Project Report
ACTA-3

COVID-19 Exposure Notification in Simulated Real-World Environments

M.C. Schiefelbein
R.C. Gervin
J. St. Germain
S.L. Mazzola

15 April 2022

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001.

This report is the result of studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

Massachusetts Institute of Technology
Lincoln Laboratory

COVID-19 Exposure Notification in Simulated Real-World Environments

M.C. Schiefelbein
Group 46

R.C. Gervin
Group 21

J. St. Germain
S.L. Mazzola
Group 76

Project Report ACTA-3

15 April 2022

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001.

Lexington Massachusetts

This page intentionally left blank.

ABSTRACT

Privacy-preserving contact tracing mobile applications, such as those that use the Google-Apple Exposure Notification (GAEN) service, have the potential to limit the spread of COVID-19 in communities, but the privacy-preserving aspects of the protocol make it difficult to assess the performance of the apps in real-world populations. To address this gap, we exercised the CovidWatch app on both Android and iOS phones in a variety of scripted real-world scenarios, relevant to the lives of university students and employees. We collected exposure data from the app and from the lower-level Android service, and compared it to the phones' actual distances and durations of exposure, to assess the sensitivity and specificity of the GAEN service configuration as of February 2021. Based on the app's reported ExposureWindows and alerting thresholds for Low and High alerts, our assessment is that the chosen configuration is highly sensitive under a range of realistic scenarios and conditions. With this configuration, the app is likely to capture many long-duration encounters, even at distances greater than six feet, which may be desirable under conditions with increased risk of airborne transmission.

This page intentionally left blank.

ACKNOWLEDGEMENTS

The efforts described in this report would not have been possible without the help of our MIT Lincoln Laboratory colleagues: Dave Maurer, Keegan Quigley, John Courtney, Bill Estabrook, Dave Sciuto, Stacy Zeder, Elizabeth Bernardo, Brent Casella, Adam Norige, and Paula Ward. Janet McIllece, Kacey Ernst, and Bruce Helming at the University of Arizona guided our approach to designing test scenarios. Our partners at WeHealth, Google, and Apple provided invaluable engineering support through granting us "developer" access and through technical discussions. Finally, we are grateful to John Hynes and McKinley Theobald of the MBTA for hosting our tests.

This page intentionally left blank.

TABLE OF CONTENTS

This page intentionally left blank.

LIST OF ILLUSTRATIONS

This page intentionally left blank.

LIST OF TABLES

This page intentionally left blank.

1. BACKGROUND

The Exposure Notification (EN) service implemented by Apple and Google in 2020 was intended to roughly estimate the aggregate risk of COVID-19 exposure for the individual carrying the phone on which the service is enabled. As different communities experience different levels of community spread of the virus, depending on factors such as current infection levels, personal mobility, and participation in masking and social distancing behaviors, the EN service was designed to be configurable by regional public health authorities. In seasons of high transmission and high risk, such as a winter surge, the service could be configured to be more sensitive, e.g., alerting a user of exposure after a shorter exposure at a moderate distance. In seasons of lower overall risk, such as after a benchmark percentage of the population is vaccinated, the service could be configured to provide a lower-level alert for a greater range of exposure distances and durations, by raising the bar for issuing high-risk alerts.

The configuration is composed of several weights and thresholds, which assign encounters into "immediate", "near", "moderate", and "other" distances, weighted by the overall duration at each range, and assign it a "high", "standard", or "none" infectiousness rating based on the date of exposure and the date of symptom onset (or test) for the sick person [1], [2]. Designing an EN configuration to have the desired sensitivity and specificity is closer to an art than a science, but scientifically estimating the effectiveness of a candidate configuration is possible. Our prior EN-prototype testing and data collection efforts had formed an essential foundation of testbed infrastructure, data processing tools, and team experience [3]. In February and March of 2021, we conducted a series of experiments with the University of Arizona's selected configuration and the CovidWatch app, which is used by faculty, staff, and students at the university to reduce the spread of SARS-CoV-2 in the campus community [4].

The Exposure Notification service relies on Bluetooth LE messages to estimate distance between phones, which causes the accuracy of its estimates to be affected by anything that affects the strength of the radio signal. This includes environmental factors, such as the orientation of the phones' antennas in space; the way a person carries a phone, such as in a pocket or backpack; the presence of RF-absorbing objects, such as people and furniture; and the presence of reflective surfaces, such as metal or tile walls. Phones may experience interference from other signals in the same frequency band, or from noise due to microwave ovens and other electromagnetic fields, that cause them to underestimate the duration of exposure. Finally, the hardware diversity in the cell phone market contributes to variations in signal strength estimates even when all other factors are held constant. The EN implementers included a calibration offset for each phone model, but not all models are well calibrated, so the calibration confidence estimate for each model must be considered as well [5]. For all of these reasons, it is important to assess the performance of the system not only under laboratory conditions, but in a variety of physical environments and physical configurations, with a variety of hardware.

This page intentionally left blank.

2. EXPERIMENTAL DESIGN

The design of our experiment needed to accommodate a range of environments, distances, durations, phone carriage states, and hardware models, yet be doable in about a month. To strike a balance between variables we could not control (such as environmental factors and hardware calibration quality), and those we could (phone placement, hardware selection, distance, and duration), we chose nine test scenarios that were the highest priority for the University, and ran three trials in each. The trials for each scenario differed only in the substitution of different hardware for the sick phone and one of its neighboring phones. (For example, if the first trial used an iPhone as its sick phone, the second would swap the iPhone with its nearest low-calibration Android neighbor, and the third trial would swap for the nearest high-calibration Android.) We would combine the results across all three trials for each scenario.

The selected scenarios were a mix of learning spaces, public transit situations, and social situations, involving as many mannequins and humans as we could physically accommodate in the space with "some" social distancing applied. For the classroom and public transit scenarios, we followed spacing and mobility guidelines [6] and the class or shuttle schedule recommended by the University. For the social scenarios, we modeled two "parties" in a smaller, two-room setting, and in a larger, single-room setting, with more moving participants than in either the classroom or transit scenarios. Appendix B: Scenario Configurations contains a gallery of test scenario photos, placement maps, and motion notes. Our testing was conducted at the Autonomous Systems Development Facility (ASDF) [7] and in lecture halls and conference rooms at MIT Lincoln Laboratory, as well as at the Massachusetts Bay Transportation Authority (MBTA) Emergency Training Center in Boston, MA.

Figure 1: RF Mannequins in the ASDF.

In order to test with an appropriate number of bodies, while complying with workforce COVID-19 safety requirements, we used mannequins covered in "lossy" foam as stand-ins for actual people. These previously were shown to have the approximate RF absorption of a human body, in tests conducted at our anechoic chamber facility in April 2020. The mannequins were mounted on robotic platforms to add autonomous movement capability (Figure 1).

All scenes had a mix of carriage states for phones (in bag, in pocket, in hand, on desk), and this was not varied across the three trials. All test durations were capped at about one hour, as that was known to be more than enough time to generate an exposure alert if the phones were "close enough" to merit one. Although we were constrained by the number of mannequins, humans, and phones we could include in each test, we would still be able to conduct realistic exposure scenarios.

TABLE 1

Physical Configurations of Test Scenes

		Population	Motion	Carriage	Exposure Time
Learning Spaces	Active Learning Space	9	2 mobile	Hand, desk, bag, pockets	50 min
	Small Hall	13	All static		60 min
	Auditorium	16	All static		50 min
Public Transit	Bus Queue	3-6	5 mobile (static sick)	Pockets	10, 20, 30 min
	MBTA Bus	15	3 mobile	Hand, bag, pockets	25, 45 min
	MBTA Train	15	3 mobile		15, 30 min
	MIT Shuttle	14	1 mobile (sick)		44 min
Social	Small Party	8	1 mobile (sick)	Hand, bag, pockets	60 min
	Larger Party	16	4 mobile (1 sick)		60 min

For these experiments, we used only a single logical configuration of the app; that is, the set of weights and thresholds deployed by the University in February and in use by its campus population. The CovidWatch app differs from more basic EN apps by its use of two alert thresholds, with different follow-on guidance for each, tailored to the infectiousness estimate of the encounter [8]. Our analysis considers the probability of a low- or a high-risk alert being generated for each phone exposed to the sick phone. We did not evaluate the infectiousness estimate or follow-on guidance.

We included iPhones, although they provided no low-level beacon logs, and for those we relied on an intermediate output from the EN Service to calculate risk scores after exposure to the trial's "sick phone".

Because the contribution of infectiousness risk to the final risk score is independent of the contribution of duration and signal strength, and can be varied mathematically at a later date, we kept the infectiousness constant in all tests. This permitted us to focus on the effects of physical configurations and the distance/duration components of the logical configuration.

3. DATA COLLECTION PROCEDURE

Prior to each exposure trial, the phones' records of previous exposures were erased from the Exposure Notification service and from the app. Each phone was placed in position in a mannequin's hand, pocket, bag, or desk area. If a human test operator was participating in the test as a "body", that person's phone was placed at their assigned location. The beginning of each exposure period was recorded from the phones' system clocks along with the time it took to activate EN on all phones. This activation time took under 5 minutes in each trial, usually only 1–2 minutes.

During the exposure period, humans and/or mannequins moved through the test space as scripted for each scenario. The overall amount of motion was roughly intended to correspond to reality; because the EN sampling rate is on the order of 2–5 minutes, however, it was not necessary to have the phones in near-constant motion. Classroom scenarios were largely stationary, while transit and social scenarios included more motion. We also varied whether the sick phone was mobile or stationary, to generate different exposure durations for tests with higher physical constraints (e.g., bus seating). In the transit scenes, the total number of phones/bodies increased or decreased during the exposure period, such as when people joined or left the bus queue, and boarded or left the bus or train during the "trip". These changes were scripted and performed at the same time for each of the three trials in those tests.

At the end of each exposure period, we deactivated EN on each phone in the same order in which it was activated. We then shared the Temporary Exposure Keys from the sick phone with each of the exposed phones, always using a "high infectiousness" date, and recorded whether the phone showed an alert to the user. The app vendor, WeHealth, had provided a function to export the ExposureWindows reported by EN, in JSON format. We saved the JSON file from each exposed phone and generated a copy of the Android system logs, which recorded the low-level timestamps and signal strength measurements of the beacons heard by each phone, and the confirmation of the key matching procedure. The iPhones did not have this low-level logging available, so on those phones, we saved only the JSON file with ExposureWindows. These data files, in combination with our ground truth positional and time data, would inform our analysis of the system performance.

The dataset from this collection exercise is available on GitHub for others to use in their research[1] [9].

[1] Note that the "PACT Exposure Notifications Beacons" database includes data from both this and a prior data collection campaign. The data from the experiments described here are labeled with `test_series=MITLL_UA` in the database.

This page intentionally left blank.

4. ANALYSIS

At the end of the data collection exercise, by using a single "sick phone" in each trial scenario, we obtained the ExposureWindows, risk score, and alert status of only 333 "exposed" phones in total. Having a rich beacon dataset from the Androids enabled us to expand the analysis to consider *every* phone as a potential sick phone. We wrapped the risk scoring algorithm published in Google's open-source repository [10] to build a tool which computes a risk score for all possible pairwise exposures in our dataset. Where possible, we inferred which RPIs came from an iPhone by process of elimination. Constrained only by the lack of beacon records for iPhones, and a few cases of beacon logging failure on Android, we achieved an order of magnitude increase in the number of samples: 2991 of the 4098 potential pairwise exposures.

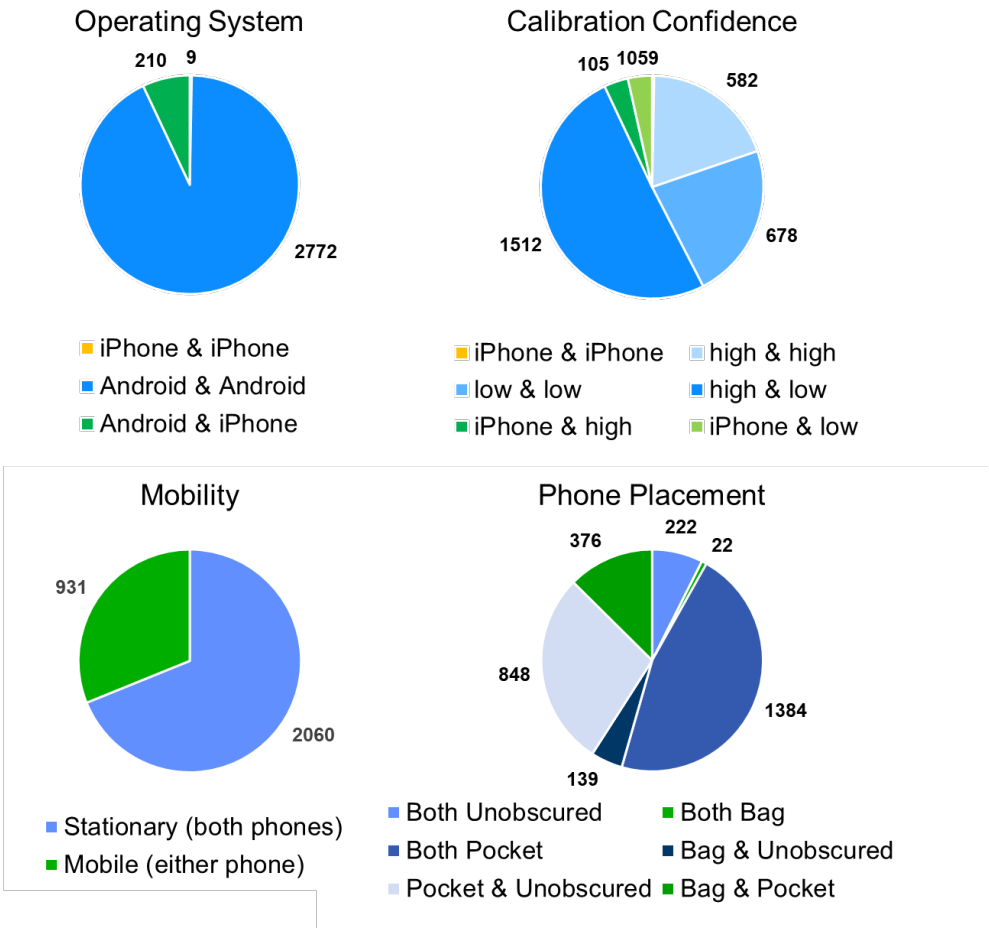


Figure 2. Phone pair demographics (all scenes combined).

As shown in Figure 2, the sample set is heavily skewed towards 2-Android pairs. Among the Androids, we tried to have an equal number of high-confidence and low-confidence models in each test. This resulted in high-low pairs making up about half the sample set overall. Likewise, the iPhone-Android pairs were equally split between high and low calibration confidence.

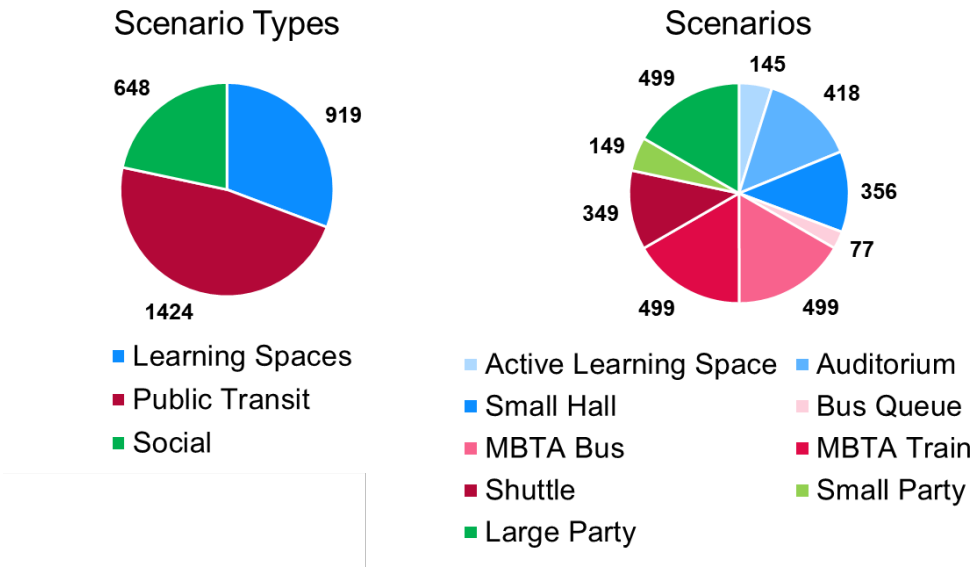


Figure 3. Scenario contributions.

In order to examine the behavior of individual phones and how it related to ground truth observations, we found it helpful to build a visualization tool. The tool combined our positional and temporal data with the beacon logs and alert reporting from each phone, and allowed us to scan forward and backward in time to see how individual signals (or lack thereof) contributed to the overall behavior of the phone. The visualization tool is described in more detail in Appendix A: Exposure Visualization Tool, and is available on GitHub [11].

Our analysis aimed to answer two questions:

1. Does the system produce the expected alert on an exposed phone, given the record of beacons heard?

2. Does the system produce an appropriate alert on an exposed phone, given the ground truth distance and duration of exposure?

To answer the first question, we focused on the 333 experimental exposures for which key matching was performed (not predicted in postprocessing). We validated the observed behavior of each Android phone (i.e. the presence of a high alert, low alert, or no alert after key matching) by recomputing the ScanInstances and ExposureWindows from the beacon data recorded in the system logs. These intermediate data structures were used to recompute the expected risk score of each encounter, and the low and high

thresholds were applied to determine the expected alert status of each phone. For iPhones, which did not have individual beacon data, we fed the reported ExposureWindows into the risk scoring algorithm to check the math. We did observe some instances in which the app displayed the incorrect risk level to the user, and determined through further analysis and conversation with WeHealth engineers, that these were artifacts of the test and export procedure and should not be observable by the real world users. For the rest of our analysis, we worked with the mathematically correct risk derived from the phone's logs, rather than the incorrect one, in these cases.

TABLE 2

Percentage of "True" TC4TL Events

Learning Spaces	Active Learning Space	14%	4%
	Small Hall	0%	
	Auditorium	5%	
Public Transit	Bus Queue	29%	18%
	MBTA Bus	12%	
	MBTA Train	8%	
	MIT Shuttle	39%	
Social	Small Party	41%	23%
	Larger Party	19%	
All Scenes Combined			15%

We answered the second question by interpreting each phone's behavior in the context of the CDC's "too close for too long" (TC4TL) contact tracing metric, namely: Was the phone really within 6 feet of the sick phone for at least 15 minutes during the exposure period [12]? Applying this metric allows us to calculate the probability of detection, or P(D), and probability of false alarm, or P(FA). We calculated these values for each of the alert thresholds, and for each of the scenarios as well as for the entire set as a whole (Table 3 and 4). Note that in the small lecture hall, the sick phone was placed with the "professor" at the podium, well over the 6 foot threshold for all students; therefore there are no "true detection" events.

TABLE 3

Probability of Detection and Probability of False Alarm for High Alert

High Alert Only		P(D)	P(FA)	Combined P(D)	Combined P(FA)
Learning Spaces	Active Learning Space	0.850	0.888	0.875	0.610
	Small Hall	n/a	0.652		
	Auditorium	0.900	0.485		
Public Transit	Bus Queue	0.591	0.109	0.953	0.569
	MBTA Bus	1.000	0.605		
	MBTA Train	1.000	0.438		
	MIT Shuttle	0.978	0.897		
Social	Small Party	0.951	0.830	0.919	0.707
	Larger Party	0.914	0.691		
All Scenes Combined				**0.935**	**0.610**

TABLE 4

Probability of Detection and Probability of False Alarm for Low or High Alert

Low or High Alert		P(D)	P(FA)	Combined P(D)	Combined P(FA)
Learning Spaces	Active Learning Space	0.950	0.936	0.950	0.746
	Small Hall	n/a	0.756		
	Auditorium	0.950	0.678		
Public Transit	Bus Queue	1	0.764	1	0.827
	MBTA Bus	1	0.802		
	MBTA Train	1	0.784		
	MIT Shuttle	1	0.986		
Social	Small Party	1	0.886	0.960	0.876
	Larger Party	0.943	0.876		
All Scenes Combined				**0.982**	**0.809**

Recall that the EN system can only alert if a sick person exists and has opted to share their keys; the "false alarms" are false only in that the exposure did not meet the TC4TL criteria, given that the exposure was to a diagnosed person. In reality, the prevalence of disease in the population will also be a factor in the perceived false alarm rate, as most encounters will be with uninfected individuals.

We believe the P(D) and P(FA) analysis is overly reductive, in the context of the broader question of whether the "6 feet and 15 minutes" rule is too simplistic to be an appropriate guideline for estimating the risk of airborne transmission of SARS-CoV-2. Therefore, we also produced a set of heat maps, showing the distribution of the average risk scores (before alert thresholds were applied) across a two-dimensional "contact grid" of ground truth distances and durations. We also produced heat maps for the probability of high alert, and the probability of low or high alert, for each scenario and across all scenarios.

This page intentionally left blank.

5. RESULTS

To build the contact grid, we assigned each encounter to a bin, using increments of 3 feet for distance and 5 minutes for duration. For phones which moved during the test, we assigned them to the (x, y) cell that represents the weighted average of their (distance, duration) tuples. For instance, if a moving mannequin spent 10 minutes while 12 feet away from the sick phone, followed by 15 minutes while 4 feet away from the sick phone, it was assigned to the cell representing (7.2 ft, 25 min). The contact grid, shown in Figure 4. Sampling distribution on contact grid (all scenes, all phones), does not have a uniform distribution, nor do the different scenarios do not contribute equally to the populated cells. During the experimental design phase, the University agreed that was an acceptable tradeoff for variety within the time available for testing, and realistic placements within each scene.

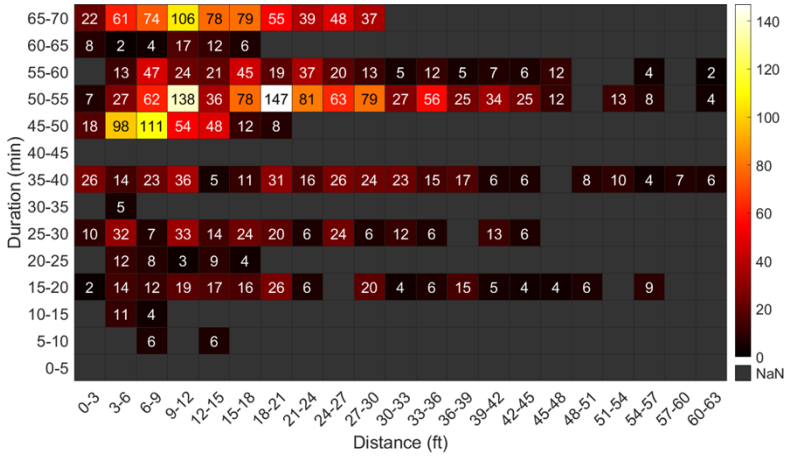


Figure 4. Sampling distribution on contact grid (all scenes, all phones).

Figure 5 shows the average risk score of each of the phone exposures assigned to the cells of the contact grid. The CDC "TC4TL" metric [12] is outlined in blue in the upper left. As we expected, there is a general trend of higher average risk scores in the upper left, in the (shorter distance, longer duration) cells. Figure 6 and Figure 7, the heat maps of the P(Alert High) and the P(Alert Low or High), respectively, show a similar gradient. This confirms that on average, the Bluetooth signal strengths and timestamps, when processed through the risk scoring algorithm and the weights and thresholds preselected by Arizona, do in fact behave as an approximate measure of cumulative exposure-minutes.

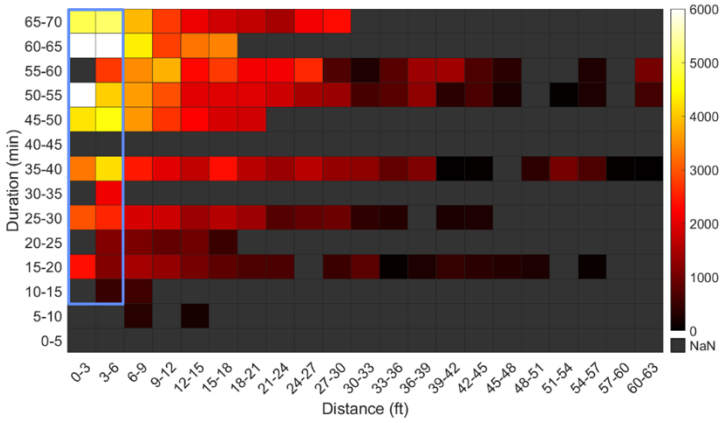


Figure 5. Average Risk Scores (all scenes, all phones). All values >= 6000 are assigned the maximum color (white).

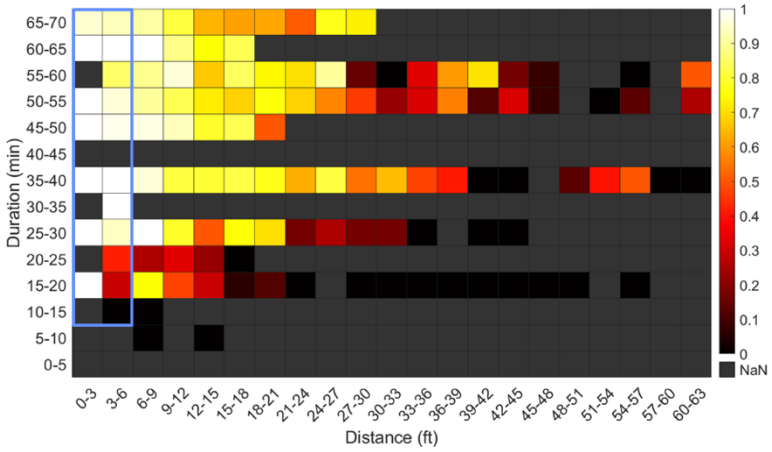


Figure 6. P(Alert $_{High}$) (all scenes, all phones).

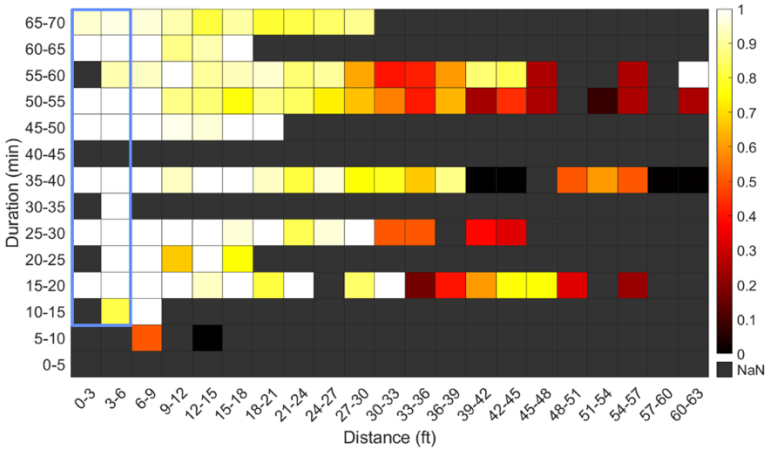


Figure 7. P(Alert $_{Low\ or\ High}$) (all scenes, all phones).

The gradients on these three heat maps are not smooth; for instance, in the (27–36 ft, 55–60 min) range, we see low risk scores to the left of a cell with a fairly high average at greater distance. Likewise, a discontinuity occurs in the 35–40 minute row at distances greater than about 40 feet. These local variations in the gradient are a combination of the non-uniform contributions of phone placement, hardware calibration, and sheer bad luck. When we inspected the phone pairs which contribute to these cells, we found instances where the exposed phone simply did not "hear" the sick phone, most likely due to interference or absorption. We expect that these discontinuities would be averaged out with more samples and a more uniform distribution of the physical experimental variables across the contact grid.

To explore whether the discontinuities could be attributed to a single factor of placement, calibration, or mobility, we examined the regions with discontinuities after "slicing" the dataset across those attributes. One might reasonably suspect the low-calibration Androids to contribute significantly to noise in the risk score and P(Alert). To our initial surprise, the heat maps for the calibration slicing showed that the low-low calibration pairs did not contribute to the discontinuity in two of the three regions. When we sliced for phone placement, we also could not identify a clear effect from phones in bags or pockets (compared to in hand or on desk), although we did see higher overall risk scores in the two-unobscured pairs across the entire grid, as expected. Slicing for mobility also showed no clear effect on the discontinuities. We infer from this exploration of the data that the risk scores exhibit the combined effect of multiple factors, possibly including ones beyond the scope of our measurement campaign, such as hardware variation within each calibration confidence level, environmental interference, and multipath effects.

With respect to contact tracing efforts, we noted large regions on the P(Alert) heat maps showing a probability equal to 100%, well outside the blue box representing the strict TC4TL metric. This is a direct result of the configuration selected by the public health authority, which includes the relative weights for very close, moderately close, far, and very far encounters, and the attenuation thresholds that define those ranges, as well as the final alert threshold(s), which are quite low relative to the range of observed risk scores.

This page intentionally left blank.

6. DISCUSSION

To judge the appropriateness of the configuration, one must first declare a preference for desired behavior of the app with respect to ground truth exposures. If the app should only alert when the exposure meets the strict TC4TL metric, these data would show an unacceptably sensitive system, which would send "too many" people into testing and quarantine, although it does show an acceptable rate of capture of the "true positive" alerts. However, if one is interested in capturing encounters deemed "risky enough" by the PHA beyond that strict threshold, these data show a system that is appropriately specific to the scenarios in which the community is most at risk.

Finally, we acknowledge that the experimental constraints mandate some caution. The pairwise exposures are heavily skewed towards Android-Android and Android-iPhone pairs, due to the number of iPhones in our testbed and to Apple's decision to hide beacon information, even from users with EN developer entitlements. This Android-heavy demographic is not well matched to the general population in many regions worldwide. The distribution of low-calibration and high-calibration Androids in our testbed may not be well matched to the actual regional distribution in Arizona or elsewhere, and it should become outdated as calibration values are updated by Google. Combining our dataset with others' EN, distance, and duration measurements likely would produce a more accurate prediction of system performance.

The perspectives of our colleagues at the University of Arizona are summarized independently in Appendix C: Practice Implications.

This page intentionally left blank.

7. CONCLUSION

A thorough performance assessment of contact-tracing apps must consider not only their system-level performance as integrated with the public health and medical communities, nor only the peer-to-peer distance metric implemented on hardware, though both of those are important contributions to the effectiveness of the EN distributed sensing network. The system performance must also be measured and assessed in small- to medium-scale interactions, reflecting the realities of how people interact while using EN, including such confounding factors as mobility, interference, absorption, and multipath effects. Our data collection efforts significantly reduced this gap in system testing and analysis, by conducting a series of realistic, scripted, repeatable experiments, across a variety of phone hardware and physical configurations. Although individual exposures showed a high variability in risk scores and the probability of alert, the integrated experience of an individual over multiple encounters, or of a large body of users, has less variance and a reasonable correlation to ground truth distance and duration of exposure. Our analysis showed that the University of Arizona's configuration makes the EN risk scoring algorithm a highly sensitive tool for estimating possible COVID-19 exposure, in common scenarios where anonymous contacts are likely and EN-compatible devices are usually present.

APPENDIX A: EXPOSURE VISUALIZATION TOOL

The Exposure Visualization tool is custom software created to help interpret the results of the scenario testing. It can also be helpful in conveying the various test settings, relative phone placements, and collected data to a general audience. All nine of the testing scenarios can be visualized by the tool to better understand the reasoning for the observed alert status of each exposed phone at the end of the testing period. Figure provides an example view of the MBTA bus tests.

When a test scenario is loaded into the visualization tool, the three corresponding tests are displayed adjacently in separate panels. Each panel contains multiple components that show information for that test which can then be compared between tests. In the center of the panel is a top-down two-dimensional sketch of the space in which the test was conducted. Circular sprites representing agents are placed in their ground truth positions within the space, with a pointed edge denoting the direction they are facing. The color of the agent indicates the alert level observed by the carried phone: gray = “none”, yellow = “low”, red= “high”. Green indicates that the phone is carried by the “sick” agent.

The test ID, number of phones in the test, and information on the sick phone is displayed on the left side. The sick phone information includes the phone ID, carrying agent ID, phone model, and phone calibration confidence. Additionally, exposed agents can be selected from the scene sketch to display similar information for that agent/phone pair, as well as the distance from the sick phone and the alert level observed at the end of the test.

When an exposed agent is selected, the attenuation graph of the carried phone is populated. This graph reports the average measured attenuation of each scan that occurred during the testing period and the range of attenuation values captured within the scan. The colored regions reflect the configured bins for the “Immediate”, “Near”, “Medium”, and “Other” attenuation ranges of the EN service.

To reflect the dynamic nature of some scenarios, the test time can be stepped through by the user. The current time is shown as a red line on the attenuation graph. The positions and rotations of the agents as well as the distance information of the selected phone is updated to reflect the test state at the stepped time.

At the present time, the visualization tool only displays data for exposures to the single experimental “sick” phone for each trial. However, it would be possible to extend it to display data for all 2991 pairwise exposures in the expanded analysis. Scenario data is read in at runtime, so the tool could also be used to visualize exposure data from complementary datasets if it is provided in the correct formats.

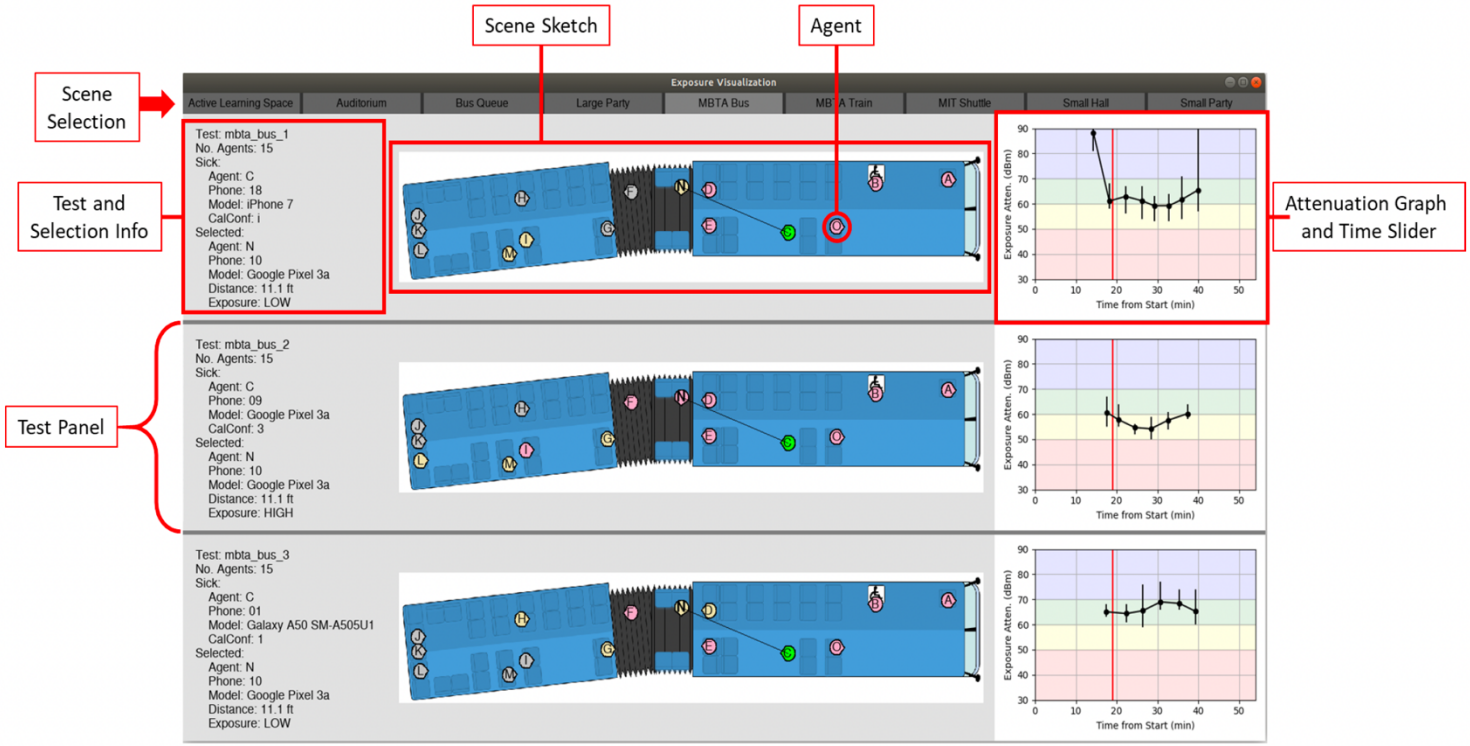
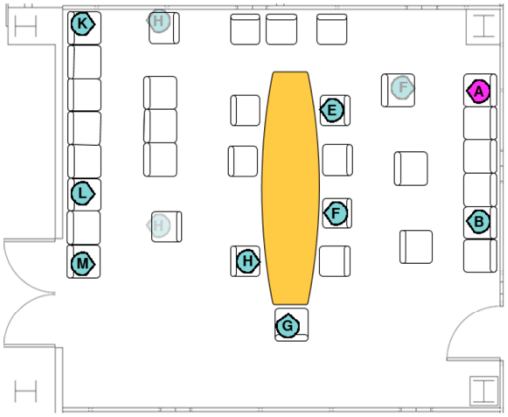


Figure 8. Example of scenario visualization across three trials.

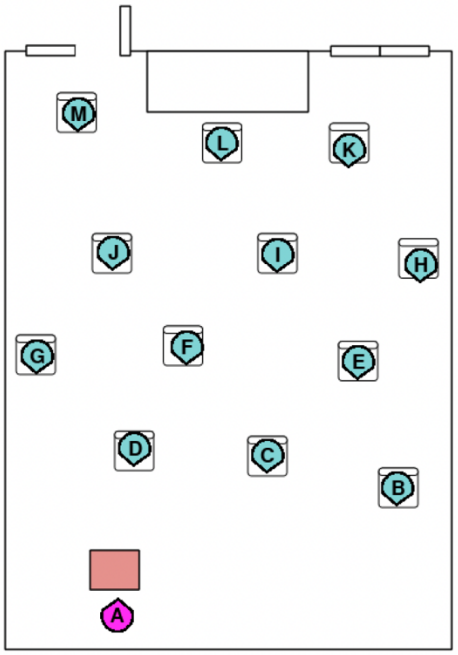
APPENDIX B: SCENARIO CONFIGURATIONS

Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	Bag (adj)	-	-	-
B	LFSP	High	High	High
E	Table	High	High	High
F	RBPP	High	High	High
G	LFPP	High	High	Low
H	RFPP	High	High	High
K	LFSP	High	High	High
L	Bag (adj)	High	High	None
M	Hand	High	High	None



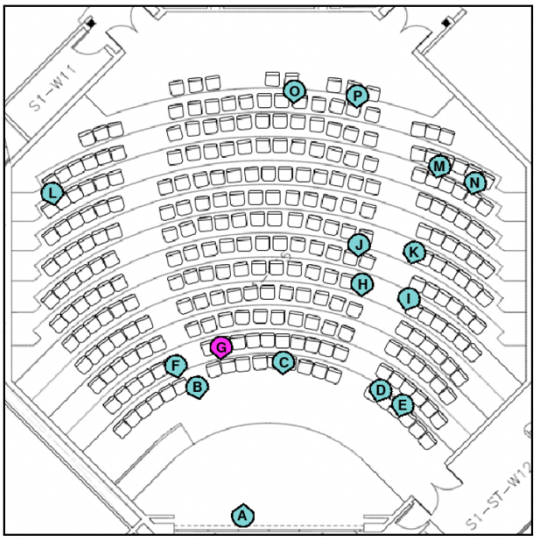
L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket; BPP = back pants pocket

Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	LIJP	-	-	-
B	RFPP	High	High	None
C	LFPP	High	High	None
D	LFSP	High	High	None
E	LFSP	High	High	None
F	LFSP	High	High	High
G	Hand	High	High	High
H	LFPP	High	None	None
I	Bag (adj)	Low	Low	None
J	LFSP	None	High	None
K	LFSP	None	High	None
L	LFSP	High	Low	None
M	Bag (adj)	None	Low	None



L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket; BPP = back pants pocket

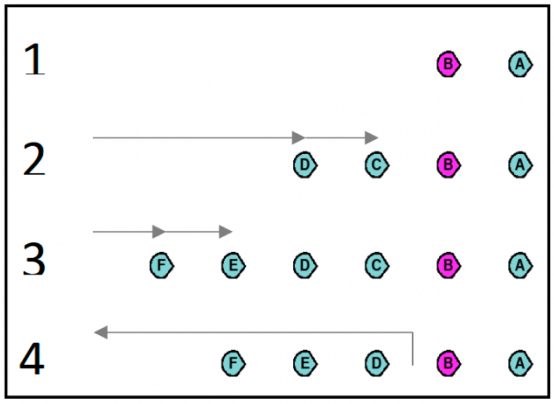
Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	LFHP	High	High	High
B	LFPP	High	High	High
C	LFSP	High	High	High
D	RFPP	High	Low	High
E	Table	High	Low	None
F	RFPP	High	Low	High
G	Table	-	-	-
H	LFSP	Low	None	High
I	Bag (adj)	High	None	High
J	Table	Low	None	High
K	Hand	Low	None	High
L	LFSP	Low	None	None
M	LFSP	High	None	Low
N	Table	High	None	High
O	LFSP	None	None	None
P	Bag (adj)	High	None	None

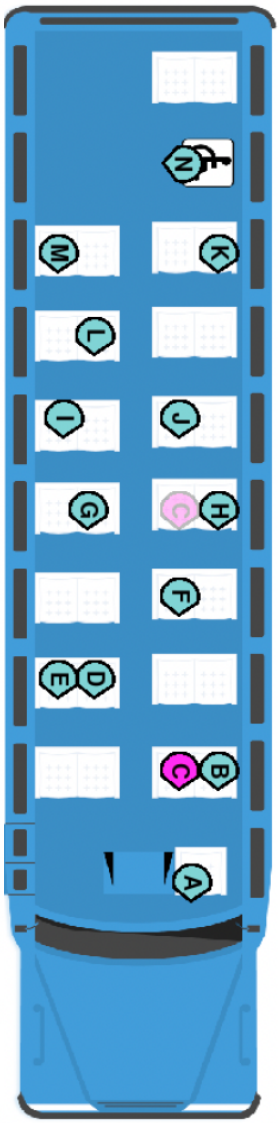


L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket; FHP = front hip pocket; BPP = back pants pocket

Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	LFSP	High	High	High
B	LFSP	-	-	-
C	LFSP	Low	Low	None
D	LFSP	High	Low	High
E	LFSP	High	Low	High
F	LFSP	High	Low	Low

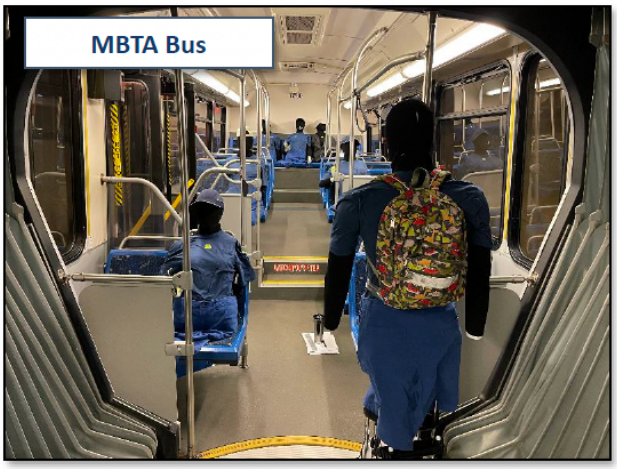
LFSP = left front shirt pocket

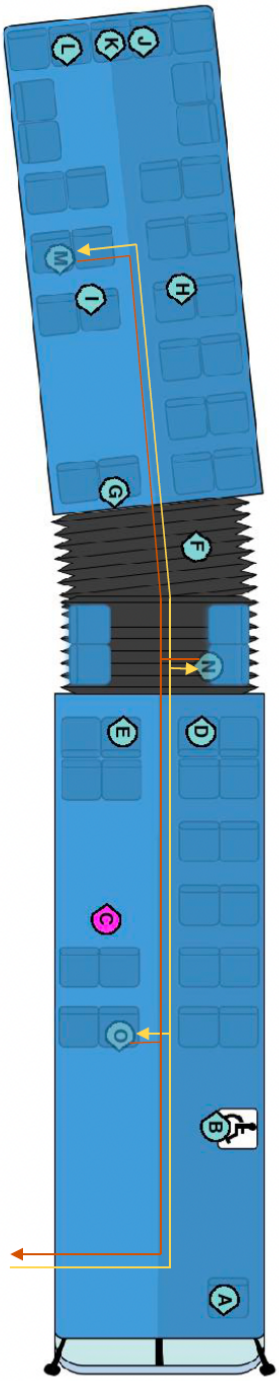




Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	LFSP	High	High	High
B	LFSP	High	High	High
C	Hand	-	-	-
D	LFHP	High	High	High
E	LFSP	High	High	High
F	Hand	High	High	High
G	LFPP	High	High	High
H	Hand	High	High	High
I	Hand	High	High	High
J	RFPP	High	High	Low
K	LFSP	High	High	High
L	RFPP	High	High	Low
M	LFPP	High	High	High
N	Bag (worn)	High	High	Low

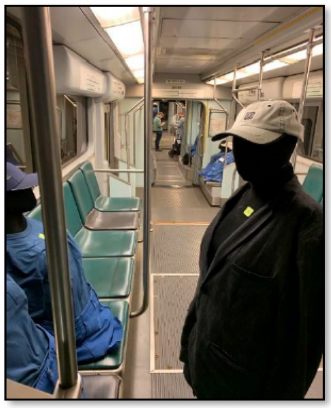
L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket;
FHP = front hip pocket; BPP = back pants pocket; HP = hoodie pocket





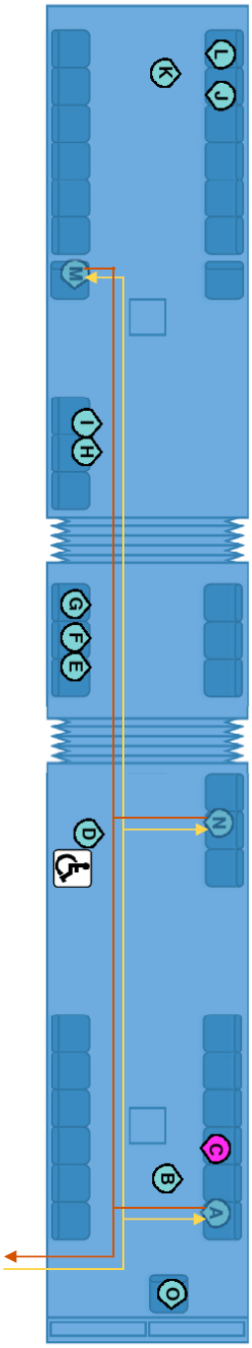
Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	RBPP	High	High	High
B	Bag (worn)	High	High	High
C	RBPP	-	-	-
D	LFSP	High	High	Low
E	Hand	High	High	High
F	Bag (worn)	None	High	High
G	Hand	None	Low	Low
H	Hand	None	None	Low
I	LFPP	Low	High	None
J	LIJP	None	None	None
K	Hand	None	None	None
L	LFSP	None	Low	None
M	Hand	Low	Low	None
N	Hand	Low	High	Low
O	Hand	High	High	High

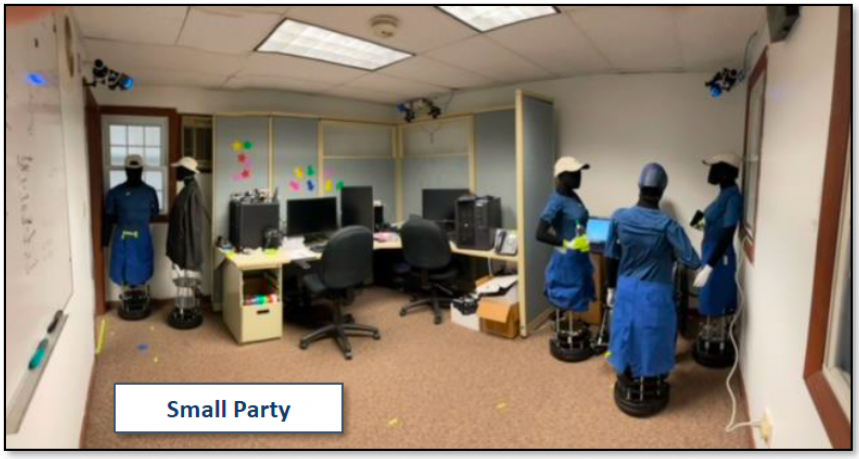
L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket;
FSP = front shirt pocket; FHP = front hip pocket;
BPP = back pants pocket; HP = hoodie pocket



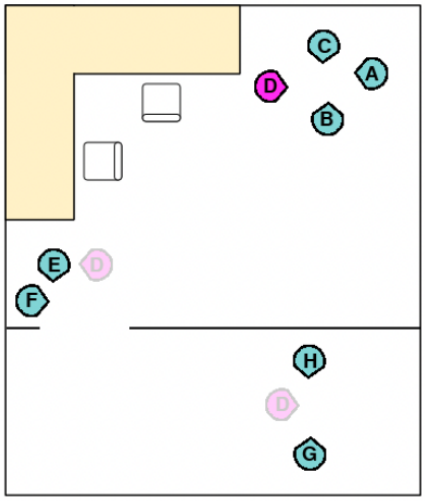
Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	HP	Low	High	Low
B	LFSP	High	High	High
C	Bag (adj)	-	-	-
D	LFSP	High	High	High
E	LFPP	None	High	None
F	Bag (adj)	High	High	Low
G	LFSP	None	High	Low
H	LFPP	None	High	High
I	LFSP	None	High	High
J	LFSP	None	Low	None
K	LIJP	None	Low	Low
L	LFPP	None	High	High
M	LFPP	None	Low	Low
N	RFPP	None	Low	Low
O	LFSP	High	High	High

L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket;
FHP = front hip pocket; BPP = back pants pocket; HP = hoodie pocket



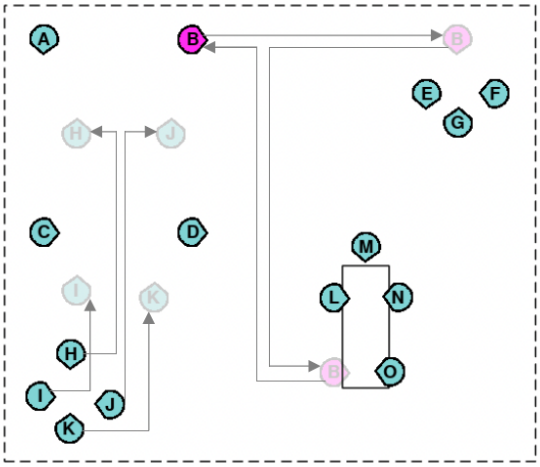


Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	RFSP	High	High	Low
B	RBPP	High	High	High
C	LFSP	High	Low	High
D	RFPP	-	-	-
E	LIJP	High	High	High
F	LFPP	High	High	High
G	Hand	High	High	High
H	RFPP	High	High	High



L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket; FHP = front hip pocket; BPP = back pants pocket; HP = hoodie pocket

Phone	Carriage	Test 1 Alert	Test 2 Alert	Test 3 Alert
A	LFSP	Low	High	High
B	LBPP	-	-	-
C	LFSP	High	High	High
D	RBPP	Low	High	High
E	Hand	Low	High	High
F	LFSP	High	High	High
G	LBPP	Low	High	High
H	Hand	Low	High	High
I	Hand	High	High	High
J	RFPP	Low	High	Low
K	LFSP	Low	Low	None
L	HP	High	High	High
M	RFPP	Low	High	High
N	RFHP	Low	High	None
O	LFSP	Low	High	Low



L* = left; R* = right; adj = adjacent to seat
IJP = interior jacket pocket; FPP = front pants pocket; FSP = front shirt pocket; FHP = front hip pocket; BPP = back pants pocket; HP = hoodie pocket

APPENDIX C: PRACTICE IMPLICATIONS

Logistics: While the University of Arizona had considered conducting their own Beta testing, there were serious safety concerns about individuals being asked to be within 6 feet of each other in different scenarios and spaces during periods of high community transmission. The MIT facility primarily used mannequins, alleviating safety concerns about the testing procedures.

Primary UAZ objectives: Results from the testing with MIT provided the University of Arizona COVID Watch team information that assisted internal and external validation of the deployed system. Prior to the testing there were concern about 1) how different model phones interacted, particularly between Android and iPhones and among older model phones which were commonly held by lower income students attending the University, 2) the ability of the technology to work in different environments such as learning environments, public transit, and social settings, 3) the ability of the technology to bin risk level of exposure appropriately. The greater concern was that there would be lower sensitivity, whereby individuals who had an exposure would be missed or that individuals who were just within the same room would be notified by even casual and temporary contact.

UAZ interpretation and utilization of results: Testing procedures indicated broadly that the exposure notification system generated higher risk scores when the duration was longer and the proximity was closer, but this appears to vary by multiple factors including model of the phone, positioning and interference. These limitations are important to define to temper expectations on the performance and limitations of the system. It demonstrated that while EN notification is an excellent complement to the current on-campus contact tracing program, they need to be deployed in concert with each other. The variability identified was very important to the interpretation of effectiveness. We communicated the results of the testing to partners within the state including Arizona Department of Health Services and county health partners.

Limitations of testing scenarios: The testing scenarios implemented were excellent at developing an understanding of system performance within situations where there were low density interactions that were modeled off common physical distancing protocols. However, there is some remaining unknown around how these systems perform in highly dense crowded environments that we are currently facing now that the COVID-19 prevention protocols have essentially been eliminated in most circumstances (as of April 2022). The performance with interference of signals between people that are close to each other but blocked by another body are unknown. Classroom and workplace settings no longer implement distancing making it challenging to understand how the results of the previous work might inform the current performance. Additional testing under these new operating conditions could be quite useful.

Kacey C. Ernst, MPH, PhD
Professor and Program Director of Epidemiology
Department of Epidemiology and Biostatistics
College of Public Health

GLOSSARY

ASDF	Autonomous Systems Development Facility
Attenuation	Reduction of signal amplitude
CDC	Centers for Disease Control and Prevention (United States)
COVID-19	Coronavirus disease caused by the SARS-CoV-2 virus
CovidWatch	University of Arizona's Exposure Notification-based mobile app
EN	Exposure Notification
GAEN	Google-Apple Exposure Notification
ID	Identifier, implicitly unique
JSON	JavaScript Object Notation
MBTA	Massachusetts Bay Transportation Authority
MIT	Massachusetts Institute of Technology
MIT LL	Massachusetts Institute of Technology Lincoln Laboratory
P(Alert)	Probability of EN alert occurring
P(D)	Probability of detection
P(FA)	Probability of false alarm
PACT	Private Automated Contact Tracing
RF	Radio frequency
RPI	Rolling Proximity Identifier, a short-lived token generated from the TEK
RSSI	Received Signal Strength Indicator
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
TC4TL	"Too close for too long" standard definition

TEK	Temporary Exposure Key, a cryptographic token generated on the smartphone once per day
UAZ	University of Arizona

REFERENCES

[1] Apple, "Configuring Exposure Notifications," [Online]. Available: https://developer.apple.com/documentation/exposurenotification/configuring_exposure_notifications. [Accessed 10 04 2022].

[2] Google, "Define Meaningful Exposures," [Online]. Available: https://developers.google.com/android/exposure-notifications/meaningful-exposures. [Accessed 10 04 2022].

[3] M. C. Schiefelbein, S. Mazzola, R. C. Gervin Jr. and J. S. Germain, "Bluetooth Low Energy (BLE) Data Collection for COVID-19 Exposure Notfication," MIT Lincoln Laboratory, Apr. 2022.

[4] M. Mace, "How a Smartphone App and Contact Tracing Helped Keep UArizona Open and Curb COVID-19 Spread," UANews, 16 12 2020. [Online]. Available: https://news.arizona.edu/story/how-smartphone-app-and-contact-tracing-helped-keep-uarizona-open-and-curb-covid-19-spread. [Accessed 10 04 2022].

[5] Google, "Exposure Notifications BLE attenuations," [Online]. Available: https://developers.google.com/android/exposure-notifications/ble-attenuation-overview. [Accessed 10 April 2022].

[6] J. Masel, A. Shilen, B. Helming, J. Rutschman, G. Windham, K. Pogreba-Brown and K. Ernst, "Quantifying meaningful adoption of a SARS-CoV-2 exposure notification app on the campus of the University of Arizona," *medRxiv,* 2020.

[7] MIT Lincoln Laboratory, "Autonomous Systems Development Facility," [Online]. Available: https://www.ll.mit.edu/about/facilities/autonomous-systems-development-facility. [Accessed 10 04 2022].

[8] A. M. Wilson, N. Aviles, J. I. Petrie, P. I. Beamer, Z. Szabo, M. Xie, J. McIllece and Y. Ce, "Quantifying SARS-CoV-2 infection risk within the Google/Apple exposure notification framework to inform quarantine recommendations," *medRxiv,* 16 April 2021.

[9] MIT Lincoln Laboratory, "GitHub - mitll/PACT-Exposure-Notification-Beacons," 27 09 2021. [Online]. Available: https://github.com/mitll/PACT-Exposure-Notification-Beacons.

[10] Google, "GitHub - google/exposure-notifications-internals," [Online]. Available: https://github.com/google/exposure-notifications-internals . [Accessed 10 04 2022].

[11] MIT Lincoln Laboratory, "GitHub - mitll/Exposure-Visualization-Tool," 05 01 2022. [Online]. Available: https://github.com/mitll/Exposure-Visualization-Tool.

[12] CDC, "Contact Tracing : Part of a Multipronged Approach to Fight the COVID-19 Pandemic," [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/php/principles-contact-tracing.html. [Accessed 30 April 2020].

This page intentionally left blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
15/04/2022	Project Report	

4. TITLE AND SUBTITLE

COVID-19 Exposure Notification in Simulated Real-World Environments

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Curran Schiefelbein, Steven Mazzola, Richard Gervin, Joe St. Germain

5d. PROJECT NUMBER
10383-2

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

ACTA-3

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Centers for Disease Control and Prevention

1600 Clifton Road, Atlanta, GA 30329 USA

10. SPONSOR/MONITOR'S ACRONYM(S)
CDC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.
This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Privacy-preserving contact tracing mobile applications, such as those that use the Google-Apple Exposure Notification (GAEN) service, have the potential to limit the spread of COVID-19 in communities, but the privacy-preserving aspects of the protocol make it difficult to assess the performance of the apps in real-world populations. To address this gap, we exercised the CovidWatch app on both Android and iOS phones in a variety of scripted real-world scenarios, relevant to the lives of university students and employees. We collected exposure data from the app and from the lower-level Android service, and compared it to the phones' actual distances and durations of exposure, to assess the sensitivity and specificity of the GAEN service configuration as of February 2021. Based on the app's reported ExposureWindows and alerting thresholds for Low and High alerts, our assessment is that the chosen configuration is highly sensitive under a range of realistic scenarios and conditions. With this configuration, the app is likely to capture many long-duration encounters, even at distances greater than six feet, which may be desirable under conditions with increased risk of airborne transmission.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		49	**19b. TELEPHONE NUMBER** *(include area code)*