United States Government Accountability Office

Report to Congressional Requesters

October 2015

BIOSURVEILLANCE

DHS Should Not Pursue BioWatch Upgrades or Enhancements Until System Capabilities Are Established

GAO-16-99

October 2015

BIOSURVEILLANCE

DHS Should Not Pursue BioWatch Upgrades or Enhancements Until System Capabilities Are Established

GAO Highlights

Highlights of GAO-16-99, a report to congressional requesters

Why GAO Did This Study

DHS's BioWatch program aims to provide early indication of an aerosolized biological weapon attack. Until April 2014, DHS pursued a next-generation autonomous detection technology (Gen-3), which aimed to enable collection and analysis of air samples in less than 6 hours, unlike the current system (Gen-2), which requires manual intervention and can take up to 36 hours to detect the presence of biological pathogens. DHS is taking steps to address the capability gap that resulted from the cancellation of Gen-3 by exploring other technology upgrades and improvements to the Gen-2 system.

GAO was asked to review (1) the technical capabilities of the currently deployed BioWatch system, (2) the Gen-3 testing effort, and (3) characteristics of autonomous detection as a possible option to replace the current BioWatch system. GAO analyzed key program documents, including test plans, test results, and modeling studies. GAO assessed Gen-3 testing against best practices, reviewed relevant literature, and discussed the BioWatch program and testing efforts with key agency officials and national laboratories staff.

What GAO Recommends

GAO recommends DHS not pursue upgrades or enhancements for Gen-2 until it reliably establishes the system's current capabilities. GAO also recommends DHS incorporate best practices for testing in conducting any system upgrades. DHS generally concurred with GAO's recommendations.

View GAO-16-99. For more information, contact Timothy M. Persons at (202) 512-6412 or personst@gao.gov or Chris Currie at (404) 679-1875 or curriec@gao.gov.

What GAO Found

The Department of Homeland Security (DHS) lacks reliable information about BioWatch Gen-2's technical capabilities to detect a biological attack and therefore lacks the basis for informed cost-benefit decisions about upgrades to the system. DHS commissioned several tests of the technical performance characteristics of the current BioWatch Gen-2 system, but has not developed performance requirements that would enable it to interpret the test results and draw conclusions about the system's ability to detect attacks. Although DHS officials said that the system can detect catastrophic attacks, which they define as attacks large enough to cause 10,000 casualties, they have not specified the performance requirements necessary to reliably meet this operational objective. In the absence of performance requirements, DHS officials said computer modeling and simulation studies support their assertion. However, none of these studies were designed to incorporate test results from the Gen-2 system and comprehensively assess the system against the stated operational objective. Additionally, DHS has not prepared an analysis that combines the modeling and simulation studies with the specific Gen-2 test results to assess the system's capabilities to detect attacks. Finally, we found limitations and uncertainties in the four key tests of the Gen-2 system's performance. Because it is not possible to test the BioWatch system directly by releasing live biothreat agents into the air in operational environments, DHS relied on chamber testing and the use of simulated biothreat agents, which limit the applicability of the results. These limitations underscore the need for a full accounting of statistical and other uncertainties, without which decision makers lack a full understanding of the Gen-2 system's capability to detect attacks of defined types and sizes and cannot make informed decisions about the value of proposed upgrades.

The actions and decisions DHS made regarding the acquisition and testing of a proposed next generation of BioWatch (Gen-3) partially aligned with best practices GAO previously identified for developmental testing of threat detection systems. For example, best practices indicate that resilience testing, or testing for vulnerabilities, can help uncover problems early. While DHS took steps to help build resilience into the Gen-3 testing, future testing could be improved by using more rigorous methods to help predict performance in different operational environments. DHS canceled the Gen-3 acquisition in April 2014, but GAO identified lessons DHS could learn by applying these best practices to the proposed Gen-2 upgrades.

According to experts and practitioners, the polymerase chain reaction (PCR), which detects genetic signatures of biothreat agents, is the most mature technology to use for an autonomous detection system. DHS is considering autonomous detection as an upgrade to Gen-2, because according to DHS, it may provide benefits such as reduction in casualties or clean-up costs. But the extent of these benefits is uncertain because of several assumptions, such as the speed of response after a detection, that are largely outside of DHS's control. As a result, the effectiveness of the response—and the number of lives that could be saved—is uncertain. Further, an autonomous detection system must address several likely challenges, including minimizing possible false positive readings, meeting sensitivity requirements, and securing information technology networks.

Contents

Figures

Abbreviations:

ADE	acquisition decision event
ADM	acquisition decision memorandum
ASP	advanced spectroscopic portal
BAND (M-BAND)	Bioagent Autonomous Network Detector
BAR	BioWatch Actionable Result
BTRA	Biological Terrorism Risk Assessment
CBP	Customs and Border Protection
CDC	Centers for Disease Control and Prevention
CONOPS	concept of operations
COTS	commercial-off-the-shelf
CRP	Critical Reagents Program
DHS	Department of Homeland Security
DNA	deoxyribonucleic acid
DNDO	Domestic Nuclear Detection Office
DOD	Department of Defense
DOE	Department of Energy
FDA	Food and Drug Administration
Fp	fraction of population protected
Gen-2	Generation-2
Gen-3	Generation-3
HSPD-10	*Homeland Security Presidential Directive 10*
KPP	key performance parameter
LLNL	Lawrence Livermore National Laboratory
LRN	Laboratory Response Network
MFR	memorandum for the record
NG-ADS	Next Gen Automated Detection System
NRC	National Research Council
OHA	Office of Health Affairs
OMB	Office of Management and Budget
ORD	operational requirements document
OTA	operational test agent
OT&E	Operational Test and Evaluation
PCR	polymerase chain reaction
Pd	probability of detection
RAM	reliability, availability, maintainability
R&D	research and development
RFP	request for proposal
Sandia	Sandia National Laboratories
S&T	Science and Technology Directorate
SBInet	Secure Border Initiative Network
SME	subject matter expert

SPADA	Stakeholder Panel on Agent Detection Assays
T&E	test and evaluation
TEMP	test and evaluation master plan
TRL	technology readiness level
TRR	Technology Readiness Review
WSLAT	whole-system live agent testing

U.S. GOVERNMENT ACCOUNTABILITY OFFICE
441 G St. N.W.
Washington, DC 20548

October 23, 2015

Congressional Requesters:

The 2001 anthrax attack—in which 22 individuals contracted anthrax, 5 of whom died, from exposure to anthrax spores sent through the mail—brought new awareness of the threat posed by bioterrorism. Given the relative speed and intensity with which a biological weapon of mass destruction could affect the nation, experts and practitioners have sought to augment traditional surveillance activities with biosurveillance programs and systems to achieve early detection and warning.[1] The Department of Homeland Security's (DHS) BioWatch program, begun in 2003, was designed to provide early indication of an aerosolized biological weapon attack. Overseen by DHS's Office of Health Affairs (OHA), the BioWatch program involves a system of aerosol collectors deployed in more than 30 cities, as well as laboratory facilities and personnel to analyze samples from these collectors. The program aims to reduce the time required to recognize and characterize potentially catastrophic aerosolized attacks by monitoring for the presence of certain biological agents considered to pose high risk for an aerosolized attack.

The currently deployed BioWatch system, known as Generation-2 (Gen-2), operated on an annual budget of nearly $87 million in fiscal year 2015. Gen-2 is designed to detect the presence of biothreat agents in 12 to 36 hours. The National Academies questioned Gen-2's technical capability in 2011, including its ability to detect attacks and the reliability of results that indicate a possible attack. Although Gen-2 has been used in the field for over a decade,[2] the National Academies stated in 2011 that the rapid

[1]Traditional disease surveillance activities involve trained professionals engaged in monitoring, investigating, confirming, and reporting in an effort to further various missions including, but not limited to, detecting signs of pathogens in humans, animals, plants, food, and the environment. The *National Strategy for Biosurveillance* defines biosurveillance as the process of gathering, integrating, interpreting, and communicating essential information related to all-hazards threats or disease activity affecting human, animal, or plant health to achieve early detection and warning, contribute to overall situational awareness of the health aspects of an incident, and enable better decision making at all levels.

[2]BioWatch Gen-1 deployed in 2003. The current system, Gen-2, refers to the increased deployment of collectors to additional jurisdictions and increased indoor monitoring capability in 2005.

initial deployment of BioWatch did not allow for sufficient testing, validation, and evaluation of the system and its components.[3]

Questions have also been raised about testing conducted during the evaluation of a next-generation technology for BioWatch, known as Generation-3 (Gen-3) Phase I testing. The goal of the Gen-3 effort was an autonomous system, that is, a single device that would collect airborne particles, analyze them, and communicate the results to decision makers automatically. Such a device, known as a lab-in-a-box, would contrast with the current BioWatch system, in which samples are retrieved manually and transported to laboratories for analysis. The new technology aimed to reduce detection time, potentially generating a result in under 6 hours, and to eliminate certain labor costs. The Gen-3 acquisition was canceled in April 2014, after testing difficulties and after an analysis of alternatives was interpreted by DHS as showing any advantages of an autonomous system over the current manual system were insufficient to justify the cost of a full technology switch. Our prior work has shown that BioWatch Gen-3 was among other DHS acquisitions that, in an effort to deploy quickly, faced challenges because of testing performance.[4]

Having canceled the Gen-3 acquisition, DHS continues to rely on the Gen-2 system for early detection of an aerosolized biological attack. Some Gen-2 equipment will reach the end of its life cycle in the next year, and DHS will need to make decisions about reinvesting in the program. Further, OHA and DHS's Science and Technology Directorate (S&T) are collaborating on next steps for the BioWatch program in an attempt to address the capability gap that Gen-3 was intended to fill, as well as other technology upgrades and improvements to the Gen-2 system. Cost-benefit decisions about future investment in BioWatch will require reliable information about Gen-2's current capabilities. More broadly, as the National Academies noted in 2011, stakeholders in the biosurveillance enterprise must consider not only the extent to which early detection can

[3]See Institute of Medicine and National Research Council, *BioWatch and Public Health Surveillance* (Washington, D.C.: National Academies Press, 2011).

[4]GAO, *Secure Border Initiative: DHS Needs to Address Testing and Performance Limitations That Place Key Technology Program at Risk*, GAO-10-158 (Washington, D.C.: Jan. 29, 2010); and *Combating Nuclear Smuggling: DHS Improved Testing of Advanced Radiation Detection Portal Monitors, but Preliminary Results Show Limits of the New Technology,* GAO-09-655 (Washington, D.C.: May 21, 2009).

make a difference given the challenges of effective response and medical countermeasures, but also the trade-offs of allocating funding across the range of biosurveillance activities. In this context, you asked us to review technical aspects of the BioWatch program. This report addresses the following questions:

1. To what extent has DHS assessed the technical capability of the currently deployed system (Gen-2) to detect a biological attack, which is necessary to inform decisions about upgrades and enhancements?
2. To what extent did DHS adhere to best practices for developmental testing during Gen-3 Phase I, and what lessons can be learned as DHS considers upgrades to Gen-2?
3. Which technology is currently most mature for an autonomous detection system as a possible upgrade from Gen-2, and what would the potential benefits and likely challenges be if DHS were to pursue an autonomous detection system in the near future?

To determine the extent to which DHS has assessed the technical capability of the Gen-2 system to detect an attack, we reviewed and analyzed test reports and other agency and agency-commissioned documents containing information on the design, development, deployment, and technical performance characteristics of the system. We also reviewed reports of modeling and simulation, conducted by Department of Energy (DOE) national laboratories for DHS, that analyzed the performance and capabilities of the system. To assess the strengths and limitations of tests and studies of the Gen-2 system, we used (1) a framework for testing and evaluation of biodetection systems developed by the National Research Council;[5] (2) leading practices in risk analysis and cost-benefit analysis;[6] and (3) judgment of internal and external experts in the fields of engineering, aerobiology, microbiology, and testing and evaluation of biodetection systems.

[5]National Research Council, *Review of Testing and Evaluation Methodology for Biological Point Detectors: Final Report* (Washington, D.C.: National Academies Press, 2004).

[6]Office of Management and Budget (OMB), *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, Circular A-94 (Oct. 29, 1992). M. Granger Morgan and Max Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (New York: Cambridge University Press, 1990). OMB, *Circular A-4* (Sep. 17, 2003).

To determine whether DHS's actions during Gen-3 Phase I adhered to best practices for developmental testing and to identify lessons that could be learned, we reviewed the best practices previously developed in conjunction with the National Academies to assess their appropriateness to our review.[7] We analyzed Gen-3 Phase I acquisition and testing documents, such as the test and evaluation master plans, individual test plans and results, and the operational requirements documents (which lay out the minimum acceptable operational performance requirements). We analyzed other DHS documentation on lessons learned, including the Post Implementation Review assessment in which DHS identified its own lessons learned on the Gen-3 acquisition. We reviewed the acquisition decision memorandum (ADM) on the cancellation of the Gen-3 acquisition.[8] We also compared the steps outlined in the test planning documents with the recommended steps described in the best practices. We consulted with internal and external experts on our evaluation of DHS's actions and decisions against the best practices. We reviewed prior GAO reports on the Gen-3 acquisition and the biosurveillance enterprise. We also reviewed prior GAO work on other DHS acquisitions that met challenges during early phases of testing to draw comparisons with other DHS acquisitions that may have benefited from more robust testing guidance.

To identify the most mature technology for autonomous detection, we reviewed a report of a 2013 workshop conducted by the National Academies that assessed the state of technologies that are potentially suitable for autonomous detection for the BioWatch program.[9] We also performed a literature review of journals and conference proceedings published since 2012.[10] Our literature review was not intended to be a

[7]GAO, *Combating Nuclear Smuggling: DHS Research and Development on Radiation Detection Technology Could Be Strengthened,* GAO-15-263 (Washington, D.C.: March 6, 2015). This report publicly provided the complete list of best practices.

[8]An ADM, the official record of the acquisition decision event, describes the decisions made and any action items to be satisfied as conditions of an Acquisition Review Board decision.

[9]Institute of Medicine and National Research Council, *Technologies to Enable Autonomous Detection for BioWatch* (Washington, D.C.: National Academies Press, 2014).

[10]We defined maturity as including not just readiness but also other characteristics required for a technology to be suitable for the BioWatch mission, including sensitivity and specificity. A full definition appears in app. I.

comprehensive examination of all technologies that might possibly be applied to BioWatch, but rather a supplement to the National Academies workshop report and a check to help ensure that the conclusions of that workshop were not affected by more recent developments in the field. To assess the potential benefits and likely challenges of autonomous detection, we analyzed reports published by the Sandia National Laboratories, as well as our prior work on the Gen-3 BioWatch system. We performed a literature review for the past 12 years on models of how response timing to a positive detection of agent release may affect response effectiveness, in terms of lives saved. We interviewed officials at the Centers for Disease Control and Prevention (CDC), at DHS's Office of Health Affairs, and at Lawrence Livermore National Laboratory (LLNL) who were familiar with BioWatch and biodetection technologies. We asked for their views on the state of autonomous detection technology and the potential benefits and likely challenges of autonomous detection in the next 5 years.

For additional insight and contextual sophistication supporting our analysis across all research objectives, we interviewed DHS officials from the BioWatch Program Office and S&T who had knowledge of the history of the BioWatch program, the Gen-2 and Gen-3 technologies and changes that had been made to the Gen-2 technology over time, and the tests and studies that had been conducted on the Gen-2 and Gen-3 systems' technical capabilities. We also collected information from these officials on DHS's actions and decisions during Phase I testing and compared that with the recommended actions outlined in the best practices for developmental testing. We interviewed officials at the DOE national laboratories who conducted or were familiar with the BioWatch testing and the modeling and simulation studies, and we interviewed officials at Department of Defense (DOD) test agencies who were familiar with the Gen-2 and Gen-3 testing. To help collect and analyze information, and to help ensure the technical accuracy of our work, we consulted with subject matter experts under contract with GAO in the fields of aerobiology, microbiology, and biodetection. A more detailed discussion of our methodology appears in appendix I.

We conducted this performance audit from December 2013 to October 2015 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Background

How the BioWatch Program Works

Currently, the BioWatch program collaborates with more than 30 BioWatch jurisdictions throughout the nation to operate approximately 600 Gen-2 aerosol collectors.[11] These units rely on a vacuum-based collection system that draws air through a filter. These filters are manually collected and transported to state and local public health laboratories for analysis using a process called polymerase chain reaction (PCR). Sometimes also called molecular photocopying, PCR is a technique used to amplify (or copy) segments of deoxyribonucleic acid (DNA), the building blocks of genetic material.[12] By targeting specific segments of genetic material, PCR can be used as the basis for a test, or assay, for the presence of genetic signatures associated with specific biological organisms, such as the five BioWatch threat agents. (The program monitors for six distinct biothreat agents, but two of these are closely related, although they cause different diseases, and the BioWatch program has treated them as a single agent. For consistency, we will treat them as a single agent and report that there are five BioWatch threat agents in total.) In the BioWatch Gen-2 system, multiple PCR assays are used for each threat agent. In an initial "screening" step, one assay is run for each threat agent. If any of these assays yields a positive result, suggesting the presence of one of the threat agents, then the analysis proceeds to a "verification" step in which multiple additional assays are run targeting different genetic signatures for that agent. If the verification step also yields a positive result, then a BioWatch Actionable Result (BAR) is declared.[13] Using this manual process, the determination of a BAR can occur from 12 to 36 hours after an agent is initially captured by the aerosol collection unit. This 36-hour timeline consists of up to 24

[11]The BioWatch program is a federally-managed, locally operated system with collectors deployed primarily in outdoor locations.

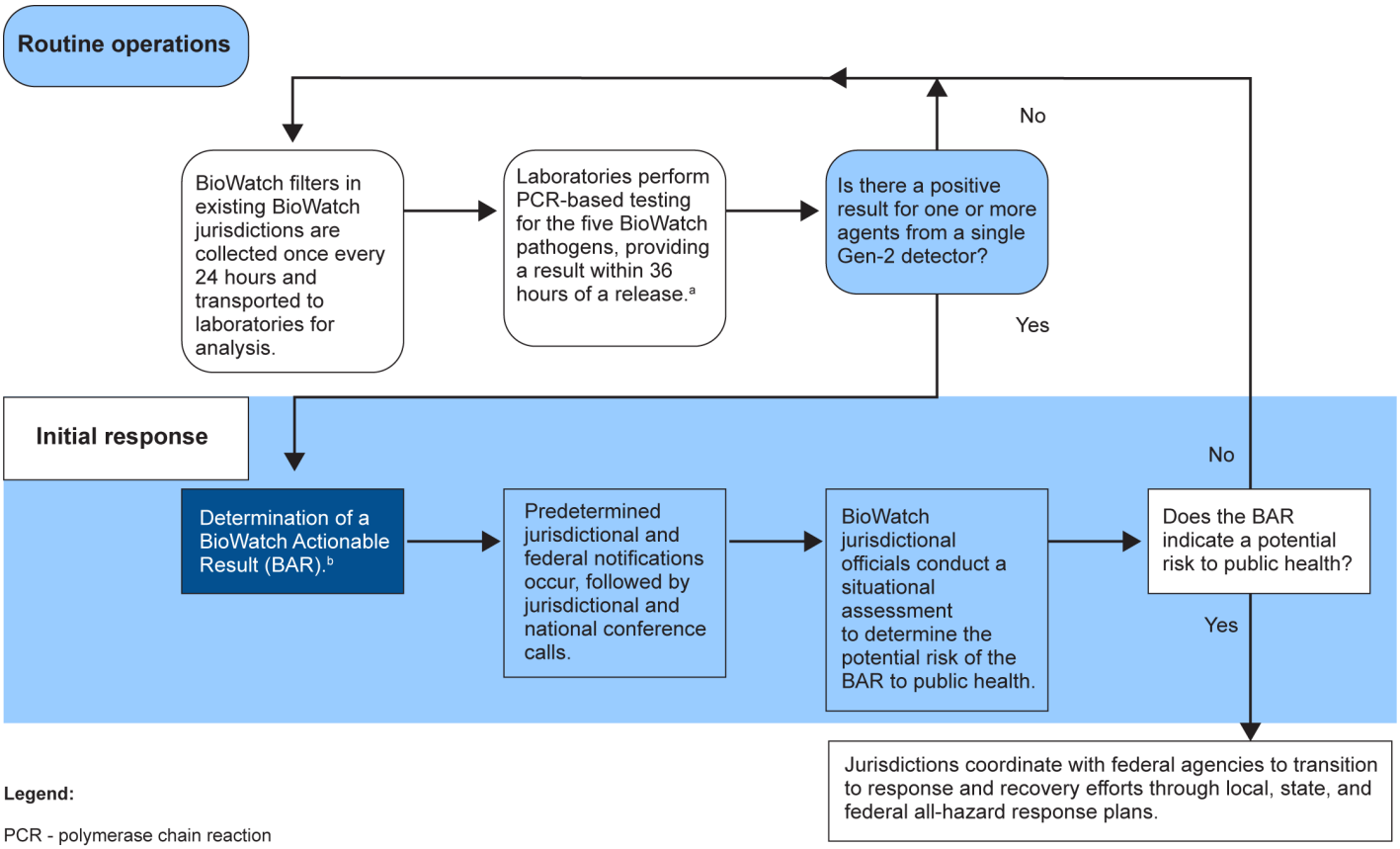
[12]To amplify a segment of DNA, the sample is heated so the DNA separates into two pieces of single-stranded DNA. Then, an enzyme builds two new strands of DNA, using the original strands as templates. This process results in the duplication of the original DNA, containing one old and one new strand of DNA. Each of these strands can be used to create two more copies. The cycle can be repeated as many as 30 or 40 times, until enough genetic material is available for analysis.

[13]The BioWatch program office defines a BAR as one or more PCR-verified positive results from a single BioWatch collector. A positive result requires multiple strands of the PCR-amplified DNA to match an algorithm that has been designed to indicate the presence of genetic material from one or more of the five agents in question.

hours for air sampling, up to 4 hours for retrieving the sample from an aerosol collection unit and transporting it to the laboratory, and up to 8 hours for laboratory testing.

Each BioWatch jurisdiction has either a BioWatch Advisory Committee or equivalent decision-making group in place, composed of public health officials, first responders, and other relevant stakeholders. The BioWatch Advisory Committee is responsible for the day-to-day BioWatch operations, including routine filter collection and laboratory analysis of filter samples. In the event of a BAR, the BioWatch Advisory Committee, in partnership with OHA and other stakeholders, is also responsible for determining whether that BAR poses a public health risk and deciding how to respond. The declaration of a BAR does not necessarily signal that a biological attack has occurred. BARs have been triggered by biological agents that occur naturally in numerous areas of the United States. From 2003 through 2014, 149 BARs were declared, but none was linked to an attack or to a public health threat. For a more detailed discussion of this issue, see appendix II. Figure 1 shows the process that local BioWatch jurisdictions are to follow when deciding how to respond to a BAR.

Figure 1: Process for Arriving at and Responding to a BioWatch Actionable Result



Legend:

PCR - polymerase chain reaction

Source: GAO analysis of BioWatch program guidance. | GAO-16-99

[a]PCR is a technique to copy deoxyribonucleic acid (DNA) for laboratory testing.

[b]The BioWatch program defines a BAR as one or more PCR-verified positive results from a single BioWatch collector. A positive result requires multiple strands of the PCR-amplified DNA to match an algorithm that has been designed to indicate the presence of genetic material from one or more of the five agents in question.

A History of BioWatch

In cooperation with other federal agencies, DHS created the BioWatch program in 2003.[14] The goal of BioWatch is to provide early warning, detection, or recognition of a biological attack. When DHS was established in 2002, a perceived urgency to deploy useful—even if immature—technologies in the face of potentially catastrophic consequences catalyzed the rapid deployment of many technologies. In the initial deployment of BioWatch—known as Generation-1—DHS deployed aerosol collectors to 20 major metropolitan areas, known as BioWatch jurisdictions, to monitor primarily outdoor spaces.[15] DHS completed the initial deployment quickly—within 80 days of the President's announcement of the BioWatch program in his 2003 State of the Union Address. To accomplish this rapid deployment, DHS adapted an existing technology that was already used for other air monitoring missions. In 2005, DHS expanded BioWatch to an additional 10 jurisdictions, for a total of more than 30. The expanded deployment—referred to as Gen-2—also included the addition of indoor monitoring capabilities in three high-threat jurisdictions and provided additional capacity for events of national significance, such as major sporting events and political conventions. The technology used in Gen-1 and Gen-2 was deployed rapidly and, according to the National Academies in 2011, without sufficient testing, validation, and evaluation of its technical capabilities.

To reduce the time required to detect biothreat agents, DHS began to develop autonomous detection capability in 2003 for the BioWatch program—known as Gen-3.[16] Envisioned as a laboratory-in-a-box, the autonomous detection system would automatically collect air samples, conduct analysis to detect the presence of biothreat agents every 4 to 6 hours, and communicate the results to public health officials via an electronic network without manual intervention. By automating the

[14]Homeland Security Presidential Directive 10 (HSPD-10) called for a national bioawareness capability providing early warning, detection, or recognition of a biological weapon attack. In response, the BioWatch program was established on January 10, 2003. BioWatch is currently managed by DHS's Office of Health Affairs. Prior to 2007, DHS's Science and Technology Directorate managed the BioWatch program.

[15]Each BioWatch jurisdiction may include various state and local government entities, such as counties or cities, or support contractors.

[16]Initially, S&T, partnering with industry, led the development of technologies to support autonomous detection. DHS's OHA has had responsibility for overseeing the acquisition of this technology since fiscal year 2007.

analysis, DHS anticipated that detection time could be reduced to 6 hours or less, making the technology more appropriate for monitoring indoor high-occupancy facilities such as transportation nodes and enabling a more rapid response to an attack. DHS also anticipated a reduction in operational costs by eliminating the program's daily manual sample retrieval and laboratory analysis. In 2008, DHS OHA initiated a competitive bid process for the first testing phase of the Gen-3 acquisition, known as Gen-3 Phase I. Five vendors responded to the request for proposal, and DHS awarded contracts to two, for technologies known as the Bioagent Autonomous Network Detector (BAND, later named M-BAND) and the Next Gen Automated Detection System (NG-ADS).

From May 2010 through June 2011, the BioWatch program conducted Phase I testing on these candidate Gen-3 technologies.[17] The testing goals included characterizing the state of available autonomous detection technology on the market and evaluating the candidate systems' abilities to meet performance requirements developed by the BioWatch program. The Phase I testing consisted of testing of individual system components, such as the aerosol sampling component (the component that collects particles from the air) and the analytical subsystem (the component that detects and identifies biothreat agents), whole system chamber testing, and an operational field test in a BioWatch jurisdiction. Characterization testing did not demonstrate the system's end-to-end ability to detect the five BioWatch threat agents in an operational environment because these agents cannot be released into the air in such environments.

Expressing concern in 2011 about the rigor of DHS's effort to help guide its Gen-3 decision making, Members of the Congress asked us to examine issues related to the Gen-3 acquisition. We released a report that evaluated the acquisition decision-making process for Gen-3 in September 2012.[18] We recommended that before continuing the Gen-3 acquisition, DHS should carry out key acquisition steps, including reevaluating the mission need and systematically analyzing alternatives

[17]NG-ADS participated in all Gen-3 Phase I test events. M-BAND participated in two test events—aerosol collection subsystem testing and assay evaluation—but did not complete all testing because it did not meet program requirements during one of the tests.

[18]GAO-12-810.

based on cost-benefit and risk information.[19] DHS subsequently commissioned an analysis of alternatives, which was interpreted by DHS as showing that any advantages of an autonomous system over the current manual system were insufficient to justify the cost of a full technology switch. DHS's April 24, 2014, ADM announced the cancellation of the Gen-3 acquisition and made Gen-2 the official program of record for aerosol biological threat detection. The ADM also directed S&T to explore development and maturation of an effective and affordable automated aerosol biodetection capability, or other operational enhancements, that meets the operational requirements of the BioWatch system.[20]

Methods and Measures for Assessing System Performance

The capabilities of the BioWatch system can be assessed at three different levels (fig. 2). At the highest level, BioWatch consists of an array of aerosol collectors deployed in an operational environment and the associated laboratory processes for analyzing samples. The operational environment might be outdoors, as in a metropolitan area (shown in the fig.); indoors, as in an airport; or a subway or other transit system. At this level, the capability of the system to detect an attack depends on factors that include the performance characteristics of the technology (including the aerosol collector and the technology used for laboratory analysis of samples), the number and locations of the collectors, the location of an attack (that is, where a biothreat agent is released into the air), and wind patterns (for an outdoor attack).

At the next level is the detector, which consists of the aerosol collector unit and the process by which samples collected by this unit are transferred to a laboratory and analyzed.[21] Performance and effectiveness at this level depend on the technical performance characteristics of the aerosol collector itself, the extent to which the

[19]The Gen-3 acquisition was in the early stages of Phase 3 (obtain the solution) when the acquisition was placed on hold.

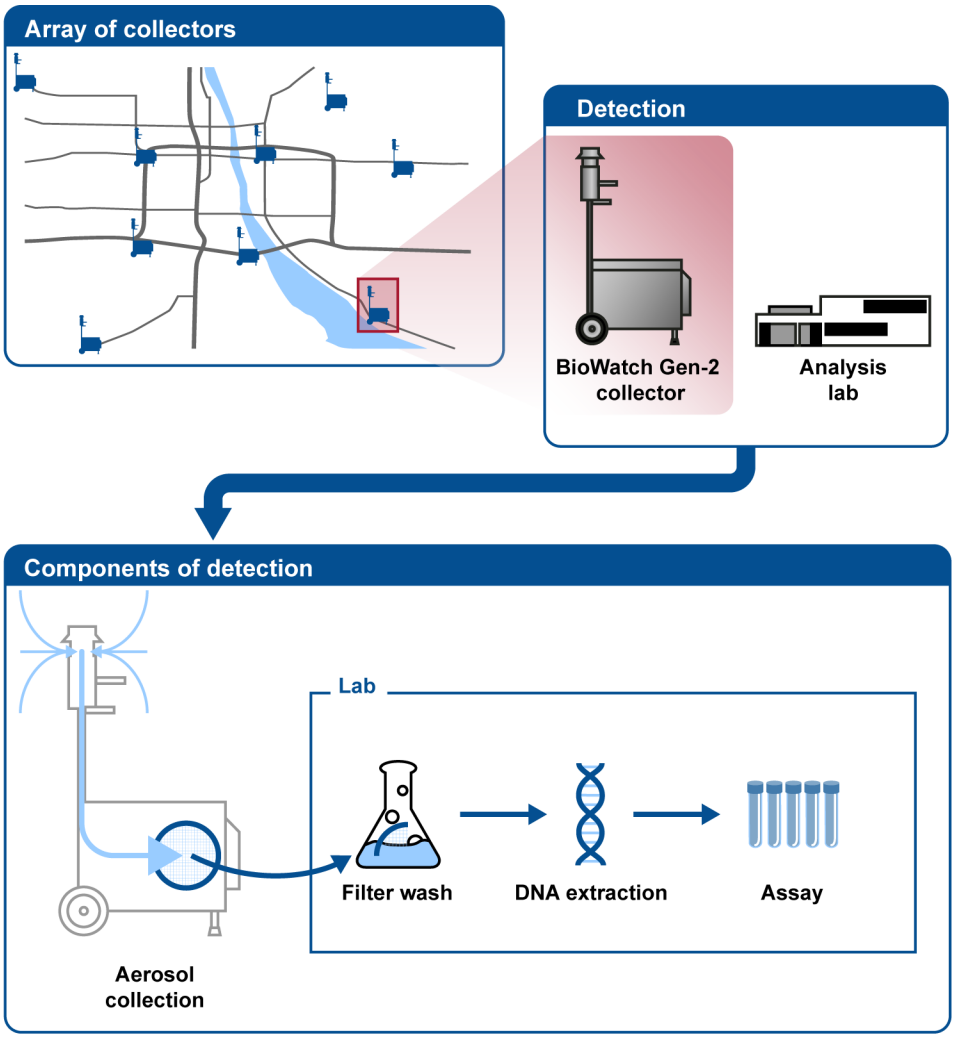
[20]DHS began to develop autonomous detection technology in 2003. Since fiscal year 2007, DHS's OHA has been responsible for overseeing the acquisition of this technology.

[21]In the BioWatch Gen-2 system, detection of a biothreat agent involves both the aerosol collection unit and the subsequent analysis of the sample, which takes place in a laboratory. For simplicity and for consistency with computational models and simulations that have been conducted of the BioWatch system, we use the term "detector" to refer to an individual aerosol collection unit and subsequent laboratory analysis of samples from this unit.

sample is preserved intact during the collection cycle and during transport to the laboratory, and the laboratory processes that are used to prepare and analyze the sample.

Finally, performance can also be assessed at the level of individual components of the detector. These include (1) the aerosol collection unit, which collects and retains aerosol particles on a filter; (2) the sample recovery process, by which samples are removed from the aerosol collector and transported to a laboratory; (3) the filter extraction process, in which aerosol particles are removed from the filter and put into a liquid solution; (4) the DNA extraction process, in which DNA is extracted from the aerosol particles in liquid solution in preparation for further analysis; and (5) the PCR assays, which are used to test for the presence of specific genetic signatures of biothreat agents (as described earlier).

Figure 2: Levels at Which BioWatch Capabilities Can Be Assessed



Source: GAO analysis/interpretation of information from DHS, Sandia National Laboratories, and the National Academies. | GAO-16-99

At the highest level—an array of detectors deployed in an operational environment—measures of performance include the system's probability of detection (*Pd*) for attacks of different types and sizes. The BioWatch program employs a variation on the *Pd* measure that is designed to assess the system's ability to detect attacks that could cause large numbers of casualties. Because BioWatch threat agents cannot be released into the air in operational environments, the performance of an

array of detectors cannot be tested directly. One method of testing that can be used to address this limitation is the use of simulants for biothreat agents. A simulant is a selected nonpathogenic organism that mimics all or some of the physical or biological characteristics of one or more pathogenic agents. Another method that has been used for BioWatch involves computer modeling and simulation of attack scenarios.

At the level of a single detector, key measures of system performance include limits of detection, probability of detection, and specificity.[22] The limits of detection are the lowest aerosol concentrations at which the system can detect the presence of a biothreat agent with a defined level of reliability. Probability of detection is the likelihood that the system will correctly detect the presence of a biothreat agent when it is present at a given aerosol concentration in the immediate vicinity of the detector. Specificity is the probability that the system will correctly yield a negative result when a biothreat agent is not present (that is, the probability that the system will not generate a "false positive"). The term specificity may also refer to the system's ability to distinguish a biothreat agent from other, similar organisms.

At this single detector level, limits of detection and probability of detection reflect the sensitivity of the system and help determine the ability of an array of detectors in an operational environment to detect attacks; all else being equal, a more sensitive system will have greater ability to detect attacks. (Note, however, that the sensitivity of a single detector reflects its ability to detect an aerosol at the location where the detector is placed; additional analysis must be done to say what this means for the ability of an array of detectors in different locations to detect attacks of defined types and sizes.) The specificity of the system may contribute to the confidence that stakeholders and decision makers have in a positive result; a system with higher specificity is less likely to generate false positives, so users can have greater confidence in a positive result.

Detector-level performance is assessed primarily through testing. Such tests may be conducted in laboratory chambers or in open air and may involve live biothreat agents or simulants. End-to-end tests using live agents are currently not possible for the BioWatch system, as the

[22]As noted earlier, we are using the term "detector" to refer to an aerosol collector unit plus subsequent laboratory analysis of samples collected by this unit.

BioWatch threat agents cannot be released in open-air environments, and at present there is no indoor chamber in which testing the Gen-2 system with live agents is technically feasible.[23] Consequently, end-to-end tests of BioWatch must rely on simulants, which may be inactivated, or killed, forms of the same agents that the system is designed to detect. Alternatively, simulants may be related organisms, either live or killed. Testing a biodetection system outdoors with killed related organisms presents the most realistic opportunity to evaluate performance in an operationally representative environment.

Individual components of a detection system, such as the aerosol collector or the assay component, may be tested under strictly defined conditions. This type of testing could support comparisons of components with live agents and simulants and provides for a tight control of test conditions and variables for a robust characterization of components. However, testing components in a laboratory or chamber setting typically excludes some factors that might affect system performance in an operational environment, such as meteorological factors and materials in the air that might interfere with system performance (called interferents). Important meteorological factors include relative humidity, temperature, and solar irradiance; important interferents include pollutants (e.g., nitrates or carbon monoxide), as well as smoke and dust, all of which can influence the performance of a biodetection system. Also, the operational environment is difficult to reproduce in a biological containment chamber in terms of the aerosol concentration, particle size distribution, aging of the agent, and dispersion dynamics. The metrics used for components of a system depend on the component being tested. For example, for an aerosol collection component, they might include efficiency (i.e., the percentage of aerosol particles successfully collected on the filter and retained intact for subsequent analysis). For laboratory tests or assays, other metrics are used, including sensitivity, limits of detection, and specificity, as well as the efficiency of the processes by which samples are removed from the filter and prepared for analysis.

[23]Whole-system live agent testing (WSLAT) would be possible in a chamber currently under construction at Dugway Proving Ground.

Biodefense, BioWatch, and the Biosurveillance Enterprise

Concerned with the threat of bioterrorism, in 2004, the White House released Homeland Security Presidential Directive 10 (HSPD-10), which outlined four pillars of the biodefense enterprise and discussed various federal efforts and responsibilities to help support it. The biodefense enterprise is the whole combination of systems at every level of government and the private sector that can contribute to protecting the nation and its citizens from potentially catastrophic effects of a biological event. It is composed of a complex collection of federal, state, local, tribal, territorial, and private resources, programs, and initiatives, designed for different purposes and dedicated to mitigating various risks, both natural and intentional. The four pillars of biodefense outlined in HSPD-10 were (1) threat awareness, (2) prevention and protection, (3) surveillance and detection, and (4) response and recovery. The BioWatch program falls under the surveillance and detection pillar, as an example of an environmental monitoring activity. Biosurveillance also includes disease monitoring and reporting to protect humans, animals, and plants from potentially catastrophic effects of intentional or natural biological events.

However, in 2011, the National Academies evaluation of BioWatch noted considerable uncertainty about the likelihood and magnitude of a biological attack, and about how the risk of a release of an aerosolized pathogen compares with risks from other potential forms of terrorism or from natural diseases. BioWatch was deployed rapidly to meet a perceived need for a system to detect catastrophic attacks. More recently, as we reported in 2012, OHA officials told us they use the Bioterrorism Risk Assessment (BTRA) to inform BioWatch because it is the most relevant risk assessment available to them and because it allows OHA to focus BioWatch detection efforts on the biological agents of significant concern. However, in 2008, the National Academies raised concerns about the methods used to develop the BTRA, particularly the methods used to assess the probability of an attack.[24] The last full BTRA was issued in 2010 and did not address all the recommendations made by the National Academies.

The National Academies' evaluation of BioWatch in 2011 also stated that to achieve its health protection goals, the BioWatch system should be better linked to a broader and more effective national biosurveillance

[24]National Research Council, *Department of Homeland Security Bioterrorism Risk Assessment: A Call for Change* (Washington, D.C.: National Academies Press, 2008).

framework that will help provide state and local public health authorities, in collaboration with the health care system, with the information they need to determine the appropriate response to a possible or confirmed attack or disease outbreak. In our earlier work, we highlighted the uncertainty about the incremental risk-mitigating benefit of the kind of environmental monitoring offered by BioWatch because of its relatively limited scope and the challenges agencies face in making investment decisions. In our June 2010 report on federal biosurveillance efforts, we recommended the Homeland Security Council direct the National Security Staff to identify a focal point to lead the development of a national biosurveillance strategy. We made this recommendation because we recognized the difficulty that decision makers and program managers in individual federal agencies face prioritizing resources to help ensure a coherent effort across a vast and dispersed interagency, intergovernmental, and intersectoral network. Therefore, we called for a strategy that would, among other things, (1) define the scope and purpose of a national capability; (2) provide goals, objectives and activities, priorities, milestones, and performance measures; and (3) assess the costs and benefits and identify resource and investment needs, including investment priorities.[25] In July 2012, the White House released the National Strategy for Biosurveillance to describe the U.S. government's approach to strengthening biosurveillance, but it did not fully meet the intent of our prior recommendations, because it did not offer a mechanism to identify resource and investment needs, including investment priorities among various biosurveillance efforts. Further, in 2005, we reported that because the nation cannot afford to protect everything against all threats, choices must be made about protection priorities given the risk and how to best allocate available resources.[26] The strategic implementation plan has not been publicly released, but according to the strategy, it will include specific actions and activity scope, designated roles and responsibilities, and a mechanism for evaluating progress. However, it is too soon to tell what effect, if any, it may have on determining resource allocation priorities across the agencies.

[25]GAO, *Biosurveillance: Efforts to Develop a National Biosurveillance Capability Need a National Strategy and a Designated Leader*, GAO-10-645 (Washington, D.C.: June 30, 2010).

[26]GAO, *21st Century Challenges: Reexamining the Base of the Federal Government*, GAO-05-325SP (Washington, D.C.: Feb. 1, 2005).

DHS Cannot Make Informed Decisions about Upgrades or Enhancements Because It Lacks Reliable Information about Gen-2's Capability to Detect an Attack

DHS lacks reliable information about BioWatch Gen-2's technical capabilities to detect a biological attack and therefore lacks the basis for informed cost-benefit decisions about possible upgrades or enhancements to the system. In order to assess Gen-2's capability to detect a biological attack, DHS would have to link test results to its conclusions about the ability of arrays of deployed detectors to detect attacks in BioWatch operational environments.[27] This would ordinarily be done by developing and validating technical performance requirements based on operational objectives, but DHS has not developed such requirements for Gen-2. In the absence of technical performance requirements, DHS officials said their assertion that the system can detect catastrophic attacks is supported by modeling and simulation studies. However, these studies have not directly and comprehensively assessed the capabilities of the Gen-2 system. Furthermore, in our review of the tests that have been conducted, we found there are limitations and uncertainties in the test results on the technical performance characteristics of the Gen-2 system.

DHS Has Not Developed Performance Requirements That Would Allow Conclusions about Gen-2's Ability to Detect Attacks

DHS commissioned four key tests of Gen-2's technical performance characteristics, but has not developed and validated performance requirements that would enable it to interpret the test results and draw conclusions about the ability of an array of detectors in an operational environment to detect attacks.[28] One test focused on the sensitivity of the whole system (that is, the aerosol collection unit and subsequent laboratory analysis of samples collected by that unit), while others focused on components of the system (table 1). None of these four tests focused on the highest level that we identified earlier—that is, an array of detectors placed in different locations. The four tests are described briefly below.

[27]As noted earlier, in the BioWatch Gen-2 system, detection of a biothreat agent involves both the aerosol collection unit and the subsequent analysis of the sample, which takes place in a laboratory. For simplicity and for consistency with computational modeling and simulation studies that we discuss in this section of the report, we use the term "detector" to refer to an individual aerosol collection unit and subsequent laboratory analysis of samples from this unit.

[28]In a later section of this report, we discuss strengths and limitations of the four tests. Here, we focus on the purposes of the tests and the extent to which DHS can interpret the results to draw conclusions about the system's capabilities.

- Dugway Proving Ground conducted a test of the sensitivity of the whole system, from the unit that collects aerosol particles on a filter through the analysis that looks for genetic material from biothreat agents. This test was designed to assess the system's ability to detect aerosols of different concentrations, and it produced estimates of the system's limits of detection (that is, the lowest aerosol concentrations that the system could detect with defined levels of reliability).

- Dugway also conducted a test of the efficiencies of particular components of the Gen-2 system—in particular, the filter wash component (where aerosol particles are recovered from the filter into a liquid solution) and the DNA extraction component (where genetic material is extracted from the aerosol particles for further analysis). Each of these components influences the overall sensitivity of the Gen-2 system.

- Edgewood Chemical Biological Center conducted a test of the aerosol collection component of the system by aerosolizing particles and measuring the system's efficiency at trapping these particles on the filter and transferring them into liquid solution, another component whose performance influences the overall sensitivity of the system.[29]

- Los Alamos National Laboratory conducted tests of the PCR assays, which included measuring the assays' sensitivity (their ability to detect different amounts of genetic material from the BioWatch threat agents), as well as their specificity (their ability to detect various strains of the BioWatch threat agents while correctly "ignoring" genetic material from other agents and interfering substances and materials commonly found in the environment).

[29]The Edgewood Chemical Biological Center test of the aerosol collection component was commissioned as a test of candidate Gen-3 systems, but included Gen-2 as a "reference system" for comparison with the Gen-3 systems.

Table 1: Four Key DHS-Commissioned Tests of the BioWatch Gen-2 System

Testing organization (year)	Parts of system tested	Performance characteristics assessed	Test environment	Agents/simulants used	Limitations
Dugway Proving Ground (2013)	Whole system	Sensitivity (limits of detection, probability of detection)	Test chamber	Killed versions of four of the five BioWatch threat agents	• Chamber not representative of operational environment; did not test environmental materials that could inhibit system sensitivity • Killed agents may not mimic live agents • Only four of five BioWatch threat agents were represented • Did not test possible decay of sample on filter during 24-hour collection cycle
Dugway Proving Ground (2013)	Filter wash, deoxyribonucleic acid (DNA) extraction	Component efficiencies, which influence system sensitivity	Test chamber	Killed versions of four of the five BioWatch threat agents	• Killed agents may not mimic live agents • Only four of five BioWatch threat agents were represented • Difficult to extrapolate to whole-system performance
Edgewood Chemical Biological Center (2010)	Aerosol collection component (aerosol collector plus filter wash)	Aerosol collection efficiency, which influences system sensitivity	Test chamber	*Bacillus globigii* (a simulant for anthrax) and synthetic particles designed to mimic biothreat agent aerosols	• Chamber not representative of operational environment • Simulant for only one of five BioWatch threat agents was used • Difficult to extrapolate to whole-system performance

Testing organization (year)	Parts of system tested	Performance characteristics assessed	Test environment	Agents/simulants used	Limitations
Los Alamos National Laboratory (multiple years)	Polymerase chain reaction (PCR) assays	Sensitivity (limits of detection), specificity	Laboratory (no aerosols used)	Purified DNA from agents, cloned DNA fragments, or DNA from near-neighbors	• Limited to the agents and environmental organisms tested, which, though based on recognized standards, may not represent all of the agents and organisms encountered in operational environments • Difficult to extrapolate to whole-system performance

Source: GAO analysis of information from DHS and testing organizations. | GAO-16-99.

In addition to these four key tests, DHS commissioned a demonstration of the system in an outdoor environment and conducts quality assurance tests on an ongoing basis. Both of these provide additional information about the system's capabilities; however, we do not include them in our list of key tests because neither was designed to produce estimates of key performance characteristics, including sensitivity, or to support conclusions about the types and sizes of attack the system can reliably detect. The outdoor demonstration, performed by the Naval Surface Warfare Center Dahlgren Division, involved releasing a simulant for one of the BioWatch threat agents and showed that the Gen-2 technology could successfully detect this simulant in an open-air environment. However, aerosol concentrations were not varied systematically and measured independently in such a way as to produce statistical estimates of the system's sensitivity. Additionally, ongoing quality assurance tests of the laboratory component of the Gen-2 system include testing filters that (1) contain potential interferents from BioWatch operational environments and (2) have been "spiked" with samples of killed biothreat agents, to verify that the system correctly detects these agents. However, these tests challenge the system with just one concentration of agent on the filter and therefore do not involve the systematic variation in concentration that is required to produce statistical estimates of the system's sensitivity. Rather than estimating the system's performance characteristics, these quality assurance tests are designed to provide confidence that system performance meets or exceeds benchmarks based on past system performance.

Under both DHS guidance and standard practice in testing and evaluation of defense systems, test results would be compared with predefined technical performance requirements. Those requirements would specify the technical performance parameters that a system must achieve in order to meet its operational objectives.[30] In other words, requirements would provide targets against which test results can be evaluated in order to assess whether the system will reliably achieve its intended purpose. Technical performance requirements for BioWatch could include the limits of detection and probability of detection that a detector needs in order for a deployed array of detectors to reliably detect attacks of particular types and sizes.

While DHS has commissioned some testing of the system's performance characteristics, officials told us they have not developed technical performance requirements, which would enable them to interpret the test results and draw conclusions about the system's ability to meet its operational objective. DHS officials told us that the system's operational objective is to detect catastrophic attacks, which they define as attacks large enough to cause 10,000 casualties, and they stated that the system is able to meet this objective.[31] However, as we have previously reported, the BioWatch system was deployed quickly in 2003 to address a perceived urgent need; it was deployed without performance requirements and, as the National Academies has reported, without sufficient testing.[32] In keeping with Office of Management and Budget (OMB) guidance on making decisions about federal programs, decisions about upgrades for BioWatch will require comprehensive information about the benefits and costs associated with the current system, including its capability to meet its operational objective.[33] However, DHS officials told us that in the 12 years since BioWatch's initial deployment, they have not developed technical performance requirements against which to measure the system's ability to meet its objective. Nevertheless, DHS has already taken steps to pursue enhancements to the Gen-2 system. Because DHS lacks targets for the current system's performance

[30]Thomas A. Cellucci, ed., *Developing Operational Requirements: Version 2.0* (Washington, D.C.: Department of Homeland Security, Nov. 2008).

[31]The term casualties refers to individuals who become ill as a result of exposure to a biothreat agent. Not all casualties necessarily result in fatalities.

[32]See GAO-12-810.

[33]OMB Circular A-94.

characteristics, including limits of detection, that would enable conclusions about the system's ability to detect attacks of defined types and sizes with specified probabilities, it also cannot ensure it has complete information to make decisions about upgrades or enhancements.

DHS-Commissioned Modeling and Simulation Studies Have Not Directly and Comprehensively Assessed Gen-2 Capabilities

In the absence of technical performance requirements for Gen-2, DHS officials said they have used modeling and simulation studies, commissioned from multiple national laboratories, to link test results to conclusions about the system's ability to detect attacks. In particular, they said that these modeling and simulation studies support their assertion that the Gen-2 system can detect catastrophic attacks, defined as attacks large enough to cause 10,000 casualties. However, while DHS officials provided reports to illustrate the modeling and simulation work that has been done, none of the studies that were provided or described to us incorporated specific test results, accounted for uncertainties in those results, and drew specific conclusions about the Gen-2 system's ability to achieve the defined operational objective. Further, according to officials, DHS has not prepared an analysis of its own that combines the modeling and simulation studies with the specific Gen-2 test results to demonstrate DHS's assertions about the system's capabilities to detect attacks of defined types and sizes.

The modeling and simulation studies were designed for purposes other than to directly and comprehensively assess Gen-2's operational capabilities. For example, one set of modeling and simulation studies, conducted by Sandia National Laboratories (Sandia) in collaboration with other national laboratories, was designed to predict the capabilities of hypothetical biodetection systems (similar to BioWatch) with different performance characteristics and deployed in different ways. For instance, these studies, which Sandia researchers called trade-space studies, assessed possible trade-offs in deploying fewer detectors with higher sensitivity or deploying more detectors with lower sensitivity. Sandia constructed models of hypothetical biodetection systems and then analyzed how these hypothetical systems would respond to simulated attacks of different sizes, using different agents, in different locations, and under different conditions (e.g., outdoor attacks with different wind speeds and directions, which affect how an aerosol disperses over an area). Because the goal was to assess hypothetical biodetection systems, Sandia analyzed ranges of hypothetical system sensitivities rather than incorporating the results of the four key tests of the performance

characteristics of Gen-2. These studies drew no conclusions about the actual capabilities of the deployed Gen-2 system.

Further, the trade-space studies did not incorporate information about the actual locations of Gen-2 collector units. Rather, these studies were designed to model hypothetical BioWatch deployments in which collectors were placed in optimal locations. If the Gen-2 collectors were not actually placed in these optimal locations, then model results might not accurately describe the capabilities of the system as currently deployed.

In addition to the trade-space studies, DHS officials described modeling and simulation work they commissioned for the purpose of selecting sites for Gen-2 collector units; however, this work also had limitations that prevent specific conclusions about the Gen-2 system's operational capabilities. Unlike the trade-space studies, the collector-siting analyses do include a test result that is meant to describe the sensitivity of the Gen-2 system. However, the test result used in this work was for just one of the five BioWatch threat agents, as decisions about collector siting are based on just this one agent. Consequently, these collector-siting analyses contain no information about the system's capabilities to detect attacks using any of the other four BioWatch threat agents. Further, the test result used in this work was not from the four key tests of Gen-2 described earlier, but from an older test from 2004. An internal DHS analysis in 2013 noted that there were differences between the system tested in 2004 and the currently deployed system that limit the ability to draw conclusions from the 2004 results. Finally, the collector-siting studies use a measure of operational capability that does not directly support conclusions about the BioWatch objective of detecting attacks large enough to cause 10,000 casualties. In general, these studies use a measure called fraction of population protected, or *Fp*. Roughly speaking, *Fp* represents a system's probability of successfully detecting simulated attacks, but calculated in a way that gives more weight to attacks that infect more people and less weight to attacks that infect fewer people. We believe this metric does not directly support conclusions about the system's ability to detect attacks causing more than 10,000 casualties. Such conclusions would be supported by another metric that has been used by Sandia but is not preferred by the BioWatch program: the probability of detection (*Pd*) for attacks causing more than 10,000 casualties.[34] DHS officials told us that they use *Fp* because BioWatch has

[34]This measure is used in limited instances in the trade-space studies described earlier.

a public health mission and so the system should be assessed in a way that reflects its ability to detect attacks that infect more people. However, *Pd* for attacks causing more than 10,000 casualties also incorporates public health impact; unlike *Fp*, it could directly support conclusions about the BioWatch operational objective, and, as noted in a Sandia report, is straightforward to communicate. Sandia officials told us that *Fp* has certain strengths and is appropriate for certain purposes. However, because the collector-siting studies focus on *Fp*, their results are not straightforward to communicate and do not support conclusions that align directly with the BioWatch operational objective.[35]

Finally, because none of the modeling and simulation work was designed to interpret Gen-2 test results and comprehensively assess the capabilities of the Gen-2 system, none of these studies has provided a full accounting of statistical and other uncertainties—meaning decision makers have no means of understanding the precision or confidence in what is known about system capabilities. Best practices in risk analysis and cost-benefit analysis require an explicit accounting of uncertainties so that decision makers can grasp the reliability of, and precision in, estimates to be used for decision making.[36] Estimates of the Gen-2 system's limits of detection, produced by the four key tests described earlier, contain multiple sources of uncertainty, which we describe in the next section of this report. None of the modeling and simulation studies that were provided or described to us incorporated information about the uncertainties associated with estimates of the system's limits of detection.[37]

We also found that these studies did not account for uncertainty in some model inputs and assumptions, including estimates of how infectious

[35]Unlike the trade-space studies described in previous paragraphs, which were provided to us as a set of published reports, the modeling and simulation work done to select collector locations is described in unpublished reports, spreadsheets, and other documents. Consequently, we were unable to perform a comprehensive review of this work, but rather reviewed illustrative documents provided to us by DHS officials and relied upon officials' descriptions of the scope and methods of this work.

[36]See Morgan and Henrion, *Uncertainty,* OMB Circular A-94, and OMB Circular A-4.

[37]Sandia officials noted that in the trade-space studies they considered ranges of possible system sensitivities, an approach that can be used to address uncertainty in model inputs. However, as noted earlier, these studies did not incorporate test results and drew no specific conclusions about the capabilities of the actual Gen-2 system.

each of the BioWatch threat agents is and how quickly each agent decays after it is released in the air.[38] For example, Sandia researchers and a subject matter expert told us that there is considerable uncertainty in even the best available estimates of the infectious dose of anthrax, as these estimates are based on data from nonhuman primates. In an earlier study, Sandia researchers and others reported that "gaps in our knowledge of the correct dose-response relationship significantly limit our ability to predict the outcome of outdoor anthrax attacks."[39] For many of the assumptions the Sandia models used, researchers dealt with uncertainty by using not just single estimates but ranges of estimates. However, for the infectious doses of the five BioWatch threat agents, researchers used single estimates that DHS provided. The uncertainty in these estimates is important because DHS officials have characterized the operational objective of the BioWatch system as detecting attacks large enough to cause 10,000 casualties. In order to assess the system's capability to achieve this objective, DHS must be able to correctly define the types and sizes of attack that fall into this category. If anthrax were less infectious than the models assumed, then DHS would be underestimating the system's ability to detect catastrophic anthrax attacks.[40] Conversely, if anthrax were more infectious than the models assumed, then DHS would be overestimating the system's ability to detect such attacks. We recognize that more precise infectious dose estimates may not exist, but this underscores the uncertainty in the ability of the BioWatch system to meet its operational objective—uncertainty that should be articulated to better inform decision makers about the capabilities of the Gen-2 system and inform cost-benefit decisions about any possible enhancements to the system.

[38]While the studies did not analyze uncertainty in agent decay rates, they did incorporate different decay rates for daytime and nighttime attacks.

[39]Sandia National Laboratories, Lawrence Livermore National Laboratory, and Washington Institute, *Catastrophic Bioterrorism Scenarios: Response Architectures and Technology Implications* (prepared for the Department of Homeland Security, Mar. 2006), 19.

[40]For example, if anthrax were less infectious than the models assumed, then greater releases of aerosolized anthrax would be required to cause 10,000 casualties. Since greater releases are easier to detect (i.e., require a less sensitive system), BioWatch would have a greater probability of detecting these attacks than predicted by the models. Conversely, if anthrax were more infectious than the models assumed, then smaller releases would have the potential to cause 10,000 casualties; detecting these smaller releases would require a more sensitive system, and BioWatch would have a lower probability of successfully detecting these attacks than predicted by the models.

Limitations and Uncertainties in Test Results

We found limitations and uncertainties in the four key tests of the Gen-2 system's performance characteristics—in particular, in the use of test chambers instead of operational environments and the use of simulants in place of live biothreat agents. As noted earlier, it is not possible to test the BioWatch system directly by releasing live biothreat agents into the air in operational environments. Because of this constraint, which is beyond DHS's control, the agency commissioned tests that involved aerosols in test chambers or were limited to components of the system for which aerosols were not necessary. Further, officials and experts told us there are no test chambers where testing the Gen-2 system with live agent is technically feasible; thus some tests have involved simulants in place of live biothreat agents.[41]

Using laboratory chambers and simulants effectively addressed certain challenges, but both introduced uncertainties into testing results. Chambers often differ from operational environments in ways that can affect a system's performance. For example, chamber environments are generally not designed to be representative of operational environments in such factors as air temperature, humidity, and, according to an expert, the presence of potential interferents in the air. Similarly, simulants may not mimic the biothreat agents that the system is designed to detect in all of the ways that matter for system performance; therefore, the system might perform differently when presented with the target biothreat agents than when tested with simulants. As a result, chambers and simulants create uncertainty as to whether test results accurately describe how the system would perform in an operational environment against live BioWatch threat agents, and this uncertainty should be clearly articulated for decision makers.

Additionally, while one of the four tests assessed the performance of the whole Gen-2 system, the three other tests were limited by their focus on components of the system, including (1) the aerosol collection component; (2) the filter wash process, in which aerosol particles are transferred from the filter into a liquid solution; (3) the DNA extraction process, in which genetic material is extracted from the particles in liquid solution; and (4) the analytical component, in which PCR assays are used

[41]Officials at Dugway Proving Ground told us that currently no chamber is large enough to contain both the BioWatch collector unit and other equipment needed for testing while also offering sufficient containment for the BioWatch threat agents. Such a chamber, known as a WSLAT facility, is currently under construction at Dugway.

to detect genetic signatures of the BioWatch threat agents. According to a National Research Council (NRC) committee, it is uncertain whether test results from individual components of a biodetection system will accurately reflect the performance of the whole system.[42] An expert told us that components may perform differently when combined than when tested separately, and parts of samples may be lost during transitions from one component to another in ways that affect the end-to-end performance of the system.

DHS took steps to mitigate the limitations associated with not testing the Gen-2 system in an operational environment with live biothreat agents, but these limitations could not be eliminated entirely. For example, to address the fact that killed agents might not perfectly mimic live biothreat agents, the Dugway tests included a direct comparison of live and killed agents, but this could be done only for the analytical component of the system (that is, the PCR assays). The Edgewood Chemical Biological Center test of the aerosol collection component included variations in temperature and humidity. This somewhat mitigated the fact that chambers may not be representative of operational environments; however, only a small number of combinations of temperature and humidity were tested, and an expert told us other characteristics that might differ between chambers and operational environments were not varied. The Los Alamos National Laboratory tests of the PCR assays included testing the assays with a set of environmental organisms and substances. However, this test was limited to the specific organisms and substances used, and results do not generalize to other organisms and substances that might occur in BioWatch operational environments.[43] In sum, although the key tests of the Gen-2 system took steps to mitigate limitations, uncertainties remain, and test results constitute only limited measures of key system performance parameters in an operational environment.

While challenges associated with testing a system like BioWatch make some limitations unavoidable, according to experts and agency officials,

[42]National Research Council, *Review of Testing and Evaluation Methodology for Biological Point Detectors.*

[43]The Los Alamos tests of the PCR assays were designed around recommendations developed by the Stakeholder Panel on Agent Detection Assays (SPADA). The environmental organisms and substances used in the test were specified in the SPADA recommendations.

we found that some limitations could likely have been mitigated. In 2004, an NRC committee proposed a framework for testing biodetection systems that was designed to minimize the uncertainties associated with laboratory chambers and simulants. In this framework, both the whole system and its components are systematically tested with simulants and, where possible, live biothreat agents. This is done in both laboratory chambers and environments that are more representative of operational environments where the system will be deployed. Importantly, this framework entails an integrated, systematic approach to testing in which some factors (e.g., agents, simulants, and environmental factors) are held constant while others are varied. The committee recommended focusing on a certain category of simulants, known as killed related strains, that have potential to mimic live biothreat agents while also enabling more realistic and frequent testing. The overall goal of the NRC framework is to home in on the true performance of the system when challenged with live agents in an operational environment—and, in so doing, to reduce the risk that the system will perform differently in the real world than it did during testing.

DHS has not systematically tested the Gen-2 system under the most realistic possible conditions. Although DHS officials said they based their approach to testing Gen-2 on the NRC framework, we found that the Dugway test of system sensitivity did not incorporate killed related strains as simulants, as recommended by the NRC. Killed related strains would offer greater flexibility for use in more operationally representative environments. Furthermore, the Dugway test did not attempt to incorporate potential environmental interferents. As noted earlier, DHS also commissioned a demonstration that the Gen-2 technology could detect a simulant in an outdoor environment. However, aerosol concentrations were not varied systematically and measured independently in such a way as to produce statistical estimates of the system's sensitivity. Furthermore, this open-air demonstration involved a simulant for just one of the five BioWatch threat agents, and, unlike the killed related strains recommended by the NRC, this simulant required that the system be modified to use a different PCR assay than is used for the actual threat agent. A GAO subject matter expert on outdoor testing of biodetection systems with simulants assessed the Dahlgren trials to be deficient in the equipment that was used to characterize the aerosols; if alternative equipment known as slit-to-agar samplers had been used, they could have provided more useful information on aerosol concentrations and exposure times. According to this expert, additional problems were associated with an inadequate dissemination system and inadvisable testing at wind speeds below 2.0 meters/second. In general,

DHS's understanding of the performance characteristics of the current system would benefit from a more systematic approach to testing under more realistic conditions.

In the next year, some Gen-2 equipment will reach the end of its lifecycle, and DHS will need to make decisions about reinvesting in the program. Further, DHS officials told us they are considering potential improvements or upgrades to the Gen-2 system. Based on OMB guidance, cost-benefit decisions about investing in new equipment and in potential system improvements will require information about current operational capabilities.[44] However, because of the limitations we have identified, decision makers are not assured of having sufficient information to ensure future investments are actually addressing a capability gap not met by the current system. As noted earlier, some limitations in testing are unavoidable given the nature and purpose of the BioWatch system. Likewise, some limitations and uncertainties are unavoidable in the modeling and simulation work that DHS has commissioned to link test results to operational capabilities (e.g., the uncertainty in infectious dose estimates for the BioWatch threat agents). These limitations underscore the need for a full accounting of statistical and other uncertainties, without which decision makers lack an understanding of the precision in what is known about the system's capability to detect attacks of defined types and sizes and cannot make informed decisions about possible upgrades to Gen-2.

[44]OMB Circular A-94.

Gen-3 Testing Partially Aligned with Best Practices, and Gen-2 Upgrades Could Benefit by Applying Lessons Learned

In 2013, in collaboration with the National Academies, we identified eight best practices for developmental testing of threat detection systems.[45] When comparing DHS's actions and decisions regarding the planned acquisition and testing of Gen-3, we found that DHS's actions partially aligned with these best practices. We also identified several lessons DHS could learn by applying these practices more systematically to improve future testing and acquisition efforts—for example, testing of possible upgrades or enhancements to Gen-2. We also highlight the testing of other DHS acquisitions that faced challenges.

Best Practices and Lessons Learned from Gen-3 Phase I Testing

Role of Developmental Testing

Developmental testing is intended to assist in identifying system performance, capabilities, limitations, and safety issues to help reduce design and programmatic risks. According to experts recruited in coordination with the National Academies, the best practices they identified apply to the process of developmental testing of binary threat detection systems; they also apply if the tested system is commercial-off-the-shelf (COTS), modified COTS, or newly developed for a specific threat detection purpose by a vendor or the government.

Source: GAO. | GAO-16-99

In previous work, in collaboration with the National Academies, we recruited experts to develop best practices for developmental testing of binary threat detection systems.[46] According to the experts, the best practices apply if the system is commercial-off-the-shelf (COTS), modified COTS, or newly developed for a specific threat detection purpose by a vendor or the government. In discussing the role of testing early in an acquisition, officials from S&T's Office of Operational Test and Evaluation (OT&E) said programs face a significant challenge in acquiring COTS products, like the BioWatch Gen-3 acquisition, because these products or systems are not designed to operate or initially tested with the same requirements needed by the federal government. Typically, OT&E officials said, program officials underestimate the testing needed to acquire a COTS product for their mission, but said testing is needed to ensure the COTS solution is reliable, scalable, and secure. We reported in 2012 that according to BioWatch program officials, developing autonomous biological detection had proved challenging, in part because some of the technology required was novel, but also because even the existing technologies—for example, the aerosol collection unit and the apparatus that reads the PCR results—had not been combined for this specific application in an operational environment before. S&T OT&E officials reflecting on the acquisition said Gen-3 Phase I testing became developmental in nature, with additional steps built into the test design to

[45]GAO-15-263. This report publicly provided the complete list of best practices. We facilitated experts' identification of best practices with pre-meeting interviews, structured questioning during the meeting, and post-meeting expert voting and ranking procedures. For more information on the methods used to develop the best practices, see app. I.

[46]Binary threat detection systems indicate whether a potential threat is present or not. They do not identify gradations of threat.

ensure the technology could progress to the next level, particularly to ensure assay detection met the program's requirements.[47]

We consider the best practices for developmental testing applicable to Phase I of the Gen-3 acquisition because they could have helped identify challenges early in the acquisition process. Additionally, applying the best practices and lessons learned during the Gen-3 testing could help mitigate the risk that DHS acquires immature technology as part of its effort to make enhancements to Gen-2. Appendix I has more information on our methodology for developing the best practices, and appendix III has a description of each best practice. Our analysis of DHS's alignment with the eight best practices for developmental testing of Gen-3 follows below and is summarized in table 2.

Table 2: Assessment of DHS's Alignment with Best Practices for Developmental Testing during Gen-3 Testing

Best practice for developmental testing	Assessment
Practice 1: Ensure that accountability and engagement in developmental testing are commensurate with the amount of risk accepted	◐
Practice 2: Include representatives from the user community in design and developmental testing teams to ensure acceptance of the system by the user community	○
Practice 3: Take a proper systems engineering view of the system prior to entering into any developmental test	◐
Practice 4: Use statistical experimenalt design methodology to establish a solid foundation for developmental testing	◐
Practice 5: Measure and characterize system performance with established procedures, methods, and metrics	◐
Practice 6: Test to build in resilience, especially in the development stages	◐
Practice 7: Use developmental tests to refine requirements	◐
Practice 8: Engage in a continuous cycle of improvement by (1) conducting developmental testing, (2) conducting operational testing, and (3) incorporating lessons learned	●

● – Fully aligned with the best practice ◐ – Partially aligned with the best practice ○ – Did not align with the best practice

Source: GAO analysis. GAO-16-99

[47]Our prior work evaluating the Gen-3 acquisition process used OHA's preferred term "characterization testing" to describe the Phase I testing for Gen-3. According to OHA officials, they did not consider the testing of the technology assessed under Phase I of the Gen-3 acquisition to be developmental testing, since they believe developmental testing is related to research and development (R&D) under the purview of DHS S&T.

Practice 1: Ensure That Accountability and Engagement in Developmental Testing Are Commensurate with the Amount of Risk Accepted

DHS's Actions Partially Aligned with Best Practice 1

DHS took some risk-mitigating steps during Gen-3 Phase I, but did not conduct a full risk assessment at the outset of the acquisition.[48] According to program officials, Phase I of the overall Gen-3 acquisition was itself a risk mitigation activity, designed to assess the capability of industry to provide mature autonomous detection solutions before committing to a rapid and extensive program to procure and field operational systems. For example, according to BioWatch program officials, DHS conducted market research to assess whether solutions potentially capable of meeting DHS performance and maturity (and therefore schedule) requirements existed.[49] In an effort to ensure accountability, OHA held a "Demonstration Day" early in the Phase I source selection process, where vendors participated in tests designed to confirm their technical maturity claims. This reduced (but did not eliminate) the risk of awarding Phase I contracts to a vendor not capable of completing the tests planned under Phase I. Additionally, officials said OHA initially planned to conduct all testing in parallel. However, because of schedule slips associated with contract issues, program officials ended up scheduling the tests incrementally to allow for the insertion of decision points on whether testing should continue. This allowed program officials to engage in the testing and evaluate test results to help reduce technical and financial risk.[50] As a result of this multi-stage Phase I testing approach, DHS identified limitations to one vendor's detection system that could not be overcome before proceeding with the next stages of testing.

DHS Could Apply Lessons Learned

An earlier evaluation of risk may have eliminated difficulties with cost and schedule estimates of the Gen-3 acquisition that in part led to its cancellation in April 2014. According to this best practice, risk needs to be assessed when the system is COTS, modified COTS, or a newly

[48]Risk refers to the technical risk that a system will provide the required performance in the required time frame utilizing specified resources.

[49]The market research identified two potential vendors potentially capable of providing viable solutions, which in turn informed the Phase I Request for Proposal (RFP).

[50]According to OHA officials, the planned acquisition decision event (ADE)-2B between Phases I and II allowed DHS to make an informed decision on whether to commit to the larger acquisition program or "return to the drawing board."

developed system, because even with commercial items, significant modifications may be needed. DHS ultimately performed a formal risk assessment of the Gen-3 acquisition but not until after Phase I testing ended. The absence of a risk assessment at the start of the acquisition led to challenges that the acquisition could not overcome because of its inflexibility. In 2012, we reported that DHS did not fully engage in the early phases of its acquisition framework to ensure that the acquisition pursued an optimal solution in the context of its costs, benefits, and risks.[51] Our prior work has found that stable parameters for performance, cost, and schedule are among the factors that are important for successfully delivering capabilities within cost and schedule expectations. Our work has also found that without the development, review, and approval of key acquisition documents, agencies are at risk of having poorly defined requirements that can negatively affect program performance and contribute to increased costs.[52] For example, despite having limited assurance that the acquisition would successfully deliver the intended capability within cost and on schedule, the Deputy Secretary approved the initial stages of the acquisition. DHS's Post Implementation Review Report, which lays out lessons learned on the Gen-3 acquisition, states that the Phase I testing demonstrated that schedule risk analyses should be used to set realistic test and evaluation schedule expectations. According to program officials, they set the acquisition schedule estimate aggressively because there was pressure to quickly deploy an autonomous detection capability. However, schedule revisions were

[51]GAO-12-810.

[52]GAO, *Department of Homeland Security: Assessments of Selected Complex Acquisitions,* GAO 10-588SP (Washington, D.C.: June 30, 2010), and *Homeland Security: DHS and TSA Face Challenges Overseeing Acquisition of Screening Technologies,* GAO-12-644T (Washington, D.C.: May 9, 2012).

needed because of significant changes in performance, deployment schedule, and cost expectations as a result of the Phase I testing.[53]

By not engaging in the initial steps of the acquisition framework to effectively account for risks early in the acquisition, DHS did not demonstrate full accountability and exceeded its cost and timeframe estimates. As a result of the Phase I testing, and on the basis of outside reviews of the Gen-3 acquisition, DHS directed OHA to conduct a more robust analysis of alternatives that included assessing risk. This led to the cancellation of the Gen-3 acquisition in April 2014. The identification of potential risks, and strategies to overcome these risks, helps ensure accountability on the part of the agency and may alleviate problems with the acquisition of threat detection systems if they are part of the early planning for testing. As DHS considers upgrades to the currently deployed Gen-2 BioWatch system or considers the acquisition of new detection technologies, early identification of risks may help better guide DHS by identifying areas for enhanced engagement that may be needed during the testing and to help ensure proper accountability for decision making during the acquisition.

[53]We reported in September 2012 that multiple officials in various DHS offices who had knowledge of Gen-3 in this early decision-making period described a climate, in the wake of the September 11, 2001, terrorist attacks and the subsequent Amerithrax attacks, in which the highest levels of the administration expressed interest in quickly deploying the early-generation BioWatch technology and subsequently improving the functionality of this technology—as quickly as possible—to allow for faster detection and an indoor capability. On the basis of this interest, officials from the multiple DHS offices said it was their understanding that the administration and departmental leadership had already determined that the existing BioWatch technology would need to be expanded and entirely replaced with an autonomous solution well before the acquisition was approved. GAO-12-810.

Practice 2: Include Representatives from the User Community in Design and Developmental Testing Teams to Ensure Acceptance of the System by the User Community

Acquisition Decision Event (ADE) 2A

An acquisition decision event, where the Acquisition Review Board—a cross-component board of senior DHS officials—determines whether a proposed acquisition has met the requirements of the relevant Acquisition Life-Cycle Framework phase and is able to proceed. ADE-2A is the culminating event for the Analyze/Select phase of the DHS acquisition framework, where DHS determines whether to authorize the acquisition to proceed to the Obtain phase, where testing and evaluation occur.

Source: GAO analysis of Department of Homeland Security (DHS) information. | GAO-16-99

DHS's Actions Did Not Align with Best Practice 2

DHS did not sufficiently involve the end user community in the development of Gen-3 system requirements or parts of Phase I testing; rather it relied on internal subject matter experts to develop requirements.[54] As we reported in 2012, the process used to set the sensitivity requirement did not reflect stakeholder consensus about how to balance mission needs with technological capabilities. Specifically, the BioWatch program did not prepare a concept of operations (CONOPS) before ADE-2A.[55] According to DHS acquisitions guidance, in developing a concept of operations, stakeholders engage in a consensus-building process regarding how to balance technological capabilities with mission needs in order to gain consensus on the use, capabilities, and benefits of a system. For BioWatch, this could include specifying the level of population protection the system should provide and then specifying the sensitivity levels needed to provide that level of protection. According to OHA officials, the high-level capability gaps documented in the mission need statement (timeliness, population coverage, time resolution, and cost-effectiveness) were representative of feedback from this user community with respect to improvements needed on the Gen-2 system, particularly for indoor deployment. Therefore, officials said they did not directly involve jurisdictional public health stakeholders in establishing the technical requirements, including sensitivity requirements, for Gen-3.

[54]OHA relied on the public health and preparedness subject matter experts they employed and other program office staff, including U.S. Public Health Service Officers and contractor support staff with direct local public health experience.

[55]According to DHS acquisition guidance, the CONOPS process is used to gain consensus among stakeholders on the uses, operating and support concepts, employment, capabilities, and benefits of an asset, capability, or system. To achieve consensus, stakeholders must collaboratively balance the desires of mission success against the realities of technology, budget, schedule and risk. The CONOPS focuses on the performance of solutions in their intended operational setting and depicts how the preferred solution would operate in the context of a real-world scenario.

Role of Subject Matter Experts (SMEs) in Developmental Testing

Distinct from the user community, SMEs provide independent technical advice and monitor the status of developmental testing to help ensure tests are conducted and analyzed properly—a practice the National Academies has supported for BioWatch. SMEs may include test designers, engineers, and statisticians. For a program like BioWatch, SMEs may have expertise in epidemiology, environmental health, public health laboratory systems, infectious diseases, genetics, and detection technology, among other disciplines. Department of Homeland Security (DHS) acquisitions policy includes guidance on using independent technical advisors as part of the test and evaluation process. For example, an operational test agent (OTA) is an independent entity that supports development of the test and evaluation master plan (TEMP) and monitors developmental testing in order to understand system performance early and determine how to execute integrated developmental and operational testing. The OTA presents objective and unbiased conclusions in reports and test readiness reviews. DHS guidance also describes the role of the Integrated Process Team, which is composed of representatives from program leadership, stakeholders, and SMEs involved in testing activities.

Source: GAO. | GAO-16-99

OHA relied on in-house experts and other department officials who they said had the expertise to convert the high-level mission needs into detailed technical performance requirements. These included requirements for system sensitivity, time needed to detect an attack, and the probability of false positives, among other things. According to OHA officials, jurisdictions (especially the four that were selected for OT&E) were kept informed through meetings with the program office, independent test agencies, and the two competing vendors in the Phase I testing. OHA officials also said they conducted numerous updates at events such as the National BioWatch Stakeholders Workshop, webinars, and invitations to all jurisdictions to submit questions to the Gen-3 or BioWatch program managers. According to OHA officials, the end user community was also invited to observe testing at two special testing events during Phase I. However, informing end users of the status of testing is not the same as including them in the development of requirements and testing. If DHS had involved end users as part of the testing process, it could have shed additional light on the potential challenges for end users to operate the technology being considered. For example, an official who tested the NG-ADS system described months-long training that was required to understand the systems being tested and evaluated. More closely involving the end users in the testing may have revealed additional end user views on their ability and willingness to use the equipment, given its complexity. Additionally, in December 2010, the Undersecretary for Management issued an ADM that, among other things, highlighted the continued need to develop a CONOPS, citing significant risk to the program because of the high-level coordination required for acting upon detection information produced by the autonomous detection system and that there was insufficient detail in the program documentation describing the necessary coordination process among end users and other stakeholders.

DHS Could Apply Lessons Learned

DHS recognized in its Post Implementation Review that stakeholders and end users need to be involved earlier in the acquisition process, including validating advanced detection systems and methods before they are fielded. To better understand the needs, concerns, and capabilities of the user community, in the future, DHS could take steps to engage with stakeholders early on in the development process. As DHS considers upgrades to the current Gen-2 system, DHS should, in accordance with DHS acquisition guidance, prepare a concept of operations and ensure end users are engaged throughout the testing of upgrades or enhancements to the Gen-2 system or new acquisitions for BioWatch.

Practice 3: Take a Systems Engineering View of the System prior to Entering into Any Developmental Test

DHS's Actions Partially Aligned with Best Practice 3

The systems engineering approach was outlined in the Gen-3 test and evaluation master plan (TEMP), which was generally clear in defining the boundaries of the system to be tested. The tests were for the most part appropriately scoped given the systems engineering view that was taken. However, testing revealed that evaluation of these detection systems may benefit from more robust testing methods, particularly to test performance against environmental contaminants. The primary purpose of the Phase I testing was characterizing the systems' performance through a series of tests that included an aerosol collection subsystem test, evaluation of assays, an analytical subsystem test, a system chamber test, and a field test. The TEMP clearly identified these test boundaries and stated that all technologies must meet the five key performance parameters (KPP) during Phase I testing before selection would be made and Phase II testing (operational testing) was initiated. KPPs included detection of the biological agents, system sensitivity, the time to detect, achieved availability (for example, the probability the detector will operate under normal conditions), and the probability of a false positive.

The TEMP recognized inherent limitations to the systems engineering testing for the Gen-3 system. For example, whereas the environmental conditions under which the Gen-3 system must operate were outlined in the TEMP, no chamber yet exists in which these requirements can be fully tested. In addition, according to DHS officials, legal constraints, public safety, and public perception limit the type of material that can be aerosolized in a realistic setting for test purposes. The TEMP outlined the systems engineering approach for Gen-3 testing by articulating the major issues that needed to be addressed in testing the system, including the key performance parameters, and accounted for the limitations given constraints on the type of testing that can be done.

DHS Could Apply Lessons Learned

Although DHS clearly articulated the boundaries of the testing for Phase I and took steps to test the autonomous detectors against likely environmental contaminants that might interfere with detection, the systems engineering approach could have benefited from more robust and comprehensive testing methods. For example, tests at Dugway also attempted to account for the possibility that environmental pollutants might interfere with the performance of the PCR assays by placing samples in liquid solutions taken from actual BioWatch filters retrieved from operational environments. However, researchers told us they were

unaware of which operational environments the filter washes had come from, and there was no sub-analysis by type of environment (e.g., outdoor versus indoor versus subway), which raises the possibility that pollutants from certain environments may not have been represented in the test or that pollutants from certain environments may have been diluted with filter wash from other environments. The pooled filter wash was used to test the Gen-3 analytical process in order to look for possible inhibition (e.g., interference) of the PCR assays.[56] Officials said that during the Gen-3 testing, the pooled filter wash used to test the analytical process never showed interference in acquiring a result to an extent that would have required additional testing steps.

However, in the final Phase I test, DHS fielded detectors in Chicago to demonstrate the performance of the candidate technology's full system in a representative environment. Results of the testing showed that the candidate system's performance was inconsistent in different operational environments. Specifically, detectors located on underground subway platforms had higher incidences of malfunction than detectors in other locations. These malfunctions may be associated with the presence of metallic brake dust, which demonstrates that different operational environments pose different challenges. Additional rigor in the testing design could have identified limitations earlier and perhaps mitigated them prior to the field testing in Chicago. In fact, the final report by the National Assessment Group on the operational assessment of the Gen-3 Phase I testing concluded that in retrospect, and based on the outcome of testing in the Chicago field test, levels of some possible environmental inhibitors, such as metallic brake dust, represented in the Dugway testing were significantly diluted or did not represent a concentration that compared with concentrations in some of the more problematic operating environments. Therefore, a more robust systems engineering approach to testing contaminants may have helped identify challenges like this earlier. The BioWatch program office agreed that the limited environmental contaminant testing was insufficient to draw any conclusions about system performance on this issue. According to program office officials, the program planned to implement robust interferent tests during Phase II

[56]Dugway officials noted that the BioWatch laboratory protocol includes, in each and every run, a check against inhibition of the PCR assays. They said a muted detection of the genetic material serves as an indication that PCR assays are being inhibited. According to the BioWatch screening protocol, if this check reveals inhibition beyond a specific threshold, then additional steps are taken, including diluting the sample.

to characterize performance against all of the operational requirements and said Phase I was only meant to characterize the state of the market for autonomous detection systems. However, the TEMP for Gen-3 states that the goal of Phase I testing was to evaluate the ability of the candidate systems to meet performance requirements specified in an operational requirements document (ORD). The ORD for Gen-3 specifically outlines the indoor and outdoor environmental conditions under which the system is expected to operate, including, but not limited to, exposure to dust, metallic dust, diesel exhaust, pollen, rain, snow, ice, wind, salt spray, as well as ranges in temperature, humidity, and altitude. DHS's plan to defer more robust testing of conditions representative of Gen-3's intended operational environment does not align with best practices, as performance problems uncovered in later stages of testing can be more costly and require additional testing.

While a basic approach to account for environmental contaminants was included in the systems engineering approach, the program office recognized that the results of the field testing highlighted additional limitations to the test approach for environmental contaminants. The BioWatch program may benefit by incorporating this lesson learned when designing future testing approaches for upgrades or enhancements to the Gen-2 program.

Practice 4: Use Statistical Experimental Design Methodology to Establish a Solid Foundation for Developmental Testing

DHS Partially Aligned with Best Practice 4

DHS included statistical experimental design in its Gen-3 testing plans in order to test performance and characterize uncertainty in the test results. However, the statistical experimental design constructed by DHS was not sufficient to estimate system performance in a realistic environment and did not link KPPs in the ORD to overall program objectives via creation and use of an appropriate model of system performance. In the analytical subsystem test, conducted at Dugway, the candidate Gen-3 system was challenged with aerosols of different concentrations in order to estimate its probability of detection for four of five BioWatch threat agents. Concentrations were systematically varied and were selected using a statistical method that was designed to yield reliable estimates of the system's probability of detection as efficiently as possible (i.e., reducing the experimental effort required to obtain sufficiently reliable information). Statistical uncertainties were calculated for the resulting estimates, and statistical modeling was used to characterize the relationship between aerosol concentration and the system's probability of correctly detecting the presence of each biothreat agent. Statistical experimental designs

were also used in the tests of the aerosol collection component, conducted at Edgewood Chemical Biological Center, and the PCR assays, conducted at Los Alamos National Laboratory. Additionally, the Gen-3 tests included experimental conditions designed to account for factors seen in operational environments, though we have identified limitations in some of these tests, as discussed earlier in the report.[57]

DHS Could Apply Lessons Learned

According to this best practice, test design should be based on a clear understanding of goals and incorporate users' needs. This could be achieved, for example, by linking KPPs in the ORD to overall program objectives and user needs via creation and use of an appropriate model of system performance, but this was not done prior to the Gen-3 testing. Operational requirements, from which the KPPs were derived, were not clearly linked to an overall mission need or program goal. The absence of these linkages among mission need, requirements, and parameters for measuring system performance means results from the Gen-3 testing cannot speak to whether the system would address an established mission need or users' needs.[58] According to program officials, the operational requirements outlined in the ORD were directly linked to the approved Mission Needs Statement for Gen-3. However, in 2012, we reported that officials were aware that the Mission Needs Statement did not reflect a systematic effort to justify a capability need and we reported that the program wrote the Mission Needs Statement later to justify a predetermined solution of acquiring an autonomous detection capability.[59]

Additionally, these tests did not cover all of the system requirements specified by the ORD. For example, one of the system requirements was a maximum false positive rate. DHS noted in the TEMP that the probability of false positives would be estimated from test results using statistical techniques; however, the TEMP did not explain how this would

[57]A comprehensive, detailed evaluation of the strengths and limitations of the specific statistical experimental designs used during Gen-3 Phase I testing was beyond the scope of our review.

[58]DHS commissioned several national laboratories to conduct modeling studies in 2011 to try to better link operational requirements, such as the number of lives saved, to specific performance parameters, such as system sensitivity thresholds.

[59]GAO-12-810.

be done. In testing, DHS did not address this. Instead DHS tested 20 times (per agent) in the assay tests and 15 times (per agent) in the Dugway tests, and drew conclusions despite not having designed experiments that would produce estimates of the system's false positive rate with defined levels of statistical precision.[60]

By using statistical experimental designs in the testing of the Gen-3 technologies, DHS was able to characterize some uncertainty in the test results. However, DHS was not able to determine with a defined level of statistical certainty the false positive rate, and thus was not able to conclude that any tested system or system component satisfies its stated requirement.[61] As DHS considers upgrades to the Gen-2 system or future technology switches for the BioWatch program, DHS could apply lessons learned from the Gen-3 testing to help develop meaningful requirements that are linked to a mission need and where operational objectives (e.g., lives saved) are linked to measurable KPPs, such as system sensitivity. DHS could also use statistical experimental design in future testing of upgrades or enhancements to the Gen-2 system to help characterize uncertainty in results and ensure a representative range of operating conditions are sufficiently used to test system performance.

[60]For example, it was concluded that environmental interferents did not increase the false positive rate. Additionally, the false positive rate requirement in the ORD was not well-defined, as noted by a National Academies committee (*BioWatch and Public Health Surveillance*, Washington, D.C.: National Academies Press, 2011). The false positive rate requirement was initially 10^{-7}, but it was not specified whether this was a rate per detection cycle, per time period, and so forth. The Dugway study of system performance did draw some conclusions about false positive rate—for example, concluding that environmental interferents did not increase the false positive rate—but these conclusions were not supported by data designed to estimate false positive rates with defined levels of statistical precision.

[61]DHS officials contested this point, stating testability must exist within reasonable cost and schedule constraints and be balanced with operational need (e.g., national security need) that is acceptable to end users. They said experimentally validating the false positive rate to the required level is impossible and implied that it does not outweigh a national security requirement. Therefore, if the false positive rate cannot be reasonably tested, DHS could consider alternative performance metrics to address this issue or remove false positive rates from the required KPPs in the ORD.

Practice 5: Measure and Characterize System Performance with Established Procedures, Methods, and Metrics

DHS's Actions Partially Aligned with Best Practice 5

DHS created and executed a well-articulated, but incomplete, plan for measuring and characterizing certain aspects of system performance using established procedures, methods, and metrics. For example, measuring system performance included ranking and scoring by agent the system's ability to detect known strains of an agent, including near-neighbor strains of an agent, and ability to detect agents with possible environmental contaminants present. System sensitivity was characterized using an adaptive methodology to determine, for a range of concentrations of agents, the probability of detection at each concentration. These probabilities were associated with confidence intervals to allow assessment of the range of performance one may expect. Both live and killed agents were used to assess detection sensitivity so that results from component testing could be extrapolated to whole system performance, because currently no facilities exist to perform whole-system live agent testing. Additionally, DHS evaluated suitability requirements by testing variables such as human factors; reliability, availability, maintainability (RAM); supportability; and survivability in an operational environment.

DHS Could Apply Lessons Learned

Whereas the TEMP and other test plans list the five KPPs for the Gen-3 BioWatch candidate systems, there is no specific link between these metrics and the mission objectives for the Gen-3 system. Further, it is not clear how the results of testing in non-representative environments would be used to determine system suitability for Gen-3 purposes. In addition to these best practices, developmental testing guidance indicates that developmental testing is intended to vet systems early to determine the suitability of the system for meeting performance requirements. By testing in a variety of modes intended to replicate an operating environment, results of the tests can be used to characterize and evaluate relevant system performance. However, the tests that were intended to account for factors seen in operational environments were of limited range and scope.[62] For example, the test plan for the Chicago field test suggested that the system would be exposed to dust, metallic dust, smoke, diesel

[62]A comprehensive, detailed evaluation of the strengths and limitations of the specific statistical experimental designs used during Gen-3 Phase I testing was beyond the scope of our review.

exhaust, and pollen, under extended temperature, altitude, and relative humidity ranges. However, there was no design of statistical tests intended to address these variations in Phase I.[63] The analytical sub-system testing did not allow for sub-analysis of pollutants from different environments (e.g., outdoor, indoor, subway), and so it was not possible to identify effects of any specific pollutant, such as subway brake dust. Additionally, the Edgewood test of the aerosol collection component included tests at different temperatures and humidities; however, relatively few combinations of temperature and humidity were tested, a fact that limited the range of environmental conditions represented and limited the utility of their results in determining how the system would perform under various operational environments.

Because DHS performed only limited testing in this regard, it was not able to draw conclusions early in the testing process to determine whether the systems would meet performance requirements. In the future, DHS could improve BioWatch testing efforts by incorporating clearly articulated measures and use of established procedures, methods, and metrics to characterize system performance. This will help ensure DHS collects the information it needs to evaluate operational suitability of upgrades or enhancements to the Gen-2 system.

Practice 6: Test to Build in Resilience, Especially in the Development Stages

DHS's Actions Partially Aligned with Best Practice 6

DHS performed some testing that could help build resilience into the system during Phase I testing, but it could improve resilience testing with more rigorous methods. According to OHA officials, the Phase I acquisition and test and evaluation (T&E) strategies were specifically designed to address the concern of identifying potential vulnerabilities in the systems under test. As part of this strategy, for example, Edgewood tested the collection efficiency of the filters under varying temperature and humidity conditions. The system was also tested in an operational environment in Chicago to assess the effects of environmental

[63]According to DHS officials, these types of tests were planned for Phase II because of results of the Phase I testing. However, as stated earlier, deferring more robust testing of conditions representative of Gen-3's intended operational environment does not align with these best practices, as performance problems uncovered in later stages of testing can be more costly and require additional testing.

interferents, among other things, to help identify vulnerabilities in operating the system. Further, DHS recognized that developing autonomous detection had proved challenging because in addition to some of the technology required being novel, even the existing technologies—for example, the aerosol collection unit and the apparatus that reads the PCR results—had not been combined for the specific application of autonomous detection in an operational environment before. By executing assay evaluation and subsystem- and system-level characterization tests incrementally, this allowed for the insertion of decision points to reduce technical and financial risk. As a result of its multi-stage approach to testing the systems, DHS identified limitations to one vendor's detection system that could not be overcome before proceeding with the next stages of Phase I testing, so that vendor did not complete the entire Phase I testing plan. DHS also included provisions in the Phase I contracts for engineering change proposals to allow the vendors, at DHS discretion, to address deficiencies found during testing to inform DHS decisions on proceeding to Phase II.

DHS Could Apply Lessons Learned

While DHS took some steps to build resilience into the Phase I testing, additional rigor could have been built into the testing design to reveal potential vulnerabilities in the performance of the systems tested. DHS's final test in Phase I, designed to test resilience, involved fielding the detectors in Chicago to demonstrate the performance of the candidate technology's full system in a representative environment. DHS was able to identify some limitations—such as environmental contaminants—to the detection systems being evaluated. However, this occurred late in Phase I testing, and represented only a limited cross-section of possible challenges, including temperature, humidity, and vibrations. For example, while the test plan for the Chicago test lists a range of temperatures and humidities the system is expected to operate under, the field testing did not reflect the entire operational range.

Additionally, other tests on temperature and humidity conditions were performed in limited combinations, such as high temperature and low humidity, but not others, such as high temperature and high humidity. Further, these temperature and humidity tests were done only on the aerosol collection unit, and not the entire Gen-3 system. Because the Gen-3 system as a whole was expected to operate continuously under such conditions, just as Gen-2 currently operates, simply testing one component, the aerosol collection unit, under a limited combination of temperatures and humidities does not adequately test the system for

purposes of building in resilience. Some of these conditions may have provided earlier indications of vulnerabilities that did not arise until near the end of Phase I testing. For example, to assess the effect of interferents, Dugway testing used pooled filter washes provided by DHS. The environments from which the filters came were not provided to Dugway and interferents from different filters were combined. In doing so, the effect of individual interferents was diluted. When the systems were tested in Chicago near the end of Phase I, DHS found that system problems occurred, attributed to petrochemicals near highways and to brake dust in subway stations. If specific interferents, such as subway brake dust, were tested prior to the Chicago testing at representative concentrations, they may have revealed issues earlier in Phase I.

Other conditions, such as network communication performance, were not tested or were tested in a limited fashion. According to the final report on the Gen-3 testing, tests of network communication performance were modified based on user needs in the Chicago area where this capability would have been demonstrated. Therefore, network performance was left undetermined at the end of Phase I testing.

As a result of not including more rigorous testing methods to test resilience, information about system failures in different environments was not revealed until late in the Phase I testing. By following this best practice in future testing of Gen-2 upgrades or enhancements, DHS could help mitigate risks the agency may likely face in acquiring these types of threat detection technologies.

Practice 7: Use Developmental Tests to Refine Performance Requirements

DHS's Actions Partially Aligned with Best Practice 7

DHS used Phase I test results to determine the likelihood Gen-3 candidate systems could meet performance requirements, but revisions to performance requirements were based on a modeling study, rather than the outcome of the Phase I testing. According to OHA officials, at the time OHA began Phase I planning and execution, there was not a robust analytical capability to determine mission-based requirements for key technical parameters (such as system sensitivity). Officials said absent that, a consensus on the technical parameters was made based on the collective best judgment of OHA BioWatch Program Office and S&T Chem Bio Defense subject matter experts. According to DHS officials, the result was more of a "technology push" requirement than a mission outcome driven requirement, and was based upon what was believed to be the state of the art for PCR-based systems. However, as we reported

in 2012, when the technologies were unable to meet the technology push requirement as determined by some Phase I testing, DHS encountered delays and uncertainty about how to move forward. In response to these concerns, the BioWatch Program Office conducted a detailed requirements analysis through a national lab consortium led by Sandia National Laboratory. The study assessed the utility of biodetection systems with varying levels of sensitivity in terms of detection timeliness, population coverage, and lives saved in a bioterror attack.[64] As we reported in 2012, the study, which was completed in January 2012, contained findings that, according to BioWatch Program officials, confirm that the sensitivity requirement could be relaxed without significantly affecting the program's public health mission.[65] As a result, DHS set a new sensitivity requirement based on the modeling studies.

DHS Could Apply Lessons Learned

According to experts, developmental testing should be viewed as a tool in helping to refine performance requirements, and a meaningful performance requirement is one that is not only achievable but also strives to maximize the fulfillment of a mission need. As we reported in 2012, according to BioWatch program officials, the original sensitivity requirement was set based on interest in pushing the limits of potential technological achievement rather than in response to a desired public health protection outcome. They said that this led to a requirement that may have been too stringent, resulting in higher costs and schedule delays without demonstrated mission imperative.[66]

Further, having requirements based on operational objectives would allow the results from Phase I testing to inform the users regarding the anticipated capabilities of the system and the likelihood that a tested

[64]This study is part of the same body of work from Sandia that we described earlier in the context of the current BioWatch system (Gen-2).

[65]Sandia National Laboratories, *BioWatch Technical Analysis of Biodetection Architecture Performance*, SAND2012-0125 (Albuquerque, NM: Jan. 2012). In response to this study, the BioWatch program submitted an updated ORD with a revised sensitivity requirement to DHS in March 2012 for approval in preparation for ADE-2B. Officials reported to us in 2014 that as part of the ORD update, the program also made changes to improve conformance with DHS acquisition guidance and flesh out requirements not fully developed prior to Phase I.

[66]The more stringent the sensitivity requirement, the lower the concentration of a pathogen that must be in the air for the system to detect its presence.

system would be able to meet mission needs. According to experts, developmental testing may help to identify the performance envelope of the system and inform decisions about refined performance requirements. Failure of the Gen-3 candidate system to meet the initial performance requirements resulted in delays and uncertainty regarding the Gen-3 acquisition. DHS took steps to refine the sensitivity requirement for Gen-3, but it was the modeling study, rather than the developmental testing protocol, that led to the change. However, we found that even the relaxed sensitivity requirement did not link system performance to a clear operational objective (e.g., to detect attacks of certain types and sizes), as discussed earlier in this report. Instead, the relaxed requirement was based on ideas about the performance characteristics of the current (Gen-2) system—the idea being that Gen-3 could be less sensitive than Gen-2 but still achieve comparable public health outcomes because of its greater speed. According to DHS, the change in sensitivity requirement was linked to casualty reduction, and the agency does not agree with our assessment. However, as noted earlier in this report, the performance of Gen-2 has not been linked to a clear operational objective; therefore, because the revised sensitivity requirement for Gen-3 was based on Gen-2, it was not grounded in an operational objective, either. In any future acquisition, upgrade, or enhancement to Gen-2, having initial requirements based more closely on mission need and operational objectives may prevent delays and uncertainty. In its Post Implementation Review, DHS also recognized the need to better communicate with stakeholders about using a flexible testing approach to refine requirements to avoid any misperception that the requirements would be adjusted to accommodate the vendor's capabilities.

Practice 8: Engage in a Continuous Cycle of Improvement by (1) Conducting Developmental Testing, (2) Conducting Operational Testing, and (3) Incorporating Lessons Learned

DHS's Actions Aligned with Best Practice 8, as Applicable

DHS took steps to engage in a continuous cycle of improvement during the Gen-3 acquisition, but not all components of this practice apply, as DHS never reached the stage of operational testing.[67] In the Gen-3 TEMP, DHS describes an integrated test approach, designed as a continuum from developmental testing through operational testing, with the previous test events acting as the foundation for the follow-on events. However, given the expectation of rapid acquisition and deployment, DHS missed opportunities to fully engage in a broader testing approach

[67]DHS completed developmental testing and assessed the lessons learned from the Gen-3 testing and acquisition.

needed for a novel system approach to biodetection. By not engaging in a more rigid testing approach, when the autonomous detection systems tested did not yield favorable outcomes, it raised too much uncertainty about cost and performance for the program. As a result, DHS canceled the Gen-3 acquisition and prepared lessons-learned documents for the Gen-3 acquisition that are intended to inform the BioWatch program's actions in the future. This could include applying the lessons learned to improvements or upgrades to the existing Gen-2 system. OHA officials reported that following the Phase I testing, the BioWatch program facilitated lessons-learned conferences that included relevant stakeholders.[68] For example, OHA officials said results from the Phase I testing indicated that the basic approach used to assess the technology worked well and appeared to answer the relevant questions as to the readiness of a technology considered for deployment. Specifically, they said testing of major subsystems independently was useful as was the chamber testing conducted using a killed (nonviable) agent. Other positive aspects included the usefulness of an interagency test and evaluation Working Integrated Product Team, and the value of independent test agencies. According to OHA, lessons learned include the need for widely accepted PCR Assay Performance standards (recently reviewed by a National Academies consensus committee) and the establishment and enforcement of operational performance criteria during testing to avoid repeated adjustments such as those that were made by one of the vendors to its agent identification algorithm and assay chemistry.[69] According to OHA officials, these adjustments delayed the testing and increased costs.

[68]The products from these conferences are Memorandums for the Record (MFR), specifically the Gen-3 Phase I Lessons Learned for Characterization Testing, dated October 18, 2011, and the Gen-3 Field Test Lessons Learned Conference Summary, dated June 15, 2011. There are also lessons learned identified in the Gen-3 Phase I Technology Readiness Review (TRR) and the Gen-3 Post Implementation Review.

[69]National Research Council and Institute of Medicine, *BioWatch PCR Assays: Building Confidence, Ensuring Reliability* (Washington, D.C.: National Academies Press, 2015).

DHS Identified Key Lessons Learned from the BioWatch Gen-3 Acquisition

1. Any policy decision to accelerate a future acquisition of detection technology should be fully documented, capturing the justification for urgency, an understanding of the limitations of current technology capabilities, and the minimum acceptable (non-technical) requirements needed to achieve the improvement envisioned.

2. Deliberations regarding biodetection research and development during an acquisition process should have pre-designated forums that are capable of resolving issues (scientific or otherwise) as they arise. An inability to reach resolution or consensus should be documented and openly acknowledged, allowing for the development of an adjusted path forward for the acquisition process, if necessary.

3. Communication efforts should be increased to Department stakeholders and leadership, both to improve efficiency of the acquisition process and to fully document an acquisition's updates to any requirements, timelines and/or procedures.

4. State and local government stakeholders, especially in locally-operated programs (such as BioWatch), should be integrated into the requirements generation process as early as possible, recognizing their ultimate role as the end-user.

Source: Department of Homeland Security (DHS). | GAO-16-99

DHS Applied Lessons Learned

In April 2014, after assessing the results of the Phase I testing outcomes, DHS canceled the Gen-3 acquisition when unfavorable testing outcomes raised too much uncertainty about cost and performance of the autonomous detection systems tested. While some might consider the cancellation of the Gen-3 acquisition a failure, by carefully weighing the pros and cons of moving forward with an acquisition that had not produced favorable results, DHS incorporated parts of the best practice of continuous improvement and avoided greater expense for a system that had not met program requirements. By applying the lessons learned that DHS identified as a result of evaluating the Gen-3 acquisition, as well as those identified through application of these best practices for testing, DHS can continue to engage in a continuous cycle of improvement for its testing and acquisition of detection technologies as it considers upgrades or enhancements to the Gen-2 system.

Lessons Learned from Other DHS Acquisitions of Threat Detection Technologies

BioWatch Gen-3 is one of several DHS technical system acquisitions that have faced challenges because of system immaturity or unreliable performance in an operational environment. In some cases, DHS canceled the acquisitions after testing did not yield favorable results. For example, we previously found the Secure Border Initiative Network (SBInet) testing did not appropriately account for risk or provide sufficient information to ensure system performance.[70] Aiming to enhance border

[70]GAO-10-158.

security and reduce illegal immigration, DHS launched SBInet to create a "virtual fence" along the border using surveillance technologies. However, as with our 2012 findings on the Gen-3 acquisition, we found the SBInet Program Office had not effectively performed key requirements of development and management practices. For example, some operational requirements for SBInet, which are the basis for all lower-level requirements, were found to be unverifiable, and we concluded that the risk of SBInet not meeting mission needs and performing as intended was increased. As noted above, the Gen-3 acquisition also did not have operational requirements directly tied to mission need.[71] Additionally, we found none of the SBInet plans for tests of system components addressed testing risks and mitigation strategies. As noted above, assessing technical and performance risk in an acquisition can help mitigate cost and schedule problems. Although we made several recommendations aimed at improving the rigor and discipline of SBInet testing, DHS canceled the SBInet acquisition in January 2011, in response to internal and external assessments that identified concerns regarding the performance, cost, and schedule for implementing the systems.

Similarly, we previously reported that a primary lesson learned regarding testing of the Domestic Nuclear Detection Office's (DNDO) advanced spectroscopic portal (ASP) was that the push to replace existing equipment with new technology led to a testing program that lacked the necessary rigor.[72] The ASP was a type of portal monitor designed to both detect radiation and identify the source to reduce both the risk of missed threats and the rate of innocent alarms. DNDO considered these to be key limitations of radiation detection equipment used by Customs and Border Protection (CBP) at U.S. ports of entry. Based on our body of work on ASP testing, one of the primary lessons to be learned is to avoid the pitfalls in testing that stem from a rush to procure new technologies. We have previously reported on the negative consequences of pressures imposed by closely linking testing and development programs with

[71]DHS officials contend that the operational requirements in the Gen-3 ORD were directly linked to the approved mission need statement, but in 2012 we reported that the mission need statement was written to justify a predetermined solution of acquiring an autonomous detection capability.

[72]GAO-09-804.

decisions to procure and deploy new technologies.[73] In the case of ASPs, as well as the Gen-3 acquisition, the push to replace existing equipment with the new portal monitors led to a testing program that initially lacked the necessary rigor. DNDO's schedule consistently underestimated the time required to conduct tests, resolve problems uncovered during testing, and complete key documents, including final test reports. We also found that testing of the ASPs lacked sufficient rigor to demonstrate performance of the detectors in an operational environment. For example, ASP testing did not include a sufficient amount of the type of materials that would mask or hide dangerous sources that ASPs would likely encounter at ports of entry, which is fundamental to the performance of radiation detectors in the field. As noted above, Gen-3 testing of possible environmental contaminants also lacked sufficient rigor to identify potential vulnerabilities in the system's detection capabilities prior to placing them in the field. Despite several recommendations we made to DHS on ways to improve the testing and management of the ASP acquisition, because of unsatisfactory test results, ASP did not pass field validation testing, which led DHS to cancel the program in October 2011.

DHS has taken several steps in recent years to improve acquisition management in response to our previous recommendations. By establishing a policy that largely reflects key program management practices, dedicating additional resources to acquisition oversight, and improving documentation of major acquisition decisions in a more transparent and consistent manner, DHS has improved its ability to manage acquisition programs. The decision to cancel the Gen-3 BioWatch acquisition is an example of improved oversight, and DHS could continue to improve by implementing some lessons learned from the Gen-3 acquisition. In April 2015, we reported that DHS's Director of Operational Test and Evaluation has expressed interest in becoming more involved in testing earlier in the development process to increase influence over program execution.[74] The Director told us that determining how test activities should inform key decisions would help mitigate risk for all types of programs. As DHS considers its options regarding the currently deployed Gen-2 BioWatch system, including possible

[73]GAO, *Best Practices: A More Constructive Test Approach Is Key to Better Weapon System Outcomes,* GAO/NSIAD-00-199 (Washington, D.C.: July 31, 2000).

[74]GAO, *Homeland Security Acquisitions: Major Program Assessments Reveal Actions Needed to Improve Accountability,* GAO-15-171SP (Washington, D.C., Apr. 22, 2015).

technology enhancements or even future technology switches, future BioWatch acquisition and testing efforts may benefit by incorporating the lessons learned from the Gen-3 Phase I testing and other recent DHS acquisitions that have faced testing challenges.

PCR Technology Is Most Mature for an Autonomous Detection System to Upgrade Gen-2, but Such Systems Face Uncertain Benefits and Several Likely Challenges

PCR is the most mature and sensitive technology for an autonomous detection system, and DHS is considering autonomous detection as an upgrade to Gen-2. While autonomous detection may provide benefits that include reduction in casualties and clean-up costs and greater cost-efficiency, the potential benefits of an autonomous system for BioWatch depend on specific assumptions, some of which are uncertain. For example, reductions in casualties would depend on a rapid, coordinated response from multiple entities at the federal, state, and local levels; whether such a response would materialize is uncertain and partially outside DHS's control.[75] Further, an autonomous detection system would have to address several likely challenges, including minimizing possible false positives, securing a networked detection and communication system, and operating under various environmental conditions.

PCR Is the Most Mature Technology for an Autonomous Detection System

The most mature analysis technology for an autonomous detection system is currently PCR, according to a National Academies report on promising technologies for autonomous detection for BioWatch and interviews with stakeholder officials, including CDC, the DHS BioWatch program manager, and other experts.[76] As mentioned earlier, while DHS canceled the Gen-3 acquisition of an autonomous detection system for BioWatch, OHA and S&T are collaborating to address the capability gap that Gen-3 intended to fill by evaluating upgrades or enhancements to the current Gen-2 system. According to DHS officials, autonomous detection

[75]DHS runs training exercises as part of its "Public Health and Preparedness" activities. However, these do not provide specific timing for responses to positive detections and do not decrease uncertainties in response timing. Further, on the basis of the summary documents provided by the DHS, many jurisdictions—more than half in some cases—report still requiring additional guidance for responding to a variety of detections, suggesting there is uncertainty in how each jurisdiction would actually respond.

[76]Near term means within the next 5 years. The National Academies report is: Institute of Medicine and National Research Council, *Technologies to Enable Autonomous Detection for BioWatch*. In addition to consulting the reported sources, we performed a literature review covering the past 3 years and could not identify a technology for autonomous biodetection that was more mature than PCR.

is among the technologies being considered. The National Academies report presented perspectives from local officials and technological assessments, gathered from a workshop that DHS requested, to explore alternative cost-effective systems that would meet the needs for a next-generation BioWatch system. This proposed next-generation system was intended to operate autonomously to detect BioWatch threat agents from aerosol samples. The National Academies report described the state of autonomous detection technology in 2013 and evaluated four broad classes of technologies (see table 3). Those technologies were PCR, immunoassays and protein signatures, genomic sequencing, and mass spectrometry.

Table 3. A National Academies Workshop Evaluation of Four Technology Classes for Autonomous BioWatch Systems

	PCR[a] (nucleic acid signatures)	Immunoassays and protein signatures	Genomic sequencing	Mass spectrometry
Technology readiness level[b]	9	6+	4	6
Sample preparation effort	Moderate	Low	Moderate	None
Sensitivity	High	Moderate	High	High
Specificity	High	Moderate	High	Moderate
Cost	Moderate	Low	High	Low
Standalone[c]	Yes	No	No	No

Source: adapted from Institute of Medicine and National Research Council, *Technologies to Enable Autonomous Detection for BioWatch* (Washington, D.C.: National Academies Press, 2014). | GAO-16-99

[a]PCR: polymerase chain reaction. PCR is currently used in the Gen-2 BioWatch system.

[b]Technology readiness level (TRL) is a method of estimating the maturity of a technology for a given purpose, ranging from TRL 1 to TRL 9. For example, a technology readiness level of 6 indicates the technology has been validated and is ready for testing in a setting representative of an operational environment. A readiness level of 9 indicates the technology has been deployed under operational mission conditions. Although the National Academies report assessed PCR at TRL 9—the highest TRL level—the PCR-based autonomous detection system tested under the Gen-3 acquisition was not assessed at that high a level for the whole system. We reported in 2012 that an independent technology readiness assessment rated all but one of the critical technology elements of that system as TRL 7—indicating a relatively high level of maturity for each technology element assessed—but lower than TRL 9.

[c]A technology is standalone if it can be used to detect and confirm the identity of an agent without the use of another technology.

Three Key Terms for Understanding Autonomous Detection Technologies

Genes: Genes are sections of nucleic acids that determine how proteins are made. Nucleic acids, such as deoxyribonucleic acid (DNA), are long chains of molecules made up of bases, of which there are four possible kinds. The order of the bases represents the sequence of the DNA. Because proteins determine much of the function of an organism, sequencing the genes can provide information about its identity. The set of all genes of an organism is known as its genome.

Protein signatures: Proteins are long chains of building blocks called amino acids. Because there are many types of amino acids, the way each protein interacts with different stimuli, such as light, can be unique. For example, shining one color of light on a protein can result in its emitting different colors. The set of colors the protein gives off can be considered an "optical fingerprint" of the protein, commonly referred to as a type of protein signature.

Antibodies: Antibodies are proteins created by the immune system that attach to specific chemicals on the surface of disease-causing organisms, resulting in the organisms' being rendered harmless. Antibodies can be designed to attach to a given target and modified to be detectable under certain conditions. Thus it is possible to use antibodies to see whether a target is present by soaking a sample with the antibodies, washing off unattached antibodies, and then measuring those that remain.

Source: GAO. | GAO-16-99

PCR, which is used to detect nucleic acid signatures, is used to amplify and detect genetic material, or nucleic acids, of organisms.[77] By amplifying (i.e., repeatedly duplicating) those sections of genes associated with certain biological agents, it is possible to distinguish the agents among various organisms. Because of the amplifying capability of PCR, small amounts of genetic material are sufficient for detection, resulting in high sensitivity for this technology. Specificity can be high if the sections of the genes being amplified are unique to the agent. However, related organisms, called genetic near-neighbors, may contain similar gene sections and lead to a PCR detection when the agent itself is not present. PCR is the method used in the current (Gen-2) BioWatch system.

Immunoassays and protein signatures use antibodies or light to identify organisms. Immunoassays use antibodies that attach to chemicals that primarily appear on certain biological agents; thus immunoassays can be tailored for high specificity. Protein signatures analyze how light interacts with different chemicals (such as proteins) on target agents, using light "signatures" emitted by specific proteins to be identified. However, neither of these methods is as specific as PCR. Also, because there is no amplification, the sensitivity of these methods is not as high as that of PCR.

Genomic sequencing provides a genetic sequence for all or part of a detected organism's genes. Because each agent contains unique genetic sequences, this method is very specific and could eventually provide information regarding antibiotic susceptibility. However, the method is not considered standalone since it depends on another method, such as PCR, to work. If used with PCR, then this method is also very sensitive. Of the four broad technologies examined by the National Academies, this method is also the least mature because of issues with systems integration—for example, developing the software to perform the analysis locally (within the device itself).

Mass spectrometry breaks apart a sample (for example, by directing a laser onto the sample) and analyzes the resulting fragments. Different

[77]While basic PCR does not yield detection results without further steps, there are modifications, such as real-time PCR, which provides a readout of the amount of material being amplified as the reaction occurs. This readout is commonly in the form of light given off by the amplified nucleic acids.

chemicals yield different types and amounts of fragments, so it is possible to reconstruct the chemical composition of the original sample. Because the chemical make-up of agents is unique, it is possible to identify their presence in the sample. Mass spectrometry is not as sensitive or as specific as PCR.

Benefits of an Autonomous Detection System Depend on Several Assumptions

We identified key potential benefits of an autonomous detection system from discussion with agency officials, a review of agency and national laboratory documentation, and a literature review. Most of these potential benefits were owing to faster detection; however, we determined that the extent to which faster detection confers benefits depends on specific assumptions, some of which are uncertain and some of which are outside of DHS's control.[78] Additionally, from our review of literature, we identified potential benefits that included decreasing user errors, such as dropped collection filters. However, since these benefits depend on the actual design and implementation of the system, it is difficult to predict the extent to which they would be realized. The benefits and challenges discussed in this report apply broadly to autonomous detection; that is, they do not depend on which of the four broad classes of technology is deployed.

According to a 2011 National Academies report, an autonomous detection system could detect agents more quickly than the Gen-2 system because of a shorter sample collection period, elimination of sample transport, and completion of the detection step within the system itself (see fig. 3). In particular, the report stated that while the current Gen-2 BioWatch system could detect agents in 10-36 hours, an autonomous detection system could detect agents in as little as 4-6 hours. Further, DHS officials and Sandia modeling studies state that faster detection enabled by automation can provide:[79]

[78]In such cases, our discussion of potential benefits and likely challenges applies to faster detection systems regardless of whether they are automated.

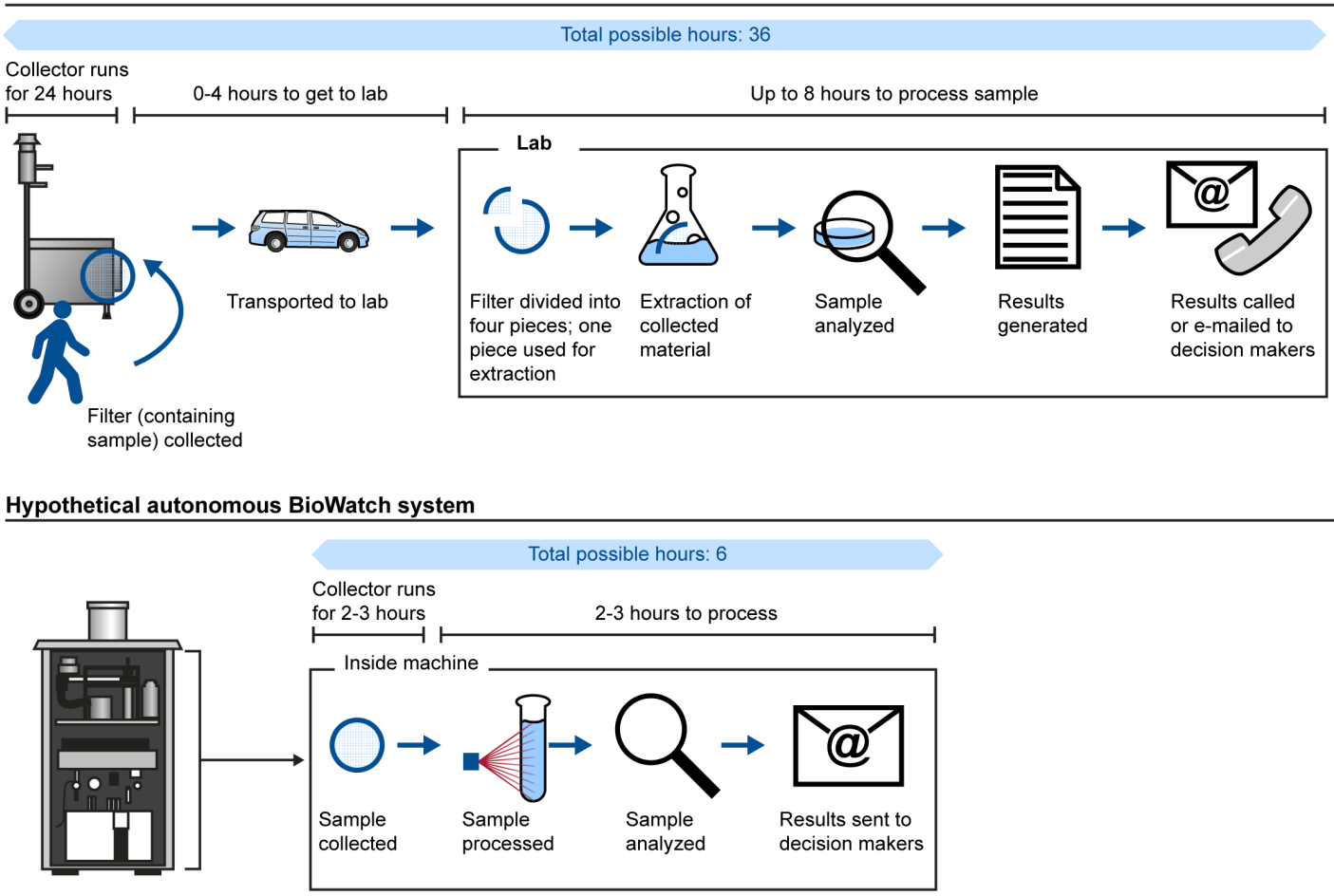
[79] DHS also states enhanced national coordination for responding to attacks, by encouraging coordination and communication among stakeholders, as a benefit, but we were unable to establish the additional benefits conferred by autonomous detection systems over those attained by the Gen-2 system.

1. reduction in casualties and/or fatalities because of faster detection and faster situational assessment;[80]
2. lowering costs, including clean-up costs by halting the entrance of transport vehicles, such as trains or airplanes, into contaminated areas; and
3. reduction in the total annual cost of the detection system, per detection cycle.[81]

[80]Faster detection compared with that of the Gen-2 BioWatch system. Situational assessment refers to activities undertaken following a BAR to determine the potential risk to public health.

[81]A detection cycle is the process of collecting an aerosol sample and analyzing the sample for the presence of a select agent.

Figure 3: Differences in Detection Time between the Gen-2 and an Autonomous BioWatch System



Source: GAO analysis/interpretation of information from DHS, Sandia National Laboratories, and the National Academies. | GAO-16-99

DHS officials told us that an autonomous detection system would offer many of these benefits but did not provide evidence to support them, saying that their assertion is "common sense." DHS officials also referred to Sandia's modeling studies. However, we determined that these benefits and the conclusions of the Sandia modeling studies depend on specific assumptions, some of which are uncertain and some of which are outside of DHS's control, although DHS officials stated that they believe the modeling assumptions are reasonable for the intended purpose. However, a CDC official cautioned against relying on models to determine program effectiveness.

Life-Saving Benefits of Early Detection Are Uncertain and Depend on Factors Outside of DHS's Control

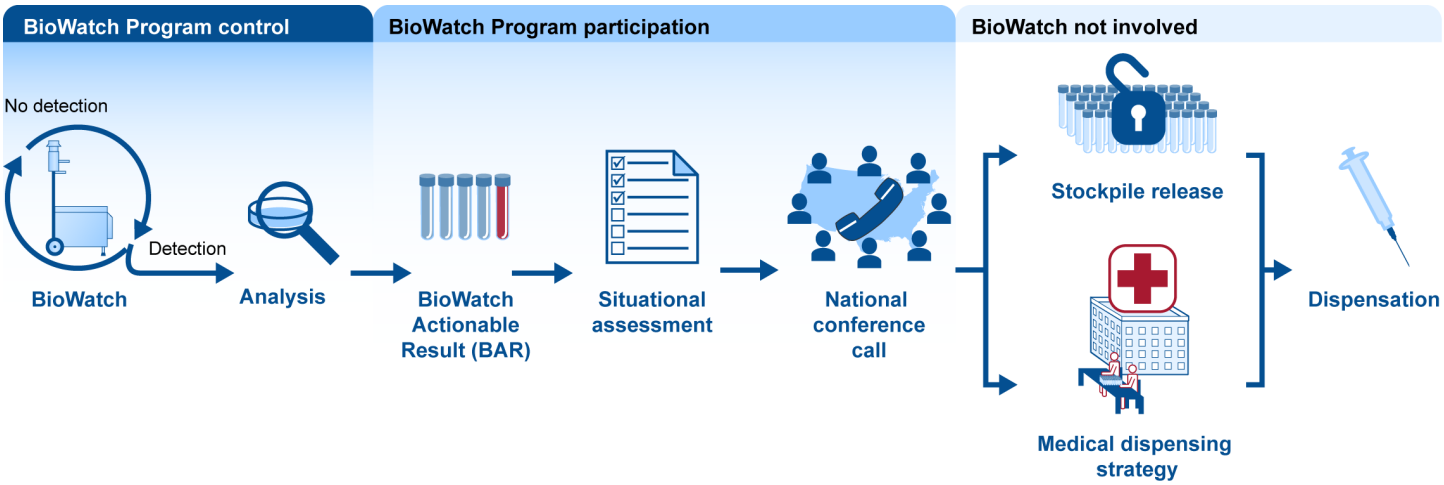
The number of lives saved from faster detection could not be determined, because some key factors affecting response time are uncertain. For example, the time it takes decision makers to determine that a detection represents a threat to public health and warrants dissemination of medical countermeasures is variable. The time between a BAR and dissemination of medical countermeasures may include the time needed to characterize the incident, determine who was exposed, make decisions regarding evacuation of contaminated regions and relocation of individuals, determine where to set up medication "points of dispensing" and to actually mobilize the medication stockpile, and distribute medication to potentially exposed people and keep track of who received medication. According to the National Academies report and current DHS guidelines, steps taken after detection, to instill confidence for requesting medication, include assessing known threats, conducting additional local lab work, and initiating a national conference call (see fig. 4).[82] There may be additional tasks such as culturing of the agents to determine their viability and antibiotic resistance. If local stakeholders follow the guidelines, there could be considerable delay prior to mobilization of stockpiled medication. For example, an official at the National Academies workshop reported that he takes an additional hour to perform an assessment prior to any national conference call. According to another National Academies report, the BioWatch national conference call usually occurs 1-2 hours after a local call of the BioWatch Advisory Committee. Thus, the time between an attack and when medication is fully distributed—and the number of lives potentially saved by minimizing this time—could vary from jurisdiction to jurisdiction.[83] DHS officials agreed that the jurisdictional response can vary and conducts exercises and training to help plan for a response. However, it is not clear what effect such exercises have on response time variability.[84]

[82]Because there are no guidelines for an autonomous detection system for BioWatch, we are using the current post-BAR protocol for illustrative purposes on actions that may be taken by local jurisdictions immediately following a BAR in an autonomous detection system.

[83]For example, public health officials at a Sandia workshop in 2009 expressed varying responses based on the information they receive from a biodetection system.

[84]We reviewed exercise and training summary documents provided by DHS but were unable to identify the effect of such exercises on response time.

Figure 4: Gen-2 Post-BAR Response from Detection to Treatment



Source: GAO analysis/interpretation of information from DHS, Sandia National Laboratories, and the National Academies. | GAO-16-99

In addition, faster detection may not be the most effective way to save lives. For example, a modeling study showed results indicating that an attack detected in 2 days, but requiring 10 days to distribute medication, would result in more deaths than an attack detected in 5 days, but requiring only 2 days to distribute medication.[85] Thus, according to this model, a jurisdiction that shifts resources from medical distribution capacity to faster detection may end up with more deaths. Decisions of resource prioritization are not under BioWatch program control, neither is the part of the response involving medication or other intervention. However, the benefits from early detection depend on such resource prioritization and effective overall responses. DHS officials agree, noting that biosurveillance is a coordinated, holistic endeavor.

Sandia ran a response model to estimate, among other things, the number of casualties and fatalities given the time that passes between the attack and detection (time to detection). Sandia's modeling studies showed that a faster detection system would reduce the number of casualties and fatalities, but that the extent of these reductions would

[85]L. M. Wein and D. L. Craft. "Evaluation of Public Health Interventions for Anthrax: A Report to the Secretary's Council on Public Health Preparedness," *Biosecurity and Bioterrorism: Biodefense Strategy, Practice and Science*, vol. 3, no. 4 (2005).

depend on assumptions in the model. One such assumption was the probability that BioWatch correctly detected the attack. As discussed earlier in this report, the probability that BioWatch correctly detects an attack depends on many factors, including the performance characteristics of the technology and the characteristics of the attack itself. Many of Sandia's estimates of the life-saving benefits of a detection system—automated or not—are downstream analyses presuming that an attack was detected. If the attack was not detected, the faster response enabled by a detection system would not occur, and there would be no life-saving benefits from operating such a detection system. Therefore, if those results are read out of context of this presumption, the expected reductions in casualties and fatalities may be overestimated. This limitation applies to autonomous detection as well as the Gen-2 system.

Estimated reductions in casualties and fatalities from faster detection also depend on assumptions about the infectivity of the BioWatch threat agents. In the Sandia modeling studies, infectivity was represented by infectious dose estimates—that is, estimates of the doses that would lead to illness; however, we found uncertainty in these estimates. As described in our earlier discussion of the current (Gen-2) system, Sandia researchers and other experts told us there is considerable uncertainty in even the best available infectious dose estimates for anthrax, as these estimates are based on data from nonhuman primates.

Finally, the life-saving benefits of faster detection that Sandia reported varied significantly, depending on the properties of the illness that the agent caused. These properties included how long it takes for symptoms to exhibit in a patient after exposure and how effectively medication can prevent death in ill people. According to Sandia, some agents act very slowly—that is, they have long incubation periods—which diminishes the effect of faster detection. For example, an agent that takes over 7 days to cause symptoms will be detected by a 36-hour and a 4-hour detection system with similar outcomes.[86] Another factor is how effectively a developed illness can be treated. According to Sandia, this factor is also

[86]While treatments may have different effectiveness based on how quickly they are applied, the Sandia model does not account for this variance. Sandia partitioned the illness into two categories—preventing illness after exposure using prophylaxis, and preventing death after illness develops using hospital treatment. A further limitation in the Sandia model is that variance in effectiveness of a response within each category was not implemented. For example, it may be the case that prophylaxes are more effective when given sooner after exposure, even if there is a long incubation period.

variable, so that for some agents, the numbers of lives saved in shifting from 36-hour to 4-hour detection change little. Thus, reducing fatalities by faster detection depends largely on the agent used in an attack.

The Lowering of Cleanup Costs Is Uncertain because of Traffic Variability

Sandia reported that faster detection improves the ability to divert transport vehicles—such as trains and airplanes— from contaminated areas so that they do not have to be subsequently cleaned up.[87] The benefit's extent is uncertain because it depends on the amount of traffic entering a given location. For example, according to Sandia, the number of subway cars entering New York City's Grand Central Terminal over a period of 5 hours can range from as few as 250 to as many as 1,750, depending on the time of day and the day of the year. Clean-up effort reduction is thus uncertain because of an attack's unpredictability. A similar analysis can be made for people entering a contaminated area—while early detection could lead to exposure prevention and mitigating the need for additional medication, the actual number of people affected is similarly variable.

We recognize that much of the uncertainty described regarding lives saved or reduction in clean-up costs is out of the control of the BioWatch program. However, when describing benefits of faster detection, particularly concerning the number of lives possibly saved by an autonomous system or any early-warning system, it is important to understand the uncertainties in these assumptions, such as response time or infectious dose of the agent. Without a comprehensive enumeration of these assumptions and their effects on the modeling results, assertions regarding the value of autonomous detection systems are questionable.

The Projected Cost Savings for an Autonomous Detection System May Not Materialize

In 2012, we found that DHS performed a limited cost trade-off assessment of switching to an autonomous detection system that focused on cost per detection cycle—that is, the cost each time an autonomous detector tests the air for agents versus the cost each time a Gen-2 filter is manually collected and tested in a laboratory.[88] We reported in 2012 that cost per detection cycle was lower with an autonomous detection system, but that overall annual program costs would increase from the current Gen-2 system program costs. From figures DHS provided in 2015

[87]The diversion of transport vehicles is sometimes referred to as rolling stock diversion.

[88]GAO-12-810.

regarding the cost of switching to an autonomous detection system with coverage comparable to that of the current Gen-2 system, our analysis yielded results similar to our 2012 findings.

To determine potential cost savings between the Gen-2 system and an autonomous system, DHS compared an autonomous system with a modified Gen-2 BioWatch system. Gen-2 generally runs one detection cycle daily, but for DHS's analysis the agency compared an autonomous system with a modified Gen-2 system which would run three daily detection cycles. We determined that total annual program costs would increase if current operations for Gen-2, which involve one detection cycle per day, were replaced with an automated detection system. Only by comparing the total annual program costs of operating Gen-2 with three detection cycles per day with the total program costs of operating an automated system were cost savings realized (see table 4). As we reported in 2012, conducting a more complete analysis of costs and benefits would help DHS develop the kind of information that would inform trade-off decisions regarding changes to BioWatch technology.[89]

Table 4: Annual BioWatch Costs by Detection Cycle Frequency

BioWatch system	Total annual program cost (Dollars in millions)
Gen-2 with one cycle per day	**$85**[a]
If Gen-2 ran three cycles per day	**$205**
Autonomous with six to eight cycles per day[b] (hypothetical)	**$96**

Source: GAO analysis of DHS data. | GAO-16-99

Note: DHS provided the estimates of costs for operating both the current system (Gen-2) with 3 daily detection cycles and a hypothetical autonomous detection system. Therefore, these figures do not represent the funding actually appropriated for the BioWatch program. This analysis is only illustrative. For more information on our methodology, see app. I.

[a]Non-operation and maintenance costs are assumed to be the same for 1 and 3 daily cycles, for the current BioWatch system. This number was derived from data DHS provided.

[b]The autonomous system cost does not include the cost of acquiring an autonomous system.

Additional Benefits of Automation May Not Be Realized

Automation may lead to additional potential benefits including fewer user errors and greater efficiency—for example, using fewer resources to accomplish the same amount of work—and greater worker safety by

[89]GAO-12-810.

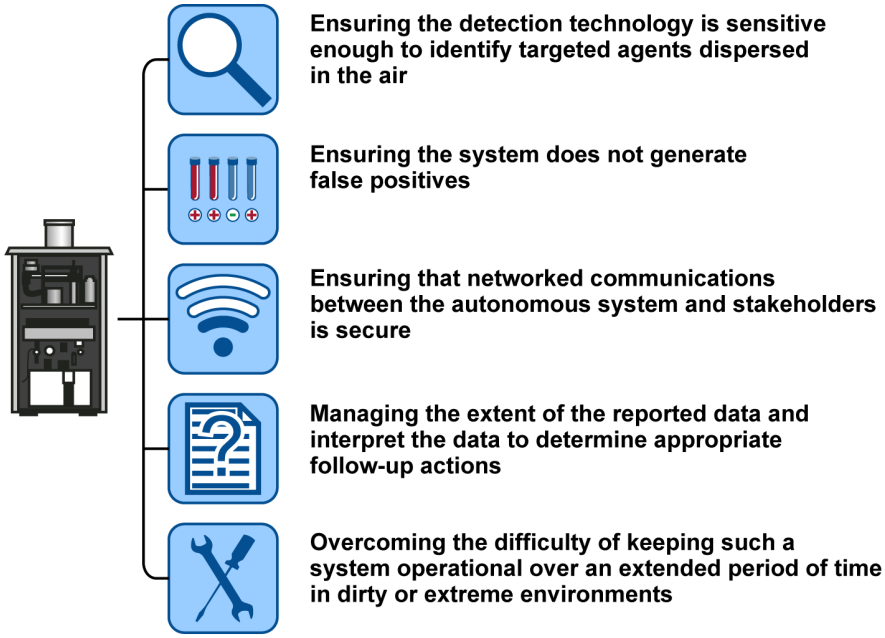
facilitating the handling of dangerous materials, according to literature on automation. However, because automated detection systems have not been deployed, assessing these benefits is difficult. Additionally, uncertainty about how the technology will work means that its benefits might be countered by new problems. For example, according to a 2007 DHS Inspector General report, transferring BioWatch system filters was done improperly several times in 2004.[90] By eliminating the need for transporting filters, automation could avoid this problem, but new problems could arise, such as system crashes. For example, repeated system crashes occurred when the BioWatch Program Office conducted a trial deployment of an autonomous detection system, in New York in 2008. Thus, it is not clear that an autonomous system would realize these benefits.

Autonomous Detection Systems Must Overcome Several Likely Challenges

The challenges an autonomous detection system must overcome include ensuring its detection sensitivity and protecting against threats to networked communications. From a National Academies workshop and interviews with agency officials, we identified five likely challenges (shown in fig. 5).

[90]Department of Homeland Security, OIG Advisory Opinion No. 7-22, OIG-07-22, (Washington, D. C.: Dec. 28, 2007).

Figure 5: Five Likely Challenges an Autonomous Detection System Faces in the Near Term



Source: GAO analysis/interpretation of information from DHS, Sandia National Laboratories, and the National Academies. | GAO-16-99

Meeting Sensitivity Requirements Has Proved Challenging

According to DHS, ensuring that the autonomous detection system meets BioWatch sensitivity requirements represents a major technical hurdle. As discussed earlier, the original sensitivity requirement for the Gen-3 system was based on a technology push because DHS lacked the analytical tools needed to generate a mission-based sensitivity requirement. DHS later revised the sensitivity requirement based on a Sandia-led modeling study. As we described earlier, the Sandia model is subject to limitations and assumptions, and how the sensitivity requirement may be linked to mission outcomes, such as detecting attacks that lead to 10,000 casualties, remains uncertain. Additionally, challenges may be associated with designing a technology to meet a given sensitivity requirement.

One way to manage sensitivity requirements is to assess whether the technology in a detection system conforms to performance standards. The standards may be subject to validation by independent groups or agencies, and constitute guidelines for the technology. For example, the number of times a test should be run and the verification of reagents are

standardized so that results can be interpreted meaningfully. With the development of newer technologies for detection, a method known as multiplexing is being increasingly used.[91] However, validating the use of multiplexing in detection systems has no performance standards. DHS commissioned the National Academies to examine performance standards for PCR, including multiplexed PCR.[92] However, the report recently released by the National Academies does not provide clear standards for multiplexing, instead noting that combining Food and Drug Administration (FDA) multiplexing guidance with certain standards, such as SPADA, which discuss multiplexing, should form a starting point for validation testing. The report also notes that changing to multiplexing (from singleplexing) reduces the sensitivity of the assay, although the effect of this reduction is unclear. Given that PCR is the most mature technology for autonomous detection systems, implementing other technologies may require similar, or greater, effort in establishing performance standards.

The Challenge of False Positives

A challenge for deploying autonomous detection systems identified by participants at the National Academies workshop is the avoidance of false positive readings—readings that indicate the presence of an agent that is not present. False positive readings can lead to major disruption from shutting down crucial transportation and economic facilities (such as airports and shopping centers—referred to as high-consequence actions) and to the unnecessary medication of an uninfected population—which can lead to adverse effects and medical stockpile waste. Local public health officials stated that false positives are likely to be a bigger issue with autonomous detection systems, because operating more detection cycles could increase their frequency.

According to the National Academies report, another common concern among public health officials is their credibility when making high-consequence decisions. At the workshop, officials stated that the integrity

[91]Multiplexing refers to running more than one test in a single tube (a test tube containing the aerosol sample in a liquid solution). Singleplexing refers to running a single test in a single tube. Multiplexing is advantageous because it conserves volumes and reduces the number of actions to be taken to obtain multiple readouts. However, multiplexing may be more susceptible than singleplexing to interference from multiple chemicals reacting with one another. The current BioWatch system uses a singleplexed PCR technology.

[92]National Research Council and Institute of Medicine, *BioWatch PCR Assays: Building Confidence, Ensuring Reliability*.

of public health is critically important, and thus they needed complete confidence in an autonomous system, which is intended to provide results without human interaction or interpretation. Similarly, according to a Sandia workshop in 2009, public health officials largely felt that wrongly taken high-consequence actions would result in loss of credibility.[93] Finally, an LLNL scientist stated that debugging a complex system to determine whether a potential false positive occurred may be an issue with some autonomous systems. He noted that false positives from naturally occurring genetic near-neighbors of the BioWatch threat agents might be a particular challenge and that DHS has made limited investments in determining background DNA signatures to address this issue. DHS officials stated they use data gathered from current operations to assess such background signatures, but it is unclear whether this approach would be effective for an autonomous detection system.

Ensuring Reliable and Secure Information Technology Networks Is Challenging

According to DHS and CDC officials, another challenge autonomous detection systems face is securing the networked communication system against interference, such as from hackers. DHS officials stated that the security of network communications for transmitting results to the local officials was an important issue for autonomous detection systems. For example, during the Gen-3 effort, DHS officials specifically planned for testing of network security as described in the TEMP. DHS officials stated that an unsecure system would be vulnerable to hackers' planting results or shutting systems down. In 2012, we reported that the 2011 Operational Assessment stated that failure to demonstrate network security may seriously inhibit user confidence in the system.[94] A CDC official also noted that network communication is an area of concern, citing previous issues with the deployment of related technologies.

Data Management and Interpreting Results Are Challenging

DHS identified data management challenges for autonomous systems, including reviewing the reported data and interpreting the data to determine appropriate follow-up actions. A participant at the National Academies workshop expressed concern over a system that would provide data every few hours, leading to strain on limited and diminishing

[93]Sandia National Laboratories, *Next Generation Threat Study: Development of a Strategy and Requirements for Detecting Next Generation Biothreat Agents,* SAND2011-0085 (Albuquerque, N.M.: Jan. 2011).

[94]GAO-12-810.

public health resources. DHS describes system data as containing information on how the detector was functioning as well as laboratory analysis data. According to DHS, data from an autonomous detection system would need to be reviewed by local or state staff across the 24/7 reporting period. Further, those staff would need to be trained in appropriate data interpretation. According to DHS officials, the cost for these local public health resources is not included in their cost projections of the autonomous BioWatch calculations.

Maintaining Autonomous Detection System Operation Is Challenging

Finally, DHS officials stated that keeping the autonomous detection system continuously functioning in a dirty environment is challenging. Additionally, an LLNL official stated that a dirty environment can contain chemicals that interfere with the technology used to detect the agent. As we reported in 2012, and according to the 2011 Operational Assessment on Gen-3, autonomous detection systems during the Gen-3 acquisition experienced malfunctions, exhibited issues with the positive controls, and required unscheduled maintenance, attributed to either traffic emissions due to proximity to an interstate, or to metallic dust generated by train brakes.[95] This underscored the challenge of an autonomous detection system needing to operate in different operational environments. In addition, according to the Analysis of Alternatives conducted by the Institute for Defense Analyses in 2013, detection systems are vulnerable to vandalism and accidents.

Conclusions

BioWatch's rapid deployment in 2003—to provide early detection of potentially catastrophic aerosolized biological attacks—did not allow for sufficient testing and evaluation against defined performance requirements to understand the system's capabilities. Since that time, DHS has commissioned tests of the system, but has not defined technical performance requirements that would link test results to conclusions about the types and sizes of attack that the Gen-2 system could reliably detect. DHS has also commissioned modeling and simulation studies, but none of these studies was designed to directly and comprehensively assess what is known about the capabilities of the currently deployed system, using specific test results and accounting for statistical and other uncertainties. Finally, while DHS has addressed certain limitations in

[95]Positive controls are pre-generated samples that are intended to demonstrate that the technology can detect the targets. Failures of positive controls often occur when the expected positive detection is absent.

testing, it has not systematically tested the system against realistic conditions, and there remains potential to reduce risk and uncertainty in what is known about the system's capabilities when deployed in a real-world environment. As a result of these gaps and limitations, considerable uncertainty remains as to the types and sizes of attack that the Gen-2 system could reliably detect. DHS officials have stated that the system's operational objective is to detect attacks large enough to cause 10,000 casualties, but DHS cannot conclude with any defined level of statistical certainty that the system can reliably achieve this objective. In the wake of the cancellation of the Gen-3 acquisition, DHS is planning for technology upgrades or improvements to the Gen-2 system, and some Gen-2 equipment is nearing the end of its life-cycle and will need to be replaced if the program is to continue. However, effective and cost-efficient decisions cannot be made regarding upgrades and reinvestments if the operational capabilities of the Gen-2 system are uncertain. Assessing the operational capabilities of the Gen-2 system against technical performance requirements directly linked to an operational objective, incorporating specific test results, and explicitly accounting for statistical and other uncertainties would help ensure that decisions about future investments are actually addressing a capability gap not met by the current system and address a clear mission need.

In recent years, DHS has canceled major acquisitions that we previously found could have been more rigorous in their test design or execution, including Gen-3. The nation's ability to detect threats against its security requires judicious use of resources directed toward systems whose capabilities can be demonstrated. Applying lessons learned from the Phase I testing of Gen-3 candidate technologies, as well as incorporating the best practices we identified, may help enable DHS to mitigate risk in future acquisitions for these types of threat detection technologies. Specifically, DHS could apply them to the BioWatch program once informed decisions have been made regarding upgrades or enhancements to Gen-2.

Furthermore, DHS officials have continued to express interest in an autonomous detection capability as a possible upgrade or enhancement to Gen-2. If DHS were to pursue an autonomous detection system in the near future, PCR would be the most mature technology available. However, the extent to which the potential benefits of such a system would materialize is uncertain, because of uncertainty in the assumptions upon which these benefits depend. Additionally, pursuit of such a system faces several likely challenges. Understanding the inherent challenges to faster detection and contextualizing the benefits of autonomous detection

technologies will help decision makers make informed decisions regarding use of limited resources. BioWatch is just one biosurveillance activity used to detect potential biological threats to our national security, and as we have previously reported, because the nation does not have unlimited resources to protect the nation from every conceivable threat, it must make risk-informed decisions regarding its homeland security approaches and strategies. In July 2012, the White House released the National Strategy for Biosurveillance to describe the U.S. government's approach to strengthening biosurveillance, but it is too soon to tell what effect the strategy and corresponding implementation plan may have on determining resource allocation priorities across the interagency. As some Gen-2 equipment reaches the end of its lifecycle, DHS will need to make decisions about investing in the future of the BioWatch program. DHS initiated the BioWatch program in 2003 to address the perceived threat at the time. Since then, numerous Bioterrorism Risk Assessments have been issued, but these have been criticized for the methodology, and none has been issued in the last 5 years. Consequently, as the National Academies has noted, there is considerable uncertainty about the likelihood and magnitude of a biological attack. Investment decisions about the future of BioWatch should be guided by the agreed-upon priorities of the various stakeholders within the biosurveillance community to help ensure investments address the current threats posed by biological hazards.

Recommendations for Executive Action

To help ensure that biosurveillance-related funding is directed to programs that can demonstrate their intended capabilities, and to help ensure sufficient information is known about the current Gen-2 system to make informed cost-benefit decisions about possible upgrades and enhancements to the system, the Secretary of Homeland Security should direct the Assistant Secretary for Health Affairs and other relevant officials within the Department to not pursue upgrades or enhancements to the current BioWatch system until OHA:

- establishes technical performance requirements, including limits of detection, necessary for a biodetection system to meet a clearly defined operational objective for the BioWatch program by detecting attacks of defined types and sizes with specified probabilities;
- assesses the Gen-2 system against these performance requirements to reliably establish its capabilities; and

- produces a full accounting of statistical and other uncertainties and limitations in what is known about the system's capability to meet its operational objectives.

To help reduce the risk of acquiring immature detection technologies, we recommend that the Secretary of Homeland Security direct the Assistant Secretary for Health Affairs, in coordination with the Under Secretary for Science and Technology, to use the best practices outlined in this report to inform test and evaluation actions for any future upgrades or changes to technology for BioWatch.

Agency Comments and Our Evaluation

In written comments provided in response to our draft report, DHS concurred with our recommendations and described actions the agency is taking to address them. DHS also provided technical comments, which we incorporated as appropriate. DOE provided technical comments, which we incorporated as appropriate. CDC and DOD reviewed the draft report and provided no comments. DHS's written comments are reproduced in full in appendix IV of this report.

DHS concurred with our recommendation to establish technical performance requirements, including limits of detection, necessary for a biodetection system to meet a clearly defined operational objective for the BioWatch program by detecting attacks of defined types and sizes with specified probabilities. DHS stated the BioWatch program has already completed a series of tests that establish the performance and capabilities of currently deployed technologies and provide baseline performance requirements for any future technological improvements. DHS also stated the BioWatch program will consider including additional measures of system performance, such as probability of detection, to augment and validate the system's ability to detect attacks, pending available resources and at a time yet to be determined.

However, using existing test results as a baseline for future technological improvements provides no information about the current—or any future—system's ability to meet a clearly defined operational objective. DHS should first establish requirements for the current system, which will enable DHS to assess its system performance measures, such as sensitivity, against its stated mission goal: to detect attacks causing 10,000 casualties. Without establishing such performance requirements, the agency does not know what the existing test results mean for the system's ability to detect attacks, and thus cannot establish the benefits of any future improvements. Further, DHS mentioned using additional system performance measures to augment and validate the system's

capability. We emphasize that the program's current preferred measure of system performance, fraction of population protected (*Fp*), does not have a clear linkage to the system's operational objective.[96] What is needed is a measure that directly supports conclusions about the system's ability to meet its objective by detecting attacks of defined types and sizes.

DHS concurred with our recommendation to assess the Gen-2 system against the performance requirements described above to reliably establish its capabilities, and stated that the results of the testing and evaluation events referred to above have already been incorporated into existing modeling and simulation studies. DHS also stated, however, that should a significant difference between two performance measures—*Fp* and probability of detection (*Pd*)—be observed, the BioWatch program would consider additional modeling and simulation studies to determine the performance capabilities of the deployed Gen-2 BioWatch system.

While DHS's response suggests that it has largely met this recommendation already by having tested the system and incorporating the test results into modeling and simulation studies, this conflicts with what we have been told by BioWatch program officials. These officials told us that they have not commissioned or produced an analysis in which the best available test results have been combined with modeling and simulation to reach specific conclusions about the system's ability to detect attacks of defined types and sizes. As we detailed in the report, modeling and simulation studies commissioned by DHS either considered ranges of hypothetical values for the system's sensitivity or else involved old test results, based on an earlier version of the system, for just one of the BioWatch threat agents. Furthermore, we identified important limitations in the tests DHS has conducted that could be addressed through a more systematic approach to reducing risk and uncertainty in what is known about the system's capabilities when deployed in a real-world environment. While it is true that the system cannot be tested directly by releasing live biothreat agents in the environments where BioWatch is deployed, both the National Research Council and subject matter experts with whom we spoke identified methods by which the system can be tested and its performance characteristics estimated in more realistic environments.

[96]In written comments, DHS refers to this measure as fraction of population covered.

DHS concurred with our recommendation to produce a full accounting of statistical and other uncertainties and limitations in what is known about the system's capability to meet its operational objectives. DHS stated it already has sufficient understanding of the statistical uncertainties and limitations associated with testing and modeling of the BioWatch system. However, DHS also agreed that there is value in consolidating this information into a single, comprehensive document and plans to do so by April 30, 2016.

As described in the report, DHS does not have a sufficiently comprehensive accounting of the uncertainties and limitations in what is known about the system's capabilities. A comprehensive analysis of uncertainties and limitations should account for how such uncertainties and limitations affect the key outcome: the system's ability to meet its operational objective by detecting attacks of defined types and sizes. Statistical uncertainties should be represented with clearly defined confidence intervals; for uncertainties that are difficult to quantify, such as uncertainties associated with testing in chambers rather than operational environments, the judgments of subject matter experts may be useful. This full accounting of uncertainties and limitations should be provided to administration and congressional decision makers so they better understand the precision in what is known about the system's capabilities in an operational context. Decision makers should be able to use this information not only when comparing the costs of the current system to the benefits it may provide, but also when weighing decisions about proposed upgrades or enhancements to the system.

DHS concurred with our recommendation to use the best practices we outline in the report to inform test and evaluation actions for any future upgrades or changes to technology for BioWatch. DHS stated that changes to BioWatch will adhere to new DHS acquisition guidance that incorporates the best practices outlined in our report.

DHS's reference to new acquisition guidance is to DHS-wide guidance that was issued in 2010, after DHS began testing the Gen-3 technology. This guidance includes additional detail on factors to consider when planning and testing new acquisitions and addresses many of the practices described within this report. However, when it comes to ensuring the acquisition will not only meet technical requirements but also perform as intended in an operational environment, we believe more robust testing earlier in the acquisition to test resilience can help reduce the risk of acquiring immature technologies. This is especially important for a system like BioWatch, which cannot be fully tested in an operational

environment. While we see proper implementation of DHS's updated acquisition guidance as a positive step towards addressing our recommendation, as we reported in April 2015, DHS's Director of Operational Test and Evaluation expressed interest in becoming more involved in testing earlier in the development process. Therefore, we believe the lessons learned on Gen-3 testing and full adoption of testing practices aimed at establishing the operational performance of a system earlier in the acquisition should also be considered to help inform future DHS decisions.

While DHS concurred with the three parts of our first recommendation, the agency did not agree with key findings that led to these recommendations; therefore, it is important to address parts of their response for clarification.

DHS took exception to our conclusion that it has not defined technical performance requirements that would link test results to conclusions about the types and sizes of attack that the Gen-2 system could reliably detect. DHS stated it uses the metric called fraction of population covered (*Fp*) to make this linkage. However, when asked about this directly, agency officials declined to explain how specific values of *Fp* would enable DHS to conclude what types and sizes of attack the system can detect. Furthermore, officials said they have not commissioned or produced an analysis in which the best available test results are used to calculate *Fp* values and draw conclusions about the system's ability to detect attacks of defined types and sizes. How a given value of *Fp* would provide information about the types and sizes of attacks BioWatch Gen-2 can detect remains uncertain, and how *Fp* relates to the probability of detecting attacks large enough to cause 10,000 casualties—DHS's stated objective for the BioWatch program—remains unclear. As we note in this report, we recognize that *Fp* is a useful metric for certain purposes, but it does not directly support conclusions that align with the BioWatch operational objective.

DHS stated that it disagreed with the conclusion that the BioWatch Program does not incorporate empirical data gathered on the current Gen-2 system to inform modeling and simulation studies. However, DHS incorrectly attributed this conclusion to us. We did not state that DHS did not use any empirical data to inform their modeling and simulation studies. We stated that (1) the modeling and simulation studies did not incorporate specific, best available test results (for example, particular estimates of the system's limits of detection) to draw specific conclusions about the BioWatch Gen-2 system's capability to detect attacks of defined

types and sizes, and (2) the modeling and simulation studies did not incorporate uncertainties in the empirical test results that are important for understanding the precision or confidence in the modeling and simulation results.

Finally, DHS acknowledged the evolving threat of bioterrorism and its continued commitment to following DHS-wide acquisition policy for any future upgrade or enhancement to the current BioWatch system. Analogous to what we reported in 2012 regarding the Gen-3 acquisition, by ensuring any future upgrades or enhancements to the BioWatch system align with the earliest steps in DHS's acquisition process, such as being grounded in a justified mission need, and reflect a systematic analysis of costs, benefits, and risks, DHS can gain assurance that it is pursuing an optimal solution. Because, as DHS stated, the threat of bioterrorism continues to evolve, and because the last full Bioterrorism Risk Assessment (BTRA) was issued in 2010, it will be important for DHS to demonstrate that any proposed upgrades or enhancement address the threat posed by the intentional release of select aerosolized biological agents at the time upgrades are considered.

As agreed with your offices, unless you publicly announce the contents of this report earlier, we plan no further distribution until 30 days from the report date. At that time, we will send copies to the Secretaries of Homeland Security, Health Human and Services, Defense, and Energy; and interested congressional committees. The report is also available at no charge on GAO's website at http://www.gao.gov.

If you or your staff members have any questions about this report, please contact Tim Persons at (202) 512-6412 or personst@gao.gov or Chris Currie at (404) 679-1875 or curriec@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made key contributions to this report are listed in appendix V.

Timothy M. Persons Ph.D., Chief Scientist

Chris Currie, Director Homeland Security and Justice

List of Requesters

The Honorable Ron Johnson
Chairman
Committee on Homeland Security and Governmental Affairs
United States Senate

The Honorable Fred Upton
Chairman
Committee on Energy and Commerce
House of Representatives

The Honorable Michael T. McCaul
Chairman
The Honorable Bennie G. Thompson
Ranking Member
Committee on Homeland Security
House of Representatives

The Honorable Tim Murphy
Chairman
The Honorable Diana DeGette
Ranking Member
Subcommittee on Oversight and Investigations
Committee on Energy and Commerce
House of Representatives

The Honorable Martha McSally
Chairman
The Honorable Donald M. Payne, Jr.
Ranking Member
Subcommittee on Emergency Preparedness, Response, and Communications
Committee on Homeland Security
House of Representatives

The Honorable Claire McCaskill
United States Senate

Appendix I: Objectives, Scope, and Methodology

The objectives of this report were to discuss: (1) the extent to which the Department of Homeland Security (DHS) has assessed the technical capability of the currently deployed system (Gen-2) to detect a biological attack; (2) the extent to which DHS adhered to best practices for developmental testing during Gen-3 Phase I, and what lessons can be learned; and (3) the most mature technology for an autonomous detection system, as well as what the potential benefits and likely challenges would be if DHS were to pursue an autonomous detection system for the BioWatch program in the near future.

To determine the extent to which DHS has assessed the technical capability of the Gen-2 system to detect an attack, we reviewed and analyzed test reports and other agency and agency-commissioned documents containing information on the design, development, deployment, and technical performance characteristics of the system. We also reviewed reports of modeling and simulation studies, conducted by Department of Energy (DOE) national laboratories for DHS, that analyzed the performance and capabilities of the system. We interviewed DHS officials from the BioWatch Program Office and from the Science and Technology Directorate (S&T) who had knowledge of the history of the program, the Gen-2 technology and changes that had been made to the technology over time, and the tests and studies that had been conducted on the Gen-2 system's technical capabilities. We also interviewed officials and researchers who conducted or were familiar with the tests and the modeling and simulation studies; these included officials and researchers at Dugway Proving Ground, Sandia National Laboratories, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory. In interviews with researchers who had conducted tests and studies, we questioned them about the scope and purposes of their work; the methods they had used; conclusions drawn, as well as any caveats on those conclusions; and the strengths and limitations of the tests and studies. We conducted a site visit to Dugway Proving Ground and saw facilities and equipment that had been used to test the Gen-2 system, as well as facilities under construction that could potentially be used for future testing of the BioWatch system. To assess the strengths and limitations of tests and studies of the Gen-2 system, we used (1) a framework for testing and evaluation of biodetection systems developed by the National Research Council,[1] (2) leading practices in risk analysis

[1]National Research Council. *Review of Testing and Evaluation Methodology for Biological Point Detectors: Final Report.* (Washington, D.C.: National Academies Press, 2004).

and cost-benefit analysis,[2] and (3) judgment of internal (GAO) and selected external experts in the fields of engineering, aerobiology, microbiology, and testing and evaluation of biodetection systems. To gather information on the field of biodetection and the strengths and limitations of alternative technologies, we attended two conferences on biodetection technologies.

To determine whether DHS's actions during Gen-3 Phase I adhered to best practices for developmental testing and to identify lessons that could be learned, we reviewed the best practices previously developed in conjunction with the National Academes to assess their appropriateness to our review. We consulted with GAO specialists familiar with the best practices for developmental testing to discuss their proper application. We determined the practices that could be applied to Gen-3 Phase I testing, as the testing was developmental in nature and presented opportunities for DHS to de-risk the Gen-3 acquisition, which is the intent of the practices—to de-risk acquisitions of binary threat detection technologies by the government. We analyzed Gen-3 Phase I acquisition and testing documents, such as the test and evaluation master plans, individual test plans and results, and the operational requirements documents. We analyzed other DHS documentation on lessons learned, including the Post Implementation Review assessment, in which DHS identified its own lessons learned on the Gen-3 acquisition. We reviewed the acquisition decision memorandum on the cancellation of the Gen-3 acquisition.

We interviewed DHS officials in the BioWatch Program Office, the Office of the Director of Test and Evaluation at the Science and Technology Directorate, and officials at the national laboratories and Department of Defense (DOD) test agencies who were familiar with the testing performed during Gen-3 Phase I. We collected information from these officials on DHS's actions and decisions during Phase I testing and compared that with the recommended actions outlined in the best practices for developmental testing. We also compared the steps outlined in the test planning documents with the recommended steps described in the best practices. We consulted with internal and external experts on our

[2]Office of Management and Budget (OMB), *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, Circular A-94 (Washington, D.C.: Oct. 29, 1992). M. Granger Morgan and Max Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (New York: Cambridge University Press, 1990). OMB, *Circular A-4* (Sep. 17, 2003).

assessment of DHS's actions and decisions compared with the best practices and acquisitions more broadly. We reviewed prior GAO reports on the Gen-3 acquisition and the biosurveillance enterprise. We also reviewed prior GAO work on other DHS acquisitions that met challenges during early phases of testing to draw comparisons with other DHS acquisitions that may have benefited from more robust testing guidance.

To develop the best practices for developmental testing of binary threat detection systems,[3] we conducted a 1-day meeting on June 4, 2013, with 12 experts we selected with assistance from the National Academies.[4] These experts were from academia, industry, and the federal government and had experience in developmental testing methodologies, binary threat detection systems, automatic target recognition, and advanced imaging technologies, from fields that included homeland security, defense, and standards development. To identify the experts, the National Academies considered experts with previous experience on appropriate National Academy studies, requested suggestions from the members of the National Academies' National Materials and Manufacturing Board and the Computer Sciences and Telecommunications Board, searched internal databases and the Web, and contacted other relevant individuals for recommendations. We facilitated the experts' identification of best practices with pre-meeting interviews, structured questioning during the meeting, and post-meeting expert voting and ranking procedures. According to the experts, the best practices apply to the process of developmental testing of binary threat detection systems; they also apply if the system is commercial-off-the-shelf (COTS), modified COTS, or newly developed for a specific threat detection purpose being created by a vendor or the government.

To identify the most mature technology for autonomous detection, we reviewed a report of a 2013 workshop conducted by the National Academies that assessed the state of technologies that are potentially

[3]Binary threat detection systems indicate whether a potential threat is present or not. They do not identify gradations of threat.

[4]A list of the experts appears in GAO, *Combating Nuclear Smuggling: DHS Research and Development on Radiation Detection Technology Could Be Strengthened*, GAO-15-263 (Washington, D.C.: Mar. 6, 2015).

suitable for autonomous detection for the BioWatch program.[5] We also interviewed officials at the Centers for Disease Control and Prevention, the Department of Homeland Security's Office of Health Affairs, Lawrence Livermore National Laboratory, and the Department of Defense who were familiar with BioWatch and biodetection technologies to gather their views on the state of autonomous detection technology. A conclusion of the National Academies workshop held in 2013 was that the polymerase chain reaction (PCR) was the most mature technology suitable for autonomous detection for BioWatch. As a check for any more recent developments that might affect this conclusion, we performed a literature review of journals and conference proceedings published since 2012 to identify any technologies potentially more mature than PCR based on the following criteria:

1. whether the detection technology is specified, meaning the technology is defined and not just referred to as biodetection or detection technology;
2. capacity to detect at least bacteria and virus;
3. detection sensitivity;
4. detection specificity;
5. processing time;
6. having both indoor and outdoor performance capabilities in realistic environments;
7. ability to detect independently (standalone);
8. technology readiness level (TRL) of 6 or higher, if reported;
9. sampled from aerosols/air;
10. whether the technology is used for disease surveillance or modeling instead of pathogen detection; and
11. whether the technology depends on, or is a variant of, PCR.

We excluded press releases and news articles, studies that did not include sufficient detail for evaluating technological detection capability, technologies that were intended for non-aerosol detection (such as for

[5]Institute of Medicine and National Research Council, *Technologies to Enable Autonomous Detection for BioWatch* (Washington, D.C.: National Academies Press, 2014).

food or clinical specimen testing), or technologies that were intended to be used alongside other technologies for detection (for example, used to supplement or verify a finding, or used as a trigger warning system). Our literature review was not intended to be a comprehensive examination of all technologies that might possibly be applied to BioWatch, but rather a supplement to the National Academies workshop report and a check to help ensure that the conclusions of that workshop were not affected by more recent developments in the field.

To assess the potential benefits and likely challenges of autonomous detection, we analyzed reports published by the Sandia National Laboratories, as well as our prior work on the Gen-3 BioWatch system. We performed a literature review for models of how response timing to a positive detection of agent release may affect response effectiveness, in terms of lives saved. We searched for models published in the last 12 years, a range that was designed to cover work done following the anthrax attacks of 2001. We interviewed officials at the Centers for Disease Control and Prevention, the Department of Homeland Security's Office of Health Affairs, and Lawrence Livermore National Laboratory to gather their views on the potential benefits and likely challenges of autonomous detection in the near future, which we defined as the next 5 years. Additionally, we reviewed Gen-3 BioWatch testing reports to identify likely challenges to autonomous detection systems.

To determine the potential cost saving benefits of an autonomous detection system, we analyzed cost data provided by DHS. The agency provided annual operation and maintenance costs and total annual program costs for the current BioWatch system under the assumption that detection cycles would be increased to three per day (up from once per day, which is the current practice in most jurisdictions), as well as total annual costs for running a hypothetical autonomous detection system with six to eight detection cycles per day, with comparable coverage. The total annual cost for operating the current BioWatch system was calculated by dividing the annual operation and maintenance costs by three but keeping the remaining costs constant under the assumption that non-operation and maintenance costs remain the same.

Our analysis of potential benefits and likely challenges represent key ones that were identified by the sources listed above, and is not intended to be comprehensive. In particular we did not assess or mention characteristics that were difficult or impossible to meaningfully discuss within the context of this report (for example, deterrent effects of a biodetection system, or finding qualified personnel to hire). For benefits,

we focused primarily on reports published by Sandia National Laboratories because they focused most directly on the BioWatch program.

To help collect and analyze information for all three of our research objectives—and to help ensure the technical accuracy of our work—we consulted with subject matter experts under contract with GAO in the fields of aerobiology, microbiology, and biodetection.

We conducted this performance audit from December 2013 to October 2015 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Appendix II: The Department of Homeland Security Has Made Adjustments to the Gen-2 System Designed to Reduce False Positives

In the Gen-2 system, if the polymerase chain reaction (PCR) assays used in both the screening step and the verification step yield positive results, suggesting the presence of a BioWatch threat agent, then a BioWatch Actionable Result (BAR) is declared. From the program's inception in 2003 through 2014, there were 149 BARs. None was found to be associated with the release of a biothreat agent, and these BARs have been termed false positives by Centers for Disease Control and Prevention (CDC) officials and others.

We found that all of the BARs from 2003 through 2014 were associated with PCR assays for two biothreat agents: *Brucella* and *Francisella tularensis*. The majority were associated with the assays for *Francisella tularensis*, and these have been attributed to detections of a non-disease-causing relative, or near-neighbor, of the *Francisella tularensis* bacterium that occurs naturally in the environment. Expert stakeholders told us that, before BioWatch, scientists had no reason or occasion to assess the presence or prevalence of these naturally occurring, non-disease-causing near-neighbors.

Department of Homeland Security (DHS) officials and other stakeholders said several adjustments were made to the Gen-2 system in 2011 and 2012 to reduce the number of false positives. In August of 2011, a stricter criterion for deciding that PCR assays revealed the presence of biothreat agents was adopted.[1] Between November of 2011 and December of 2012, BioWatch adopted new PCR assays for the screening step of analysis. Previously, assays developed by CDC for use in its Laboratory Response Network (LRN) had been used for both screening and verification. The new assays were from the Department of Defense's (DOD) Critical Reagents Program (CRP) and were designed to look for different genetic signatures of the BioWatch threat agents. In general, assays designed to detect greater numbers of unique genetic signatures

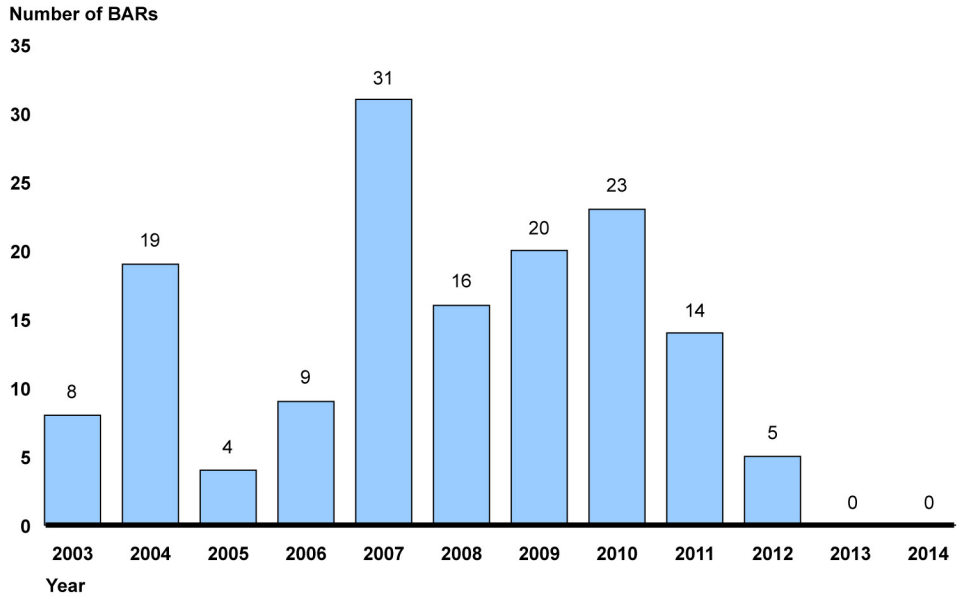
[1]This criterion is known as a cycle threshold (Ct) cutoff, and it determines how many rounds, or cycles, of "molecular photocopying" of a sample are permitted before the assay is determined to have produced a negative result. A higher Ct cutoff enables more rounds to be performed and thus generally results in an assay that has greater sensitivity (higher probability of detecting the presence of a targeted agent) but also lower specificity (higher probability of a false positive result, which could happen because PCR sometimes "photocopies" genetic material that does not belong to the target agent). By analogous reasoning, a lower Ct cutoff generally results in an assay with lower sensitivity and greater specificity.

will provide greater specificity—that is, greater ability to distinguish between the agents of interest and other, genetically similar agents. In December of 2012, BioWatch adopted new PCR assays specifically intended to distinguish between disease-causing and non-disease-causing species of *Francisella*; under the new analysis protocol, these new assays are run in the verification step if the screening step returned a positive result for *Francisella*.

Another adjustment to the system was not made in order to reduce the number of false positives but likely had this effect. In March, 2008, the PCR assays for *Brucella* were discontinued. According to BioWatch officials, this was because CDC had reclassified *Brucella* into a lower-threat category. Some of the BARs prior to March of 2008 were associated with the PCR assays for *Brucella*, and such BARs were no longer possible after this agent was discontinued as a BioWatch threat agent.

The annual number of BARs decreased during the years when the adjustments designed to reduce false positives were made, and after the final adjustment (the adoption of the new *Francisella* assays) there were no BARs through 2014 (see fig. 6). This decrease is consistent with the possibility that the adjustments have provided greater specificity, as intended; however, there was large, unexplained variability in the annual numbers of BARs in earlier years, and we did not conduct an independent analysis to assess the extent to which the decrease since 2010 might be associated with the specific adjustments DHS made to the system.

Figure 6: Number of BioWatch Actionable Results (BAR) per Year, 2003–2014



Source: Department of Homeland Security. | GAO-16-99

According to a recent report by the National Academies, there is no once-and-for-all solution to the problem of false positives for a biodetection system based on PCR assays.[2] This is because biological agents continue to evolve, and new strains and near-neighbors continue to arise. Consequently, a biodetection system based on PCR assays, such as BioWatch, will likely require ongoing adjustments to manage or prevent false positives, and the National Academies recommended that the BioWatch program continue to test its assays against panels of near-neighbors as these panels are reviewed and updated over time.

[2]National Research Council and Institute of Medicine, *BioWatch PCR Assays: Building Confidence, Ensuring Reliability* (Washington, D.C.: National Academies Press, 2015).

Appendix III: Best Practices for Developmental Testing

In 2013, in collaboration with the National Academies, we identified eight best practices for developmental testing of threat detection systems.[1] According to the experts who deliberated on the best practices, the practices apply to the process of developmental testing of binary threat detection systems; they also apply if the system is commercial-off-the-shelf (COTS), modified COTS, or newly developed for a specific threat detection purpose being created by a vendor or the government. For additional information on the methods used to determine the best practices, see appendix I.

The identified eight best practices for developmental testing of binary threat detection systems are described below, often using the context of the BioWatch program.

Practice 1: Ensure that accountability and engagement in developmental testing are commensurate with the amount of risk accepted.

According to experts, the level of government involvement in the development of a given system should be commensurate with the level of risk it is accepting.[2] Risk needs to be assessed when the system is COTS, modified COTS, or a newly developed system, because even with commercial items, significant modifications may be needed. Experts also told us that relying solely on the vendor and holding the vendor responsible for any problem that arises is not consistent with the accountability and engagement required for acquisitions where the government is accepting significant risk. Further, an understanding of the technical risk associated with the development of a given system is important, or in the case of the purchase of a commercial item,

[1]GAO, *Combating Nuclear Smuggling: DHS Research and Development on Radiation Detection Technology Could Be Strengthened,* GAO-15-263 (Washington, D.C.: Mar. 6, 2015).

[2]Experts referenced are those who collaborated to establish the list of best practices used in this report.

understanding the modifications needed to an existing system to accommodate government-specific needs, is important.[3]

Practice 2: Include representatives from the user community in design and developmental testing teams to ensure acceptance of the system by the user community.

Design and developmental testing teams need to understand the needs, concerns, and capabilities of the user community or they run the risk of designing and testing a system that, in the case of autonomous biological detection systems, (1) operators may have difficulty operating, or (2) decision makers may have difficulty with when interpreting results. According to experts, the user community may have suggestions that could improve the system or make the developmental tests more realistic. Distinct from subject matter experts that monitor developmental testing, these representatives are integral parts of the design and developmental testing teams. The role of these team members is to make sure that the needs, concerns, and capabilities of the user community are considered throughout design and developmental testing efforts.

Practice 3: Take a systems engineering view of the system prior to entering into any developmental test.

According to experts, to take a systems engineering view of a system, the tester must understand the boundaries of what it is being tested prior to developmental testing. For example, would DHS plan to test just the assays or analytical components of the Gen-3 candidate systems or would it plan to test the whole end-to-end system (i.e., collection, extraction, analysis, and communication of result), and was that plan communicated in the test and evaluation master plan (TEMP)? This is critical, since different system boundaries impose different testing methods and constraints.

Practice 4: Use statistical experimental design methodology to establish a solid foundation for developmental testing.

According to experts, use of statistical experimental design methodology ensures that a test has been designed with a clear understanding of goals and acceptable limitations, that the test is clearly documented, and that the test results are rigorously analyzed. Experts said statistical experimental design is the tool used to define the test goals, limitations,

[3]We define technical risk as the risk that a system will provide the required performance in the required time frame utilizing specified resources. For commercial items, this means that the required performance already exists, it does not have to be developed, or modifications to the items are minor; the availability is limited only by production time; and specified resources are the negotiated price. This is consistent with internal controls in the federal government, which advocates use of risk assessment in federal activities, as well as DOD risk management guidance, which states that the level of government involvement in the development of a given system should be commensurate with the level of risk it is accepting. See Department of Defense*, Risk Management Guide for DOD Acquisition*, Sixth Edition(Version 1.0), August 2006.

Testing for Resilience

Resilience means resilience against failures. It entails building robustness into the system by eliminating as much vulnerability as possible so that the system performs according to requirements, even when faced with unforeseen obstacles.

Source: GAO. | GAO-16-99

and procedures, and establishes a detailed plan for conducting the experiment. Further, experts stated that the creation and use of an appropriate model against which system performance can be evaluated is fundamentally important when establishing the statistical experimental design. According to experts, the system's operational objectives and user's needs should be identified before designing the experiment, and uncertainties should be characterized and reported with all system performance estimates. Experts told us that well-chosen statistical experimental designs maximize the amount of information that can be obtained for a given amount of experimental effort.

Practice 5: Measure and characterize system performance with established procedures, methods, and metrics.

According to experts, binary threat detection systems have an established body of statistically based methods and procedures used to evaluate and characterize them. Further, experts stated it is important to use certain objective metrics to characterize system performance.

Practice 6: Test to build in resilience, especially in the development stages.

According to experts, one way to improve resilience is to uncover vulnerabilities as early as possible through rigorous and comprehensive testing of the system against various scenarios. For example, the BioWatch system operates in a number of different environmental settings with varying contaminants. Settings vary from warm to cold climates, dry to wet climates, and indoor and outdoor settings. The BioWatch sensors might be exposed to dust, metallic dust, smoke, and diesel exhaust in indoor environments, as well as to rain, fog, snow, ice, wind, salt spray, sand, and pollen in outdoor environments. To the extent possible, these potential contaminants should be part of the testing of a detection system to help identify vulnerabilities to performance in these environments.

According to experts, the further the system moves down the development path, the more fixed the design becomes. Thus, when the developmental testing team uncovers an error (i.e., the system failed a test), it is increasingly expensive to fix. Any time there is a change in the design, everything that worked before needs to be re-tested to make sure the change did not undo something that already has been shown to work. Therefore, according to experts, agencies should focus on building in resilience during early and intermediate developmental testing so as to minimize the number of hidden failures found in the later stages of testing.

Practice 7: Use developmental tests to refine performance requirements.	According to experts, developmental testing should be viewed as a critical tool in helping to refine performance requirements. Experts told us that a meaningful performance requirement is one that not only is achievable but also strives to maximize the fulfillment of a mission need—which in the case of BioWatch might be the number of lives saved. While the minimum required performance thresholds may be achievable, they may fall short of the maximum achievable performance. Experts informed us that the maximum achievable performance can be uncovered only by understanding what the system is actually capable of doing through comprehensive developmental testing that unrestrainedly explores the performance boundaries of the system.
Practice 8: Engage in a continuous cycle of improvement by (1) conducting developmental testing, (2) conducting operational testing, and (3) incorporating lessons learned.	Experts told us that it is important to consider developmental testing and operational testing as a continuum by defining developmental testing broadly to cover, for example, operational test activities that traditionally have been viewed as post-development, rather than artificially limiting the development of a system to a fixed stage. Experts also said it is important to use lessons learned on preceding tests to improve the probability of success (proper system performance) on following tests and to use lessons learned from test failures as feedback into the design process to continuously improve system performance.

Appendix IV: Comments from the U.S. Department of Homeland Security

U.S. Department of Homeland Security
Washington, DC 20528

Homeland Security

September 28, 2015

Chris P. Currie
Director, Homeland Security and Justice
U.S. Government Accountability Office
441 G Street, NW
Washington, DC 20548

Re: Draft Report GAO-15-736, "BIOSURVEILLANCE: DHS Should Not Pursue Biowatch Upgrades Until System Capabilities Are Established"

Dear Mr. Currie:

Thank you for the opportunity to review and comment on this draft report. The U.S. Department of Homeland Security (DHS) appreciates the U.S. Government Accountability Office's (GAO) work in planning and conducting its review and issuing this report.

DHS recognizes the time and effort GAO has devoted to better understanding the sometimes complex intersection of homeland security and public health. Bioterrorism remains a continuing threat to the security of our nation. The threat to our environment is constantly evolving and the early detection of a biological attack is an essential part of an effective biodefense posture.

Biological attacks can begin without overt signs; therefore early warning and detection capabilities are essential for mitigating consequences and saving lives. The BioWatch Program is the Nation's only federally managed, locally operated nationwide biosurveillance system designed to detect the intentional release of select aerosolized biological agents. The program provides public health officials with a warning of potentially hazardous biological agent release before exposed individuals would typically develop symptoms of illness. BioWatch is a key part of the Nation's layered approach for protection against a catastrophic biological terror attack, and each hour gained in detection translates into lives saved. The Office of Health Affairs (OHA), which oversees the BioWatch Program, is committed to providing medical and health expertise in support of the DHS vision to ensure a homeland that is safe, secure, and resilient against terrorism and other hazards.

It is important to note, however, that DHS does not necessarily agree with all of GAO's characterizations of the BioWatch efforts described in this report, even though the Department concurs with the three recommendations in the report. For example, DHS takes exception to GAO's conclusion that "DHS has not defined technical performance

requirements that would link test results to conclusions about the types and sizes of attacks." To the contrary, the BioWatch Program uses the measure "*Fraction of Population Covered*" to make this linkage, which considers the known sensitivity of the BioWatch Gen-2 system and focuses on maximizing the number of lives protected in multiple attack scenarios.

In addition, DHS disagrees with GAO's conclusion that the BioWatch Program does not incorporate empirical data gathered on the currently deployed Gen-2 system to inform modeling and simulation studies. To the contrary, BioWatch sponsors modeling and simulation studies that use a range of empirically derived sensitivity values and evaluate every component of the Gen-2 system to provide confidence in the precision and reliability of the system's capabilities.

While there may not be full agreement of the Gen-2 system performance, which was fielded prior to the implementation of DHS-wide acquisition policy and processes, DHS leadership will ensure that the system's replacement or recapitalization complies with this guidance. This will require full vetting and approval of the operational and performance requirements, analysis of alternatives and all required acquisition documentation before moving forward, as appropriate.

The draft report contains four recommendations, with which the Department concurs. Specifically, GAO recommends that the Secretary of Homeland Security direct the Assistant Secretary for Health Affairs and other relevant officials within the Department to not pursue upgrades or enhancements to the current BioWatch system until OHA:

Recommendation 1: Establishes technical performance requirements, including limits of detection, necessary for a biodetection system to meet a clearly defined operational objective for the BioWatch program by detecting attacks of defined types and sizes with specified probabilities.

Response: Concur. The BioWatch Program has already completed a series of tests and evaluations that establish the performance and capabilities of currently deployed technologies. These tests include chamber testing against killed agents and an operational demonstration against live simulants. Results of these tests provide the baseline performance requirements for any future technological improvements. The BioWatch Program will also consider, pending available resources, including additional system performance measures, such as the *Probability of Detection* that the draft report references, to augment and validate the system's ability to detect attacks. Estimated Completion Date (ECD): To Be Determined (TBD).

2

Recommendation 2: Assesses the Gen-2 system against these performance requirements to reliably establish its capabilities.

Response: Concur. The results of testing and evaluation events referred to in the response to Recommendation 1 above have already been incorporated into existing modeling and simulation studies. However, should a significant difference between *Fraction of Population Covered* and *Probability of Detection* be observed, the BioWatch Program will consider additional modeling and simulation studies to determine the performance capabilities of the operationally deployed Gen-2 BioWatch System relative to the new operational objective in specific BioWatch jurisdictions. ECD: TBD.

Recommendation 3: Produce a full accounting of statistical and other uncertainties and limitations in what is known about the system's capability to meet its operational objectives.

Response: Concur. The BioWatch Program believes it already sufficiently understands the statistical uncertainties and limitations associated with fully testing and documenting the system capabilities, as described in test plans and guidance documents that have been provided to the GAO. However, the Program agrees that there is value in consolidating this information into a single, comprehensive document, which it will do. It is important to recognize levels of uncertainty and limitations are inherent in any complex technical system and the BioWatch system has undergone extensive testing and evaluation during the last decade and achieved proven operational success in the detection of airborne biological organisms of public concern, which can speed response to and recovery from a bioterrorism event. ECD: April 30, 2016.

In addition, GAO recommends that the Secretary of Homeland Security direct the Assistant Secretary for Health Affairs in coordination with the Under Secretary for Science and Technology to:

Recommendation 4: Use the best practices outlined in this report to inform test and evaluation actions for any future upgrades or changes to technology for BioWatch.

Response: Concur. Any future upgrades or changes to BioWatch technology will adhere to the new acquisition management guidance provided by the DHS Office of Program Accountability and Risk Management, which incorporates best practice elements outlined in this report. ECD: TBD.

3

Again, thank you for the opportunity to review and comment on this draft report. Technical comments were previously provided under separate cover. Please contact me if you have any questions. We look forward to working with you in the future.

Sincerely,

Jim H. Crumpacker, CIA, CFE
Director
Departmental GAO-OIG Liaison Office

4

Appendix V: GAO Contacts and Staff Acknowledgments

GAO Contacts

Timothy M. Persons, (202) 512-6412 or personst@gao.gov
Chris Currie, (404) 679-1875 or curriec@gao.gov

Staff Acknowledgments

In addition to the individuals named above, Edward George (Assistant Director), Sushil Sharma (Assistant Director), Russ Burnett, Kendall Childers, Eric Hauswirth, Hayden Huang, Susanna Kuebler, Jack Melling, Jeff Mohr, Rebecca Shea, and Katherine Trimble made key contributions to this report.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (http://www.gao.gov). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to http://www.gao.gov and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, http://www.gao.gov/ordering.htm.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on Facebook, Flickr, Twitter, and YouTube.
Subscribe to our RSS Feeds or E-mail Updates.
Listen to our Podcasts and read The Watchblog.
Visit GAO on the web at www.gao.gov.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Website: http://www.gao.gov/fraudnet/fraudnet.htm
E-mail: fraudnet@gao.gov
Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Katherine Siggerud, Managing Director, siggerudk@gao.gov, (202) 512-4400, U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, DC 20548