particularly in resource constrained environments where the burden of infectious diseases is highest [gbd2019childandadolescentcommunicablediseasecollaboratorsUnfinishedAgendaCommunicable2023][roserBurdenDisease2023], cases are counted and if a pre-determined threshold is breached, be that weekly, monthly, or some combination of the two, an alert is triggered that may launch a further investigation and/or a response [worldhealthorganizationMeaslesOutbreakGuide2022][worldhealthorganizationOperationalThresholds2014]. In effect, this discretizes a distinctly continuous phenomenon (observed cases) into a binary measure, outbreak or no outbreak, for decision making purposes. Reactive interventions, such as vaccination campaigns and non-pharmaceutical based interventions, designed to reduce transmission or limit and suppress outbreaks, early action has the potential to avert the most cases [atkinsAnticipatingFutureLearning2020][taoLogisticalConstraintsLead][graisTimeEssenceExploring2008][ferrariTimeStillEssence2014][worldhealthorganizationConfirmingInvestigatingManaging2009][minettiLessonsChallengesMeasles2013]. While this framing would point towards a sensitive (i.e., early alert) surveillance system being optimal, each action comes with both direct and indirect financial and opportunity costs stemming from unnecessary activities that limit resources for future response capabilities. Just as the balance of sensitivity and specificity of a test for an individual must be carefully evaluated, so must the balance at the outbreak level.

The concept of using incidence-based alert triggers to define the discrete event of an “outbreak” with characteristics analogous to individual tests has been well documented in the case of meningitis, measles, and malaria [worldhealthorganizationMeaslesOutbreakGuide2022][lewisTimelyDetectionMeningococcal2001][worldhealthorganizationConfirmingInvestigatingManaging2009][trotterResponseThresholdsEpidemic2015][cooperReactiveVaccinationControl2019][zalwangoEvaluationMalariaOutbreak2024][kanindaEffectivenessIncidenceThresholds2000]. However, an overlooked, yet critical, aspect of an outbreak detection system is the interplay between the individual test and outbreak alert characteristics. With their success within malaria surveillance systems, and particularly since the COVID-19 pandemic, rapid diagnostic tests (RDTs) have garnered wider acceptance, and their potential for use in other disease systems has been gaining interest [warrenerEvaluationRapidDiagnostic2023]. Despite concerns of their lower diagnostic accuracy slowing their adoption until recently [millerAddressingBarriersDevelopment2015], RDTs generally have reduced cold-chain requirements [brownRapidDiagnosticTests2020] and faster speed of result that has been show to outweigh the cost of false positive/negative results in some settings [warrenerEvaluationRapidDiagnostic2023][mcmorrowMalariaRapidDiagnostic2011][larremoreTestSensitivitySecondary2021][middletonModelingTransmissionMitigation2023].

In this paper we examine how the use of imperfect diagnostic tests affects the performance of outbreak detection for measles outbreaks in the context of a febrile rash surveillance system that includes both measles and non-measles cases. Because measles symptoms are non-specific, it is important to account for non-measles sources of febrile rash e.g., rubella, parvovirus, varicella, etc., producing the potential for false positive results in the context of imperfect tests. Currently, measles outbreaks are declared on the basis of either suspected measles cases (i.e., an individual with fever and maculopapular rash [essentialprogrammeonimmunizationepiimmunizationvaccinesandbiologicalsivbMeasles2018]) alone, cases confirmed by enzyme immunoassay / enzyme-linked immunosorbent assay, from here-on in referred to as ELISA, to detect the presence of measles-specific IgM antibodies, or a combination of the two, depending on the elimination status of the region, with countries nearer elimination increasing the use of ELISA diagnostic confirmation for suspected cases [essentialprogrammeonimmunizationepiimmunizationvaccinesandbiologicalsivbMeasles2018]. Each of these detection systems have their flaws. Although clinical case definition is very fast and requires limited resources, it is highly sensitive, and in the face of high “background noise” from non-measles sources of febrile rash, can lead to low positive predictive value (PPV), i.e., the probability that an alert accurately captures an outbreak [hutchinsEvaluationMeaslesClinical2004]. And while ELISA confirmation is the standard diagnostic test for measles surveillance and has higher specificity [essentialprogrammeonimmunizationepiimmunizationvaccinesandbiologicalsivbMeasles2018], the training and facility requirements generally mean that samples must be transported from the point of care to a separate laboratory, which incurs both costs and delays [essentialprogrammeonimmunizationepiimmunizationClinicalSpecimensLaboratory2018][worldhealthorganizationMeaslesOutbreakGuide2022][brownRapidDiagnosticTests2020] In resource-poor settings these delays may be days to weeks. A rapid diagnostic test may meet the WHO’s definition and requirements for using in a surveillance setting, if not for individual patient care [worldhealthorganizationSelectionUseEssential2021]. Recent developments show encouraging signs in the field as well as in theory, providing a compromise between diagnostic accuracy and timeliness in most [shonhaiInvestigationMeaslesOutbreak2015][warrenerEvaluationRapidDiagnostic2023][brownRapidDiagnosticTests2020], though not all [seninMeaslesIgMRapid2024], settings.

By examining the combination of alert threshold and individual test characteristic in a modeling study that explicitly incorporates dynamical background noise, we aim to illustrate the need to develop a TPP for the whole detection system, not just one component. To evaluate the alert system performance, we develop a set of outbreak definition criteria and surveillance metrics, drawing inspiration from acceptance sampling, ecological surveillance systems, and epidemiological surveillance systems guidelines and reviews [yemshanovAcceptanceSamplingCosteffective2020][christensenHerdlevelInterpretationTest2000][muratoEvaluationSamplingMethods2020][sternAutomatedOutbreakDetection1999][calbaSurveillanceSystemsEvaluation2015][guidelinesworkinggroupUpdatedGuidelinesEvaluating2001]. Using these metrics we overcome issues encountered by early warning systems that rely on dynamical values such as in defining outbreaks [jombartRealtimeMonitoringCOVID192021][proverbioPerformanceEarlyWarning2022][stolermanUsingDigitalTraces2023][brettAnticipatingEpidemicTransitions2018][salmonMonitoringCountTime2016], for example, characterizing the end of an epidemic period is important in a time series where multiple outbreaks will occur.

# Methods

## Model Structure

We constructed a stochastic compartmental non-age structured Susceptible-Exposed-Infected-Recovered (SEIR) model of measles and simulated using a modified Tau-leaping algorithm with a time step of 1 day, utilizing binomial draws to ensure compartment sizes remained positive valued [chatterjeeBinomialDistributionBased2005][gillespieApproximateAcceleratedStochastic2001]. We assumed that transmission rate () is sinusoidal with a period of one year and 20% seasonal amplitude. was set to 16, with a latent period of 10 days and infectious period of 8 days. The population was initialized with 500,000 individuals with Ghana-like birth and vaccination rates, and the final results were scaled up to the approximate 2022 population size of Ghana (33 million) [worldbankGhana]. We assumed commuter-style imports at each time step to avoid extinction; the number of imports each day were drawn from a Poisson distribution with mean proportional to the size of the population and [keelingModelingInfectiousDiseases2008]. The full table of parameters can be found in [[tbl-model-parameters]](#tbl-model-parameters). All simulations and analysis was completed in Julia version 1.10.5 [bezansonJuliaFreshApproach2017], with all code stored at <https://github.com/arnold-c/OutbreakDetection>.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | |  |  |  | | Importation rate |  | | | Population size (N) | 500,000, scaled to 33M | | | Initial proportion susceptible | 0.05 | | | Initial proportion exposed | 0.0 | | | Initial proportion infected | 0.0 | | | Initial proportion recovered | 0.95 | | |

To examine the sensitivity of the detection system to background noise, we layered the measles incidence time series with one of four noise time series structures: Poisson-only noise; or dynamical noise with rubella-like parameters that could be in- or out-of-phase, or with independent seasonality to the measles dynamics. For Poisson-only noise, the number of non-measles febrile rash cases each day were independent draws from a Poisson distribution with mean . For dynamical noise, we generated time series of cases from an SEIR model that matched the measles model in structure, but had , mean latent period of 7 days, and mean infectious period of 14 days, and added some additional noise independently drawn from a Poisson distribution with mean equal to 15% of the average daily rubella incidence from the SEIR time series to account for non-rubella sources of febrile rash ([[tbl-model-parameters]](#tbl-model-parameters)) [papadopoulosEstimatesBasicReproduction2022][RubellaCDCYellow]. The seasonality for the dynamical noise was assumed to be in-phase with measles, anti-phase with measles (peak timing 6 months later), or non-seasonal. Only dynamical in-phase noise and Poisson-only noise are presented in the main text; the anti-phase and non-seasonal dynamical noise scenarios are presented in the supplement.

For each noise structure, we simulated five different magnitudes of noise (), representing the average daily noise incidence. was calculated as a multiple () of the average daily measles incidence (): . Noise magnitudes will be denoted as for the rest of the magnitude e.g., to denote simulations where the average noise incidence is 8 times that of the average measles incidence. For the Poisson-noise scenarios, independent draws from a Poisson distribution with mean were simulated to produce the noise time series i.e., . For the dynamical noise scenarios, the rubella vaccination rate at birth was set to 85.38%, 73.83%, 50.88%, 27.89%, and 4.92% to produce equivalent values of (to within 2 decimal places): . 100 time series of 100 years were simulated for each scenario, before summarizing the distributions of outbreak detection methods.

## Defining Outbreaks

It is common to use expert review to define outbreaks when examining empirical data, but this is not feasible in a modeling study where tens of thousands of years are being simulated. To account for this, many studies only simulate a single outbreak within a time series (repeating this short stochastic simulation multiple times to ensemble results), define an outbreak as a period where > 1, or use a threshold of > 2 standard deviations (s.d.) over the mean seasonal incidence observed in empirical data (or from a ‘burn-in” period of the simulation) [sternAutomatedOutbreakDetection1999][jombartRealtimeMonitoringCOVID192021][stolermanUsingDigitalTraces2023][salmonMonitoringCountTime2016][teklehaimanotAlertThresholdAlgorithms2004][leclereAutomatedDetectionHospital2017]. Each method has its uses, but to evaluate the performance of an outbreak detection system in an endemic region where multiple sequential epidemics are expected it is important to clearly define the bounds of the outbreak, which can only be achieved by 2 s.d. > mean ( will be less than 1 after an outbreak’s peak, but still within what can be reasonably defined as the outbreak’s bounds). This, however, assumes strong seasonal forcing and regular periodicity of incidence to produce a smooth enough baseline, which is not present as countries near measles elimination status [grahamMeaslesCanonicalPath2019]. Here we define a true measles outbreak as a region of the time series that meets the following three criteria:

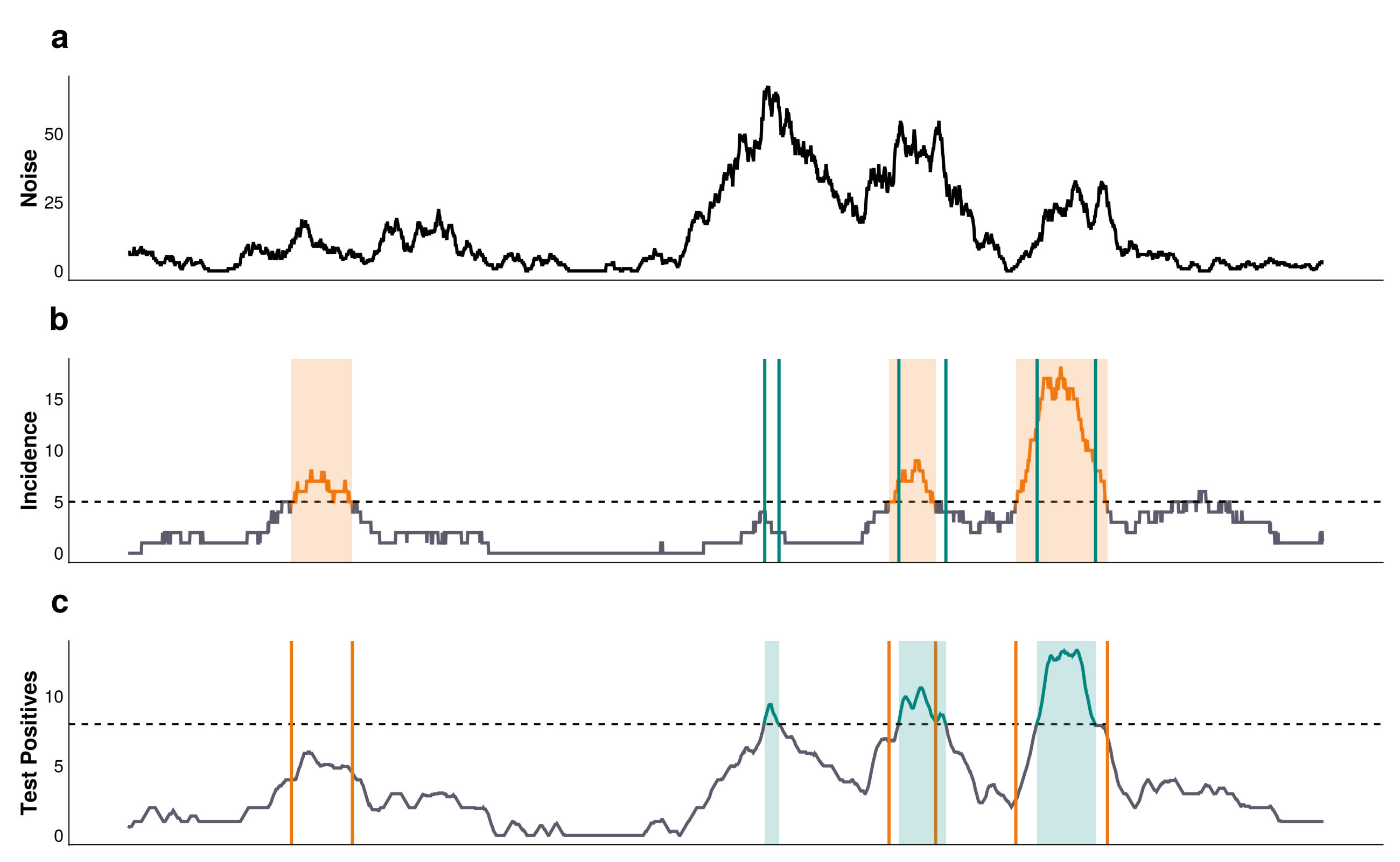
* The daily measles incidence must be greater than, or equal to, 5 cases
* The daily measles incidence must remain above 5 cases for greater than, or equal to, 30 consecutive days
* The total measles incidence must be great than, or equal to, 500 cases within the bounds of the outbreak

Only events meeting all 3 criteria are classified as outbreaks The incidence of non-measles febrile rash (i.e., noise) does not affect the outbreak status of a region but may affect the alert status triggered by the testing protocol.

Each day, 60% of the measles and non-measles febrile rash cases visit the clinic for treatment, and a percentage (P) of these clinic visits are tested, as all clinic visits are deemed to be suspected measles cases because they meet the clinical case definition. This percentage of clinic visits that are tested is varied between 10% and 60%, in 10% increments, for all combinations of diagnostic test (except clinical case definition) and alert threshold, defining the “testing scenario”. Each testing scenario uses one of the following tests:

* An RDT equivalent with 85% sensitivity and specificity, and 0-day lag in result return. That is, 85% of true measles cases will be correctly labelled as positive, and 15% of non-measles febrile rash individuals that are tested will be incorrectly labelled as positive for measles. This acts as a lower bound of acceptability for a new measles RDT [20240613\_tpp\_measles\_rubell\_FV\_EN] ***NOTE: TPP document used here was potentially influenced by the prelim results of this paper***
* An RDT equivalent with 90% sensitivity and specificity, and 0-day lag in result return [brownRapidDiagnosticTests2020].
* An ELISA-like perfect test with 100% sensitivity and specificity, and a 0-day test result delay. This is more accurate than is observed for current ELISA tests [hiebertEvaluationDiagnosticAccuracy2021], but it used to evaluate the theoretical best-case scenario
* An ELISA-like perfect test with 100% sensitivity and specificity, and a 14-day test result delay

The time series of “test-positive cases” is the daily count of those tested cases that return a positive test result. Thus, for each non-measles noise ([[fig-outbreak-schematic]](#fig-outbreak-schematic) a) and measles ([[fig-outbreak-schematic]](#fig-outbreak-schematic) b) time series, we have 1 time series of outbreaks ([[fig-outbreak-schematic]](#fig-outbreak-schematic) c) and 5 possible time series of test-positive cases ([[fig-outbreak-schematic]](#fig-outbreak-schematic) c) (all clinically compatible cases, plus the 4 testing scenarios), which will include false positive and negative cases resulting from imperfect diagnostic tests, that can be used to trigger outbreak alerts.



## Triggering Alerts

We define an “alert” as any consecutive string of 1 or more days where the 7-day moving average of the test-positive cases is greater than, or equal to, a pre-specified alert threshold, T. For each time series of test-positive cases, we calculate the percentage of alerts that are “correct”, defined as any overlap of 1 or more days between the alert and outbreak period ([[fig-outbreak-schematic]](#fig-outbreak-schematic) c). This is analogous to the PPV of the alert system, which is how it will be referred to for the rest of the manuscript. Note that it is possible to have multiple alerts within a single outbreak, if the 7-day moving average of test positive cases drops below the threshold, and each would be considered correct. For all outbreaks in the measles time series, we calculate the percentage that contain at least 1 alert within the outbreak’s start and end dates ([[fig-outbreak-schematic]](#fig-outbreak-schematic) b). We refer to this as the sensitivity of the alert system. We also calculate the detection delay as the time from the start of an outbreak to the start of its first alert. If the alert starts before the outbreak and continues past the start date of the outbreak, this could be considered a correct alert with a negative delay i.e., an early warning triggered by a false positive. Finally, for each time series we calculate the number of unavoidable and avoidable outbreak cases. Avoidable cases are defined as those that occur within an outbreak after a correct alert is first triggered i.e., cases that could theoretically be prevented with a perfectly effective and timely response. Unavoidable cases are the inverse: those that occur before a correct alert, or those that occur in an undetected outbreak. In practice, not all cases deemed avoidable are (due to imperfect and delays in responses), but to minimize the sensitivity of the results to the response implementation and operational constraints we are counting them as such.

We define the accuracy of the surveillance system for a given time series as the mean of the system’s sensitivity and PPV. To examine the interaction of test and surveillance systems we varied the alert threshold, between 1 and 15 cases per day, and identified the threshold T for each combination of individual test, testing rate, noise structure and magnitude that maximizes accuracy.

For each combination of diagnostic test and testing rate, the optimal alert threshold is calculated by selecting the threshold that produces the highest accuracy. Each of the 100 simulations per scenario produces an accuracy and the median accuracy of this distribution is used to determine the optimal alert threshold. We then compare testing scenarios (the combination of the diagnostic test, testing rate, and alert threshold) at the optimal threshold that is specific for the scenario. This allows for conclusions to be made about the surveillance system as a whole, instead of just single components.

# Results

The threshold that maximized outbreak detection accuracy depends on diagnostic test characteristics, the testing rate, and the structure of the non-measles noise ([[tbl-optimal-thresholds]](#tbl-optimal-thresholds)). When the average noise incidence was 8 times higher than the average measles incidence (), the optimal threshold ranged between 1 and 7 test-positive cases per day. Not surprisingly, the biggest driver of this difference was the testing rate; as a large fraction of suspected cases are tested, the optimal threshold increases monotonically for all test and noise types ([[tbl-optimal-thresholds]](#tbl-optimal-thresholds)).

The maximal attainable outbreak detection accuracy at the optimal threshold depends strongly on the structure and magnitude of the background noise. For Poisson noise, at all magnitudes, the maximum outbreak detection accuracy increases rapidly from 65% at 10% of suspected cases tested to 90% accuracy at 20% testing for all test types ([[fig-accuracy]](#fig-accuracy)). For dynamical SEIR noise, the ELISA-like tests perform similarly to the Poisson noise case at all magnitudes ([[fig-accuracy]](#fig-accuracy)), as well as to the perfect tests with complete discrimination. For RDT-like tests, which have lower individual sensitivity and specificity, the maximal attainable accuracy is lower than the ELISA-like test for all testing rates at noise magnitude ([[fig-accuracy]](#fig-accuracy)). Notably, the maximal attainable accuracy declines with increasing noise and, at all noise levels, is not improved with higher testing rates as the signal becomes increasingly dominated by false positive tests ([[fig-accuracy]](#fig-accuracy)).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | Test Characteristic | | Testing Rate | | | | | | | Noise Type | Test Type | Test Lag | 10% | 20% | 30% | 40% | 50% | 60% | | Dynamical noise: in-phase | RDT Equivalent (85.0%) | 0 | 1 | 2 | 3 | 4 | 4 | 5 | | Dynamical noise: in-phase | RDT Equivalent (90.0%) | 0 | 1 | 2 | 3 | 4 | 4 | 5 | | Poisson noise | RDT Equivalent (85.0%) | 0 | 1 | 3 | 4 | 5 | 6 | 7 | | Poisson noise | RDT Equivalent (90.0%) | 0 | 1 | 2 | 4 | 5 | 5 | 6 | | All noise structures | Perfect Test | 0 | 1 | 2 | 3 | 4 | 4 | 5 | | All noise structures | Perfect Test | 14 | 1 | 2 | 3 | 4 | 4 | 5 | |

|  |
| --- |
| image(“manuscript\_files/figure-typst/fig-accuracy-output-1.png”) |

Introducing a lag in test result reporting necessarily decreases outbreak detection accuracy because an alert can only begin once the test results are in-hand, which increases the chance that an outbreak will end before the result. For the conditions simulated here, introducing a 14-day lag in test reporting for an ELISA-like test reduces the outbreak detection accuracy by 3%. For all simulated scenarios, this is consistent with, or higher than the accuracy achievable with an RDT-like test. This always leads to an increase in the median delay from outbreak start to alert relative to an ELISA-like test with no detection and frequently leads to a detection delay relative to an RDT-like test ([[fig-delay]](#fig-delay)).

|  |
| --- |
| image(“manuscript\_files/figure-typst/fig-delay-output-1.png”) |

It is notable that outbreak detection accuracy and detection delays do not increase monotonically with an increase in testing rate, and this holds regardless of the type of test. The reason behind this unintuitive results stems from the use of integer-valued thresholds. An integer valued threshold can result in step-changes of accuracy between two threshold values, and the expected increase in the alert system’s PPV from a higher threshold value is outweighed by the loss in the alert system’s sensitivity to detecting outbreaks. Even with a perfect test, the alert system must discriminate between endemic/imported cases and epidemic cases; increasing the testing rate will result in higher numbers of test positive individuals, and a lower threshold can result in an overly sensitive alert system, triggering for measles infections, but not those within an outbreak. A higher threshold will face the opposite issue; not triggering for smaller outbreaks, which is more likely to be an issue at lower testing rates.

In general, the increase in accuracy with higher testing rates, is accompanied with longer testing delays. This reflects the change from highly sensitive systems with low thresholds to more specific systems with higher thresholds at higher testing rates. For Poisson noise, similar detection delays are observed for all test and noise magnitudes, with variation by testing rate (mean of -3.7 to 36.1 days). Under dynamical noise, there are clearer differences in the performance of ELISA and RDTs, with the separation of outcomes occurring later than observed for detection accuracy (8 time noise magnitude vs. 2 times, respectively) ([[fig-accuracy]](#fig-accuracy)). With large amounts of dynamical noise (8 times the measles incidence), the mean detection delay of the 90% and 85% RDTs range from -17.5 days to 3.2 days, and from -25.2 days to -3.4 days, respectively. Long detection delays manifest as large numbers of unavoidable cases (i.e., cases that occur between the outbreak start and its detection. Given the initial exponential rate of increase of outbreaks, the pattern of unavoidable cases follows the same shape as for detection delays, but more exaggerated. Negative delays indicate that alerts are being triggered before the start of the outbreak and is correlated with the proportion of the time series that is under alert, with larger negative delays associated with more and/or longer alert periods ([[fig-alert-proportion]](#fig-alert-proportion), Supplemental Figure 2).

|  |
| --- |
| image(“manuscript\_files/figure-typst/fig-alert-proportion-output-1.png”) |

The non-monotic relationship between testing rate and both detection delay and unavoidable cases can be explained by the step changes in sensitivity due to integer thresholds and the overall sensitivity of the system. Notably, for all noise and testing combinations, the detection delay and number of unavoidable cases is lowest at low testing rates. This occurs because the system performs best with a highly sensitive threshold (1 test positive case), which means that the system in “alert” 15-30% of the time as a results of longer and more frequent alerts ([[fig-alert-proportion]](#fig-alert-proportion), Supplemental Figures 2 & 3). Higher testing rates give rise to more a more specific outbreak detection system and a decrease in the proportion of the time series that is designated as alert. Note that for high testing rates and high ratios of dynamical noise to signal RDTs dominate the system with false positives leading to a correspondingly high proportion of the time series in alert ([[fig-alert-proportion]](#fig-alert-proportion)).

Examining the number of unavoidable cases ([[fig-unavoidable]](#fig-unavoidable)) and the detection delays ([[fig-delay]](#fig-delay)), we can see that the decrease in accuracy results from a decrease in the sensitivity of the alert system: the detection delay increases by 6 days, and the unavoidable cases by c. 2300. Although the testing rate increases, so does the optimal threshold, from 3 test positives to 4 test positives per day. The subsequent change in testing rate (40 - 50%) is associated with no change in the optimal threshold for the ELISA, and as expected, the number of unavoidable cases and detection delay decrease, resulting in a more sensitive alert system, with a higher system accuracy indicating that the increase more than offset the decrease in the system’s PPV.

|  |
| --- |
| image(“manuscript\_files/figure-typst/fig-unavoidable-output-1.png”) |

# Discussion

The performance of an outbreak detection system is highly sensitive to the structure and level of background noise in the simulation. Despite the mean daily noise incidence set to equivalent values between the dynamical and Poisson-only simulations, drastically different results are observed.

Under the assumption that non-measles febrile rash is relatively static in time (Poisson), RDTs can perform as well, if not better than ELISA tests at moderate to high testing rates, and at a fraction of the cost. However, if it is expected that the noise is non-stationary, imperfect tests cannot overcome their accuracy limitations through higher testing rates, saturating at c. 74% accuracy, relative to ELISA’s 93%. This occurs because, despite the same average incidence of noise in each (comparable) scenario, the relative proportion of measles to noise varies differently throughout the time series, exacerbating the effects of imperfect diagnostic tests that will produce higher rates of false positives and negatives than ELISA-like diagnostics. Imperfect tests are also more susceptible to incorrect results due to the varying prevalence throughout the time series, so the PPV of the system will not be static throughout the time series.

* *Make point about balance of false positive/negative outbreak detection*
* *Analysis isn’t true optimization:*
  + *Requires explicit decisions about preference for speed / false alerts vs higher PPV*
  + *Surveillance is counting for action (WHO quote)*

The purpose of routine surveillance is to characterize the infection landscape. With strong public health infrastructure and infectious disease surveillance programs, it is possible to develop a strong understanding of the shape of febrile rash cases, regardless of source. With this information, countries can tailor their future activities to rely more or less heavily upon RDTs, depending on the dynamics of the target disease and its relationship to background noise, favoring RDTs when there are low levels of noise and ELISAs during large rubella outbreaks, for example.

## Limitations and Strengths

To our knowledge, this is one of the first simulation studies to examine the relationship between individual test characteristics and the wider surveillance program. By explicitly modeling the interaction between the two, we make a case that surveillance systems should take a holistic approach; ignoring one component can lead to drastically different, and suboptimal, results. Additionally, by defining outbreak bounds concretely we have been able to calculate metrics of outbreak detection performance that draw parallels to those used when evaluating individual diagnostic tests, allowing for intuitive and simple implementation of this method in resource-constrained environments, something that is not possible with most outbreak detection and early warning system simulations in the literature. An evaluation of all outbreak detection algorithms is beyond the scope of this work, but a more computationally expensive approach based on nowcasting incidence may help overcome the shortcomings of RDTs in high-noise scenarios.

For computational simplicity, this paper did not include demography in the model structure. And while a simulation-based approach allows for complete determination of true infection status i.e., measles vs non-measles febrile rash cases, and therefore an accurate accounting of the outbreak and alert bounds, these simulations do not specifically represent any real-world setting. The evaluation of empirical data does provide this opportunity, but at the cost of knowing the true infection status of individuals and confounding of multiple variables, limiting analysis to only those who are observed (i.e., not those in the community who do not visit a healthcare center), and removing the possibility to explore the sensitivity of the results to parameters of interest to a surveillance program e.g., testing rate, and the test itself.

Additionally, is has been well documented that the performance of an individual test is highly sensitive to its timing within a person’s infection cycle [gastanaduyMeasles2019][larremoreTestSensitivitySecondary2021][middletonModelingTransmissionMitigation2023][kisslerViralDynamicsAcute2021][ratnamPerformanceIndirectImmunoglobulin2000], so it is possible that different conclusions would be drawn if temporal information about the test acquisition was included in the simulation.

Finally, the optimal threshold for a testing scenario depends heavily on the costs ascribed to incorrect actions, be that failing to detect an outbreak or incorrectly mounting a response for an outbreak that doesn’t exist. In the simulations we have weighted them equally, but it is likely that they should not be deemed equivalent: missing an outbreak leads to many thousands of cases and associated DALYs, whereas an unnecessary alert would generally launch an initial low-cost investigation for full determination of the outbreak status. This is particularly important in countries with vast heterogeneity in transmission: different weightings should be applied to higher vs. lower priority/risk regions to account for discrepancies in consequences of incorrect decisions. For example, a high priority zone could still benefit from a false alert as many high-risk healthcare regions within a country are targeted for Supplemental Immunization Activities, so a false alert would just hasten the (proactive) vaccination response.

Given these limitations, the explicit values (i.e., optimal thresholds, accuracies etc.) should be interpreted with caution, and the exact results observed in the real-world will likely be highly dependent on unseen factors, such as the proportion of measles and non-measles sources of febrile rash that seek healthcare. However, the general patterns should hold, and more importantly, the analysis framework provides a consistent and holistic approach to evaluating the trade-off between individual level tests and the alert system enacted to detect outbreaks.

# Funding

* *Something about GAVI/Gates*

This work was supported by funding from the Office of the Provost and the Clinical and Translational Science Institute, Huck Life Sciences Institute, and Social Science Research Institutes at the Pennsylvania State University. The project described was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR002014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding sources had no role in the collection, analysis, interpretation, or writing of the report.

# Acknowledgements

## Author Contributions

*Conceptualization:* CA, MJF

*Data curation:* MJF, CA

*Formal analysis:* CA, MJF

*Funding acquisition:* MJF, WM, AW

*Investigation:* CA, MJF

*Methodology:* CA, MJF

*Project administration:* MJF

*Software:* CA

*Supervision:* MJF, WM, AW, BP

*Validation:* CA, MJF

*Visualization:* CA

*Writing - original draft:* CA, MJF

*Writing - review and editing:* all authors.

## Conflicts of Interest and Financial Disclosures

The authors declare no conflicts of interest.

## Data Access, Responsibility, and Analysis

Callum Arnold and Dr. Matthew J. Ferrari had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Callum Arnold and Dr. Matthew J. Ferrari (Department of Biology, Pennsylvania State University) conducted the data analysis.

## Data Availability

All code and data for the simulations can be found at <https://github.com/arnold-c/OutbreakDetection>

1. Corresponding author. Callum R.K. Arnold. Address: Department of Biology, Pennsylvania State University, University Park, PA, USA 16802. Email: [contact@callumarnold.com](mailto:contact@callumarnold.com). [↑](#footnote-ref-20)