Duke University Snowball Respondent-Driven Sampling Study

SNOWBALL STUDY TOOLKIT

Respondent-Driven Sampling for Respiratory Disease Surveillance

Table of Contents

Introduction to the Snowball Toolkit

The Snowball Toolkit consists of four components: (1) the Snowball Technology Platform that was built to support the study; (2) the Standard Operating Procedures for safe and effective sample collection for SARS-CoV-2; (3) the social contact models utilized by the study; and (4) the predictive model developed with biometric data.

The Snowball Toolkit was designed to be sufficiently general to be applied across different settings. Although the details described are specific to community transmission of SARS-CoV-2, each individual piece can be modified for application to different infections. For instance, the Snowball Platform has a customizable survey to collect information relevant to the infection (or intervention) of interest. The sampling SOP can serve as a guide for developing protocols for sampling for infections with different modes of transmission or biosafety concerns. The social contact models can be tailored to the specific route(s) of transmission and population(s) at risk and the biometric monitoring algorithm can be modified based on the prodrome stage and incubation time of any infection.

Each component can stand alone or be used in conjunction with any or all other Snowball Toolkit pieces as needed. The Snowball study provides a use case for the deployment of these tools, but they are designed to have utility for future applications in different research and public health settings and in response to a broad range of infectious diseases.

For questions about these materials or interest in utilizing any of the Toolkit components, please feel free to contact Snowball Study Principal Investigator, Dr. Dana Pasquale, at dana.pasquale@duke.edu.

Snowball Technology Platform

The Snowball technology platform served as the foundation of our study activities, including outreach and enrollment, study data management, and reporting. Over the first year of our project period, we developed two versions of the platform, which we describe below.

Snowball Platform v1.0

Snowball platform v1.0 was developed specifically for the Snowball study and launched during the first quarter of the project period. It was based on the Duke Clinical Research Model and designed to meet all of the requirements for safely conducting human subjects research at Duke. It was also built to integrate with the Duke Health electronic health record for seed case ascertainment and with the Research Electronic Data Capture (REDCap)[1,2] application hosted at Duke for electronic consent and for the participant survey.

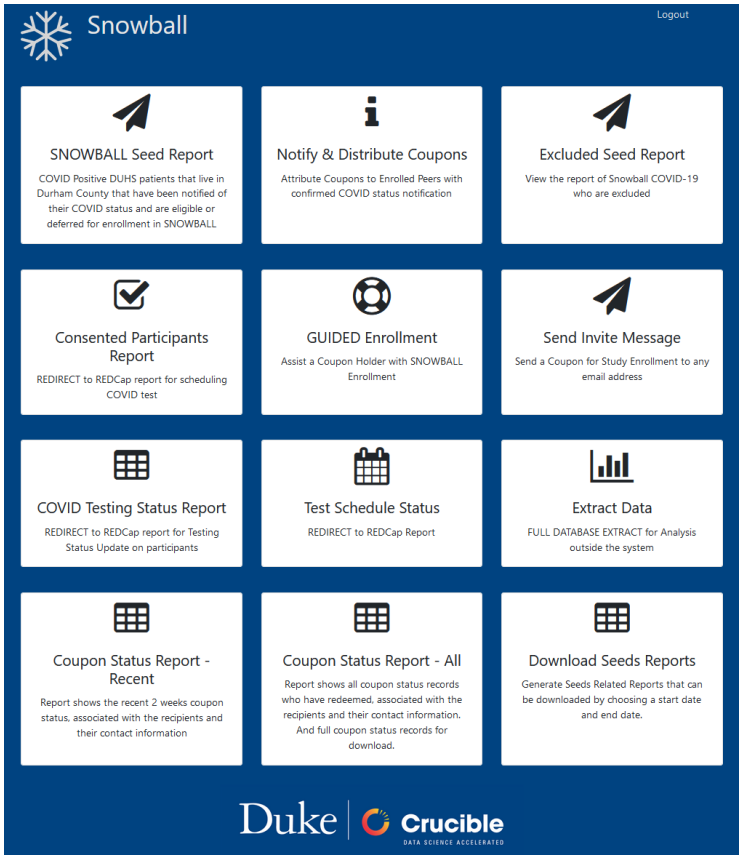

Figure 1. Screenshot from the Snowball platform v1.0 homepage

Through the platform, we received a daily seed report, which included all eligible persons who had recently tested positive within the Duke University Health System (Duke Health). The study's principal investigator (PI) reviewed all prospective seed cases and determined whether to include, exclude, or defer them from the study. Persons who were included were then sent a system-generated email inviting them to participate in the study and providing them with a unique 4-word coupon code that they

[1] PA Harris, R Taylor, R Thielke, J Payne, N Gonzalez, JG. Conde, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009 Apr;42(2):377-81.

[2] PA Harris, R Taylor, BL Minor, V Elliott, M Fernandez, L O'Neal, L McLeod, G Delacqua, F Delacqua, J Kirby, SN Duda, REDCap Consortium, The REDCap consortium: Building an international community of software partners, J Biomed Inform. 2019 May 9 [doi: 10.1016/j.jbi.2019.103208]

could use to enroll. Participants then visited the study website to learn more about the study and elect to enroll, at which point they were forwarded to REDCap to validate their coupon, review and sign the study's electronic informed consent form, and complete the social network survey.

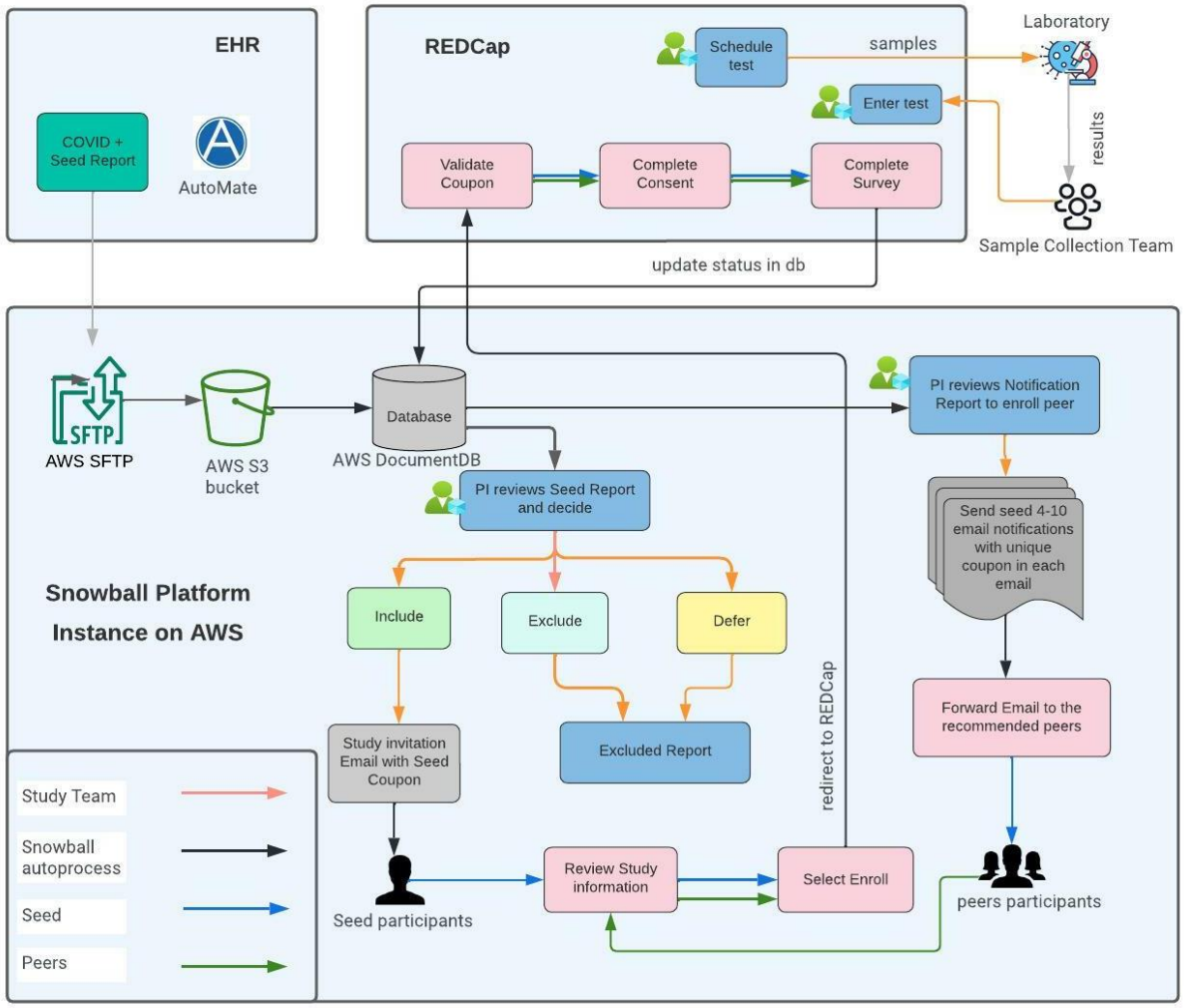

Figure 2. Snowball Platform v1.0 Participant Enrollment Process.

As participants moved through the enrollment process, the platform synchronized with REDCap to pull the latest study milestone for each participant. Once a seed participant completed their survey, they were added to a report in the platform so that the PI could review their survey results and release, via the platform, automated emails with 4-10 coupons for the seed case to distribute to their peers to recruit them to the study. Prospective peer cases that receive a coupon would then go through the same enrollment process as the seed cases, complete the REDCap survey, and be contacted by the sampling team for COVID-19 testing.

Critically, the platform generates all of the unique 4-word codes shared for enrollment and uses them to track all links between seed and peer cases as the chains of enrollment develop. Study investigators use this data to conduct the social network analyses.

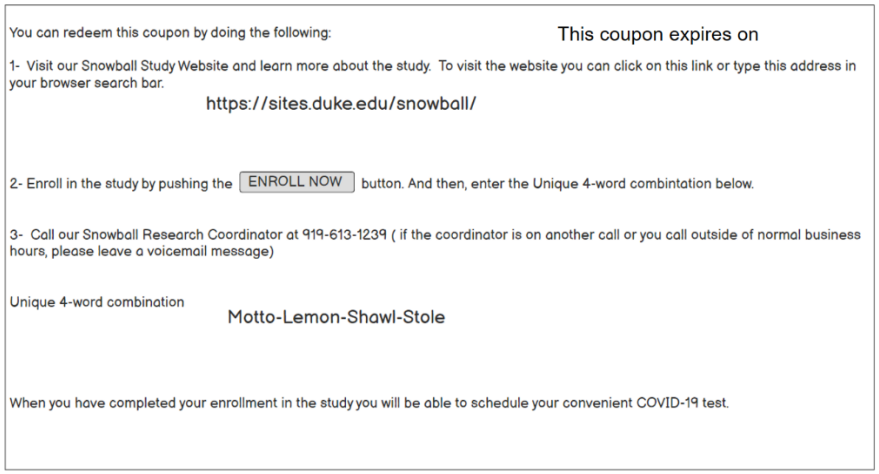

Figure 3. Example email invitation to participate in Snowball with 4-word coupon code

Finally, the platform also included a number of reporting functions that generated tailored reports used to share information with the study's investigators, sponsor, and institutional review board (IRB), as well as for analysis by the study team.

General Release Version

In addition to platform v1.0 deployed by the study team, a key deliverable of the Snowball project was to develop an open-source, General Release version of the platform that could be made publicly available. To achieve this objective, we used what we learned from the experience of developing and implementing v1.0 to inform the design of a more flexible version that could support respondent-driven sampling methods in response to COVID-19 or other infectious disease outbreaks.

While v1.0 runs on a virtual private cloud with Amazon Web Services as its cloud provider, the General Release version is designed to be "cloud-agnostic." To demonstrate this feature, the demo version of the General Release accessible via the Duke Crucible website runs on a Microsoft Azure cloud subscription.

The General Release version is also designed with the capability to integrate with an organization's existing electronic health record similar to how v1.0 was deployed at Duke, but does not require integration for seed ascertainment. Instead, the General Release is designed with functions for uploading seed reports as CSV files or for directly adding individual new seed cases. The General Release version also enables but does not require integration with REDCap for consent and completion of the social network survey. Instead, the General Release version includes a feature to upload participant consent forms directly to the platform and also a fully customizable survey tool.

Figure 4. Screenshot from the Snowball Platform page on the Duke Crucible website.

With these features and design elements in place, we hope that the General Release version will be downloaded and deployed by a broad range of organizations including other research institutions, public health agencies, or health systems that are interested in using the tool as part of their contact tracing efforts in response to COVID-19 or other infectious diseases.

We completed our development of the General Release by the end of the first year of the study period and made all source code and documentation publicly available via two GitHub repositories. Both repositories include README files with quick start instructions and information regarding the development and deployment (including the environment configuration) of the interfaces. They can be accessed via the following links:

- The Snowball General Release user interface (UI) can be found here: https://github.com/duke-crucible/Snowball-UI.
- The Snowball General Release application programming interface (API) can be found here: https://github.com/duke-crucible/Snowball-API.

Links to the repositories, along with a publicly available demo instance of the platform, are available via the Duke Crucible website: https://crucible.duke.edu/products/snowball/.

Back to top

Standard Operating Procedures

SNOWBALL Toolkit | **Sampling Best Practices: Standard Operating Procedure (SOP)**

Revision Date: 02-December-2022 **SOP Reference #** 0001

Description: The process below provides the step-by-step process for sampling community members who may be positive for SARS-CoV-2 or another aerosol-spread respiratory infection

Frequency: **See details below to determine frequency**

Timing: **See details below to determine timing**

The minimal sample for study participation is 1 SST tube and an NP swab. If a potential subject refuses the nasopharyngeal swab but has already had a sample collected for routine care, the clinical research coordinator (CRC) may still enroll that subject ONLY if they are able to obtain the residual NP swab sample from the clinical lab (follow site procedures to obtain this sample) within 24 hours. The subject may NOT be enrolled without collecting an NP swab sample (clinical or study related) for study use.

If the NP swab sample is obtained from the residual clinical sample, at least 600µL must be available to be split into two aliquots of ~300µL.

Specimen Collection and Management:

Specimen Collection:

Blood Collection

Supplies needed for Blood Specimen Collection

Evacuated Tube Holders	Alcohol Wipes
Evacuated tubes: PAXgene Blood RNA Vacutainer Tube	Gauze Pads
Serum Tubes	Bandages or Tape
4mL EDTA Tubes (for plasma, whole blood)	Tourniquets (Check for Latex Allergies); No Latex products used in Pediatrics
Needles	Wet ice
Butterflies	Luer Adapters

- Please refer to Appendix SOP-A, Procedure for Collection of Blood Specimens by Venipuncture.
- To prevent contamination of tubes during blood draw, the serum SST (Gold/Tiger top) should be collected first and the plasma with EDTA (Lavender/purple top) last.

    - There may be special situations where blood collection must be prioritized and this order may not be followed (see **Note* below).
- PAXgene Blood RNA tubes must be immediately inverted 10X and stored according to instructions. See Appendix SOP-B: Procedure for Storage of PAXgene Blood RNA Tubes.
- EDTA vacutainers should be placed on wet ice immediately after collection. All cryovials for plasma collection should be kept on ice during aliquoting and labeling until transferred to -80°C.
- SST vacutainers should be left at room temperature for 30 min prior to centrifugation to allow clotting to occur.
- Both EDTA and SST vacutainers should be spun down no more than 4 hours after collection.
- Store samples on wet ice or in refrigerator if they are not being processed immediately.

Nasopharyngeal (NP) Swab and Throat/Oropharyngeal (OP) Swab Collection

Supplies needed for Nasopharyngeal (NP) Swab and Throat/Oropharyngeal (OP) Swab Collection

Nasopharyngeal/Throat Swabs	Specimen rack
Wet ice and container	Flocked Minitip Swab with 3 mL Universal Transport Media
Throat (polyester) Swab	Tongue Depressor
N95 respirator and gloves	Goggles
Gown	Heavy duty scissors

- Please refer to Appendix SOP-C: Procedure for Collection of Nasopharyngeal Swab and Oropharyngeal Swab.
- If both NP and OP swabs are collected, place tip of the OP swab into the same tube as the NP swab and cut off the applicator tip using heavy duty scissors. NP/OP swabs will be processed into aliquots together according to Appendix SOP-C: Procedure for Collection of Nasopharyngeal Swab.

Rapid Antigen Tests

Commercially available antigen-tests should be used, when available, to provide peers with preliminary information about the status of their COVID-19 diagnosis. Follow the instructions of the commercially available test for instructions for use. Results from the test should be recorded on the visit's data collection form.

Specimen Processing

Supplies Needed for Specimen Processing:

- Freezer boxes 9x9 – each holds 81 samples (for 2mL cryovials)
- Freezer boxes 7x7- each holds 49 samples (for 5mL cryovials and tubes)
- 1mL pipette and sterile, filtered tips or sterile pipette
- Sample acquisition forms (SAF) for biorepository (contains visit date, subject identifier, selection of samples collected, date and time of each collection, and aliquot identifiers)
- 2mL cryovials
- 5 mL cryovials
- Barcode labels with unique identifiers, in duplicate
- Wet ice
- Dry ice

To process blood samples:

- Process the PAXgene Blood RNA tubes for storage according to Appendix SOP-B: Procedure for Storage of PAXgene Blood RNA Tubes. Make sure to label tubes.
- Process the serum according to Appendix SOP-D: Procedure to Aliquot Serum from Whole Blood. Make four approximately 0.5mL aliquots. Label the SAF and aliquots and freeze at -80°C in the 9x9 freezer box.
- Process the plasma according to Appendix SOP-E: Procedure to Aliquot Plasma from Whole Blood. Make four approximately 0.5mL aliquots. Label SAF and aliquots and freeze at -80°C in the 9x9 freezer box.
- Plasma and serum samples may be centrifuged together as they have the same processing requirements (1300 RCF in a 4°C centrifuge for 10 minutes). Label and freeze whole blood (PAXgene and EDTA) at -80°C in the 7x7 freezer box.

To process nasopharyngeal swab/oropharyngeal (throat) swab:

- If both a nasopharyngeal and throat swab are present, process them together. On the SAF, there is an entry only for a Nasopharyngeal Swab. If a Throat Swab is also taken, the "NP/OP" box in the Nasopharyngeal Swab section must be checked to specify that this specimen contains biological information from both anatomical sites.
- Process the swab according to Appendix SOP-F: Procedure to Process and Aliquot Nasopharyngeal Swab/Oropharyngeal Swab Fluids. Make five approximately 0.5mL aliquots.
- Label the SAF and aliquots and freeze at -80°C in the 9x9 freezer box.
- Freeze the five aliquots in the 9x9 freezer box.

Specimen Management and Shipping	After the samples have been processed and/or aliquoted, transfer all samples to a deep freeze (80°C) for storage until ready for shipping or transfer into the specimen biobank. All SAF's (and photocopies) associated with the specimen collections should be kept in locked filing cabinets until ready for shipment/transfer for data entry to specimen biobank. The site study coordinator will make necessary arrangements with the specimen biobank coordinator to transfer samples to the specimen biobank. **Shipping Instructions for Frozen Specimens:** • Contact recipient prior to shipping to ensure your sample will not be arriving on a day on which staff will not be present to receive samples. If possible, sites should plan to ship early in the week (Mon-Wed) in case of shipping delays. A sample manifest should be provided for each shipment container of what is expected in the shipment. The manifest must be submitted electronically by email to the Biobank Coordinator and a hard copy included in each shipping container. • Samples will be shipped on dry ice and placed in adequate shipping containers with all necessary labeling and packaging requirements as defined by federal, industry, and U.S.P.S. authorities. • All samples shipped for the purposes of this study should be considered Infectious Substance, Category B. The following links provide FedEx and IATA guidelines for shipping Category B samples: http://www.fedex.com/us/services/pdf/How_To_Pack.pdf (refer to page 32 of the linked PDF for UN3373 information) https://www.iata.org/whatwedo/cargo/dgr/Documents/packing-instruction-650-DGR56-en.pdf

Back to top

Appendix SOP-A: Procedure for Collection of Blood Specimens by Venipuncture

Purpose

The purpose of the standard operating procedure (SOP) is to provide a consistent procedure for the collection of blood by venipuncture following standard protocols. Samples should be considered biohazardous and handled appropriately.

Background

Failure to follow correct procedure results in errors in the collection process. This may cause interference in the analysis of laboratory specimens and impact laboratory results. Collection errors include incorrect patient identification, failure to mix specimens correctly, hemolyzed or clotted specimens, incorrect order of draw, and the use of an incorrect anticoagulant during specimen collection.

Materials/Reagents

- Needles
- Butterflies
- Luer Adapters
- Evacuated Tube Holders
- Evacuated tubes
- Alcohol Wipes
- Gauze Pads
- Bandages or Tape
- Tourniquets (Check for Latex Allergies) No Latex products used in Pediatrics
- Refrigerant or Hot Packs if indicated

Procedure

1. Verify correct specimens for selected patient.
2. Wash hands and don clean pair of gloves.
3. Greet patient (and family), identify self and explain purpose of encounter.
4. Properly identify patient using two (2) patient identifiers (call patient name and verify by study documents).
5. Assemble and properly position equipment.
6. Properly apply tourniquet 3-4 inches above the intended venipuncture site. Tourniquet should not remain on patient for more than one (1) minute.
7. If only available site is an arm with an IV, the IV must be turned off for two (2) minutes. Draw blood below the IV site (only nurse may do this).
8. If blood products are being administered, venipuncture should be performed in the opposite arm. If no site is available, then following the completion of blood product administration, collect sample.
9. Select a vein for the venipuncture. Palpate with your index finger to determine the size, depth and direction of the vein. A vein that is large and well anchored (does

not move to side or roll easily) is usually the best choice. Make sure that the equipment you have chosen is appropriate for the size of the vein selected.
10. Clean site using appropriate cleansing agent using a circular motion from the center to the periphery. *Do not touch the site after it has been cleaned.*
11. Allow site to air dry.
12. Perform the venipuncture. Line the needle up with the vein, with the **bevel** of the needle facing **upward.**
13. Allow evacuated tubes to fill completely or at least to the minimum fill line of the tube.
14. Collect specimens according to appropriate order of draw, mixing specimens gently and thoroughly as required. Correct order of draw will be study specific and is designed to minimize carryover from tube additives. For RNA tubes (PAXgene RNA): it is essential to invert this tube 10 times immediately after collection. Do not shake!
15. Release tourniquet and remove needle from vein.
16. Apply pressure to venipuncture site.
17. Apply pressure to bandage. For pediatrics, follow age specific guidelines for type of bandage used.
18. All items are single use items and should be disposed of according to institutional policy.
19. Apply labels to appropriate tubes (PAXgenes and whole blood EDTA) orienting labels in the correct position in the presence of the patient.
20. Remove gloves and cleanse hands.

Note: If you are not able to obtain blood after two attempts, ask another Phlebotomist/nurse/tech to try. After two people have unsuccessfully attempted to draw a subject's blood, samples will not be taken.

Back to top

Appendix SOP-B: Procedure for Storage of PAXgene RNA Tubes

Purpose

The purpose of the standard operating procedure (SOP) is to describe the procedure for handling and storage of whole blood sample collected in PAXgene RNA tube for the purpose of stabilizing total RNA from peripheral whole blood (WB).

Background

The PAXgene™ Blood RNA System is used to collect a human whole blood sample for isolation of RNA (ribonucleic acid). RNA is a molecule found in cells that translates genetic information from DNA to proteins produced by the cell. RNA can be used in lab tests to evaluate gene expression levels that may facilitate diagnosis of disease or disease condition.

RNA is sensitive to multiple extreme temperature changes and is unstable in the presence of RNases that may be present in whole blood samples. To mitigate against RNA degradation during collection of whole blood, specific collection procedures must be maintained as described. In particular, it is critical to invert tubes immediately following whole blood collection to inactivate potential RNases present and to chill the sample on wet ice. Expected WB volume is 2.5 mL per tube.

Materials/Reagents

- Test tube rack

Procedure

1. Immediately after blood has been collected in the PAXgene RNA tube, invert the tube deliberately 10 times. Do not shake the tube.
2. Once the blood has been collected in the PAXgene RNA tube, label each tube with a 2-D barcode according to the procedure below, referred to as, "**Procedure for Labeling Samples for the Biobank with 2-Dimensional Barcodes**."
    a. Do not place the bar code label over a paper label, as the paper label may detach and fall off under extreme cold storage conditions.
    b. Note that when filling out the Sample Acquisition Form, the "processing time" should note when the PAX tubes were placed into the freezer.
3. Following collection and prior to freezing, PAXgene™ tubes must be set out for a minimum of 2 hours at ambient temperature (18°C – 25°C). Holding at ambient for 8 hours to overnight may work best with workflow, but less than 24 hours is best practice.
    a. The PAXgene™ circular indicates tubes may be held at 18°C – 25°C for up to 72 hours before freezing, but less than 24 hours is preferred. In cases where storage at room temperature must approach 72 hours, this is acceptable but should not be regular practice.

4. Controlled freezing is recommended by the product users manual to minimize breakage. It is best practice to the place the PAXgene™ tube upright in an open wire rack or similar at -20°C overnight. **Do not use Styrofoam trays for freezing.**
5. After ~12-24 hours at -20°C, transfer to long-term storage at -80°C. These storage conditions are sufficient to maintain total RNA integrity for future analysis.
6. If a -20°C freezer is not available, samples may be placed directly in a -80°C. The risk of doing this is that the tubes may break - to reduce this risk, store the tubes upright in a covered storage box prior to freezing.

Note: If one-step freezing is used, please check the tubes after freezing for cracks.

Back to top

Appendix SOP-C: Procedure for Collection of Nasopharyngeal Swab and Oropharyngeal Swab

Purpose

To obtain a nasopharyngeal swab or oropharyngeal swab using an efficient, consistent method.

Background

A Nasopharyngeal (NP) Swab and associated Universal Transport Medium (UTM) can contain detectable levels of virus and/or virus particles. In addition, nasopharyngeal swab/UTM specimen can contain detectable levels of proteins and other biomarkers used in various diagnostic and molecular techniques; therefore, proper handling and storage is critical. Oropharyngeal (OP) swab performed at the same time and included in the UTM can enhance detection of bacterial and viral pathogens associated with ARI. Expected yield is 3 mL.

Materials/Reagents for NP Swab

- Specimen rack
- Wet ice and container
- Flocked Minitip Swab with 3 mL Universal Transport Media. *Cotton or calcium alginate swabs are not acceptable.* PCR assays may be inhibited by residues present in these materials.
- N95 respirator and gloves
- Goggles
- Gown

Procedure for NP Swab

Important considerations:

- If exhaled breath condensate (EBC) is to be obtained at the same time as nasal fluids, they should always be collected prior to the nasal fluid collection.
- Follow recommended infection control (IC) precautions including putting on N95 respirator, goggles, gown and gloves before proceeding.

1. Label **one** UTM tube with the subject study ID, date/time of collection, and by whom the sample was collected.
2. If possible, have patient sit with head against a wall or the bed (as patients have a tendency to pull away during this procedure).
3. Insert swab into one nostril straight back (not upwards) and continue along the floor of the nasal passage for several centimeters until reaching the nasopharynx (resistance will be met).
    a. The distance from the nose to the ear gives an estimate of the distance the swab should be inserted.

   b. Do not force swab, if obstruction is encountered before reaching the nasopharynx, remove swab and try the other side.
4. Rotate the swab gently for 2-3 seconds to loosen the epithelial cells.
5. Remove swab and immediately inoculate universal transport media by inserting the swab at least 1⁄2 inch below the surface of the media. Bend or clip the swab handle to fit the transport medium tube and reattach the cap securely.
6. Immediately place the tube upright on **wet ice** until delivered to the laboratory.
7. As soon as possible, prepare the sample for longer-term storage according to the SOG **'Procedure to Aliquot and Process Nasopharyngeal Swab/Oropharyngeal Swab Fluid"**, which should be completed within **5 hours** of collection while maintaining the samples on wet ice.

Notes: Avoid repeated freeze-thaw cycles to preserve the integrity of the sample. Handling the specimens from collection to freezing should be done as quickly as possible, but delays of a few hours will probably not have an adverse effect on virus isolation rates so long as the specimens are kept cold in the interval; however, timing is potentially more critical for detection of biomarkers.

Materials/Reagents for OP Swab

- Specimen rack
- Wet ice and container
- Polyester swab. *Cotton or calcium alginate swabs are not acceptable.* PCR assays may be inhibited by residues present in these materials.
- Tongue Depressor
- N95 respirator and gloves
- Goggles
- Gown

Procedure for OP Swab

1. Ask patient to tilt head back and say "aaahhh," using tongue depressor to gently hold first 1/3 of the tongue down.
2. Insert swab into the back of the throat, avoiding tongue, uvula, teeth, lips and gums.
3. Swab the both tonsils and the back of the throat in all four quadrants gently. It is not uncommon for patients to experience a gagging sensation. Use of the tongue depressor and working efficiently during this portion will minimize this.
4. Remove swab and immediately inoculate universal transport media by inserting the swab at least 1⁄2 inch below the surface of the media. Bend or clip the swab handle to fit the transport medium tube and reattach the cap securely.
5. Immediately place the tube upright on **wet ice** until delivered to the laboratory.
6. As soon as possible, prepare the sample for longer-term storage according to the SOP-F: Procedure to Aliquot and Process Nasopharyngeal Swab/Oropharyngeal

Swab Fluid, which should be completed within **5 hours** of collection while maintaining the samples on wet ice.

Back to top

Appendix SOP-D: Procedure to Aliquot Serum from Whole Blood

Purpose

The purpose of the working guideline is to describe the appropriate procedure for collecting blood of sufficient quantity to obtain high quality serum for the purposes of proteomics analysis.

Background

Expected yield is dependent on the size of the collection tube (collection tubes come in a variety of sizes and yield will vary accordingly).

Materials/Reagents

- 1.0 mL pipette and sterile, filtered tips or sterile pipette
- 2.0 mL Cryovials

Procedure

The blood should have had sufficient time to clot since collection before the aliquot procedure can proceed.

Clotting

1. Recommended times are based upon an intact clotting process. Patients with abnormal clotting due to disease, or those receiving anticoagulant therapy require more time for complete clot formation.
2. In addition, different types of tubes have different required lag times between collection and processing. Please follow the package insert for the tube that you using to collect the serum. For SST (BD) tubes, minimum clotting time is 30 minutes at room temperature.
3. Once the blood sample has clotted, place into refrigerator or onto wet ice until processed.

Sample Processing and Centrifugation

Note: Read precautions below before proceeding attempting centrifugation.

1. Upon processing sample, proceed to centrifuge sample as indicated below.
2. Samples should be processed as quickly as possible, and within 5 hours of collection. Please indicate the time of processing on the specimen acquisition form (SAF).
3. Centrifuge the clotted samples in a refrigerated (4° C) centrifuge at 1300 RCF for 10 minutes.
4. Aliquot 0.5 mL of serum sample into each of the 2.0 mL cryovials until you exhaust the sample. **Partial aliquots are not acceptable.**
5. Label 2.0 mL cryovials according to the procedure described in Appendix SOP-B titled **'Procedure for Labeling Samples for the Biobank with 2-Dimensional Barcodes.'**
6. Immediately transfer samples to a deep freeze (-80°C) for storage.

Precautions

- Do not centrifuge glass tubes at forces above 2,200 RCF in a horizontal head (swinging bucket) centrifuge as breakage may occur. Glass tubes may break if centrifuged above 1,300 RCF in fixed angle centrifuge heads.
- Balance tubes to minimize the chance of glass breakage. Match tubes to tubes of the same fill level. Note: At this g force, Vacutainer tubes are less likely to break in a swinging-bucket centrifuge than in a fixed-angle rotor.
- Ensure that tubes are properly seated in the centrifuge carrier. Incomplete seating could result in separation of the BD Hemogard™ Closures from the tube or extension of the tube above the carrier. Tubes extending above the carrier could catch on centrifuge head, resulting in breakage.
- Always allow centrifuge to come to a complete stop before attempting to remove tubes. When centrifuge head has stopped, open the lid and examine for possible broken tubes. If breakage is indicated, use mechanical device such as forceps or hemostat to remove tubes. Caution: Do not remove broken tubes by hand.
- See centrifuge instruction manual for disinfection instructions.
- Always use appropriate carriers or inserts.
- Use of tubes with cracks or chips or excessive centrifugation speed may cause tube breakage, with release of sample, droplets, and an aerosol into the centrifuge bowl.
- Release of these potentially hazardous materials can be avoided by using specially designed sealed containers in which tubes are held during centrifugation.
- Centrifuge carriers and inserts should be of the size specific to the tubes used. Use of carriers too large or too small for the tube may result in breakage.
- Storage of glass tubes containing blood at or below 0°C may result in tube breakage.
- Do not remove conventional rubber stoppers by rolling with thumb. Remove stoppers with a twist and pull motion.
- Do not use tubes or needles if foreign matter is present.
- CTAD tubes must be protected from artificial and natural light during storage. Accumulated light exposure in excess of 12 hours can cause additive inactivation.
- BD Vacutainer® Plus Serum Tubes with clot activator are not to be used as a discard tube for coagulation studies.

Back to top

Appendix SOP-E: Procedure to Aliquot Plasma from Whole Blood

Purpose

The purpose of the working guideline is to describe the appropriate procedure for collecting blood of sufficient quantity for the purposes of obtaining plasma for proteomics, metabolomics and other analyses.

Background

Plasma is prepared using EDTA tubes.

Materials/Reagents

- 1.0 mL pipette and sterile, filtered tips
- 2.0 mL Cryovials with internal threads and silicone O-ring or washer

Procedure

Sample Processing and Centrifugation

Note: Read precautions below before proceeding to centrifugation.

1. Maintain the samples on wet ice at all times after collection until processing.
2. Sites should strive to process samples and freeze plasma aliquots as quickly as possible, and within 5 hours of collection (Please indicate the time of processing on the specimen acquisition form SAF).
3. Centrifuge the samples in a refrigerated (4°C) centrifuge at 1300 RCF (g) for 10 minutes.
   a. After centrifugation, three layers - Plasma, Buffy Coat (nucleated cells) and Red Blood Cells - should be clearly separated, top to bottom, respectively. Handle carefully to avoid disturbing layers.
   b. In some subjects, Lipids may cause the plasma to be cloudy or even form a "butter" layer on top of the plasma. Avoid any solids when pipetting.
   c. Hemolysis of Red Blood Cells may cause the Plasma to take on a red color (continue processing and note Hemolysis on SAF).
      i. If a specimen has severe hemolysis, a repeat sample may be needed. Notify and consult with Clinical staff as necessary.
4. Aliquot 0.5 mL of sample into each 2.0 mL cryovial as needed.
   a. Gently draw up the Plasma, taking care to avoid the buffy coat, or disturb the red blood cells.
   b. If the layers are compromised, repeat centrifugation step and continue recovering plasma.
   c. The number of aliquots will vary depending upon the specific protocol.
      i. Note: if the final aliquot is less than 0.5mL, distribute the residual volume as evenly as possible across the full aliquots.
5. Label 2.0 mL cryovials according to the protocol in Appendix SOP-B "**Procedure for Labeling Samples for the Biobank with 2-Dimensional Barcodes**."
6. Transfer samples to a deep freeze (-80°C) for storage as soon as possible. (Once samples are in aliquots, maintain on wet ice until frozen)

Precautions

- Do not centrifuge glass tubes at forces above 2,200 RCF in a horizontal head (swinging bucket) centrifuge as breakage may occur. Glass tubes may break if centrifuged above 1,300 RCF in fixed angle centrifuge heads.
- Balance tubes to minimize the chance of glass breakage. Match tubes to tubes of the same fill level. Note: At this g force, Vacutainer tubes are less likely to break in a swinging-bucket centrifuge than in a fixed-angle rotor.
- Ensure that tubes are properly seated in the centrifuge carrier. Incomplete seating could result in separation of the BD Hemogard™ Closures from the tube or extension of the tube above the carrier. Tubes extending above the carrier could catch on centrifuge head, resulting in breakage.
- Always allow centrifuge to come to a complete stop before attempting to remove tubes. When centrifuge head has stopped, open the lid and examine for possible broken tubes. If breakage is indicated, use mechanical device such as forceps or hemostat to remove tubes. Caution: Do not remove broken tubes by hand.
- See centrifuge instruction manual for disinfection instructions.
- Always use appropriate carriers or inserts.
- Use of tubes with cracks or chips or excessive centrifugation speed may cause tube breakage, with release of sample, droplets, and an aerosol into the centrifuge bowl.
- Release of these potentially hazardous materials can be avoided by using specially designed sealed containers in which tubes are held during centrifugation.
- Centrifuge carriers and inserts should be of the size specific to the tubes used. Use of carriers too large or too small for the tube may result in breakage.
- Storage of glass tubes containing blood at or below 0°C may result in tube breakage.
- Do not remove conventional rubber stoppers by rolling with thumb. Remove stoppers with a twist and pull motion.
- Do not use tubes or needles if foreign matter is present.
- BD Vacutainer® Plus Serum Tubes with clot activator are not to be used as a discard tube for coagulation studies.

Back to top

Appendix SOP-F: Procedure to Process and Aliquot Nasopharyngeal Swab/ Oropharyngeal Swab Fluid

Purpose

Nasopharyngeal and oropharyngeal swabs are stored in the same tube and processed together when both are available. The swab contents collected herein might be used for the detection of the presence of viral infection using various commercially available test kits such as the BinaxNow® RSV and BinaxNow® Influenza test kits and/or biomarkers such as small proteins; therefore, proper handling and storage is critical. Samples should be considered biohazardous and handled appropriately.

Background

A Nasopharyngeal Swab and associated Universal Transport Medium (UTM) can contain detectable levels of virus and/or virus particles. In addition, nasal swab/UTM specimen can contain detectable levels of proteins and other biomarkers used in various diagnostic and molecular techniques; therefore, proper handling and storage is critical. Expected yield is 3 mL.

Materials/Reagents

- 2.0 mL Cryovials
- 1.0 mL pipette and sterile, filtered tips or sterile pipette
- Collection Swab (Flocked mini-tip swabs are required)
- Universal transport medium

Procedure

1. Sample processing should take place within **5 hours** of collection. Please record the time of processing on the specimen acquisition form (SAF).
2. Pulse vortex the tube containing the swab submerged in transport medium to dislodge material from the swab.
3. Aliquot **0.5 mL** of nasal swab fluid into 2.0 mL cryovials (create 5 aliquots in 0.5 mL increments).
    a. Partial aliquots (< 0.5 ml) are not acceptable.
4. Label cryovials with 2-D barcodes according to the protocol in Appendix SOP-B '**Procedure for Labeling Samples for the Biobank with 2 Dimensional Barcodes**'.
5. As soon as possible, transfer samples upright to a deep freezer at 80ºC.
6. Dispose of the nasopharyngeal swab and any excess from the sample according to regulatory guidelines.

Note: Avoid repeated freeze-thaw cycles to preserve the integrity of the sample. No chemical additives are desired.

Back to top

Social Contact Models

Overview

Social ties are not formed randomly. The principle of forming ties with someone alike is "homophily" [1]. Social connections can exhibit homophily on both demographics and behaviors, and homophily tends to hold for both observed and unobserved traits. This is why the principle of clustering [2], in which *social* contacts of people with new STI diagnoses are offered testing and counseling, can be effective in some settings, as it captures people who are not necessarily first-degree transmission but instead are people who quite likely due to homophily share the same behavioral and social risks as the index case, and thus may have a similar likelihood of infection.

The people with whom a person forms closer connections—close contacts such as family or close friends joined by *strong* ties—tend to share ties among the contacts as well, leading to a dense area of overlap among contacts; we also form a large number of *weak* ties who might be thought of as acquaintances [3,4]. It is along these social ties that diffusion of infections (or interventions) occurs [5]. In sociology and social network analysis, strong ties are traditionally those with frequent, regular, close contact—close family and friends, sometimes close workplace, community, or neighborhood contacts—and weak ties are those with less regular contact.

Generally, human social networks tend to form as a set of denser (often homophilous) clusters, among people who share strong ties with each other, forming dense clusters or "communities," as they are called in network analysis, with weak ties bridging the communities and bringing infection (or information or ideas) in and out [6]. This has several implications for epidemic potential.

First, it can lead to bursty epidemics where small peaks are observed in case counts as a result of increasing cases moving from group to group [7]. Cases increase first within one network community as the infection spins up between the strong ties; the infection then spreads out along weak ties into a new community [8], which then has a peak. This was observed with the SARS-CoV-2 pandemic as early peaks formed within different ethnic and racial groups at different times.

Second, it can result in a failure to notice an outbreak until it has established itself in the population, as with the HIV outbreak among injecting drug users in Indiana a few years ago. Because the group was insular and traditionally does not access frequent care, the outbreak was out of control before it was noticed.

For infectious transmission to occur, an infectious and a susceptible person must have an *effective contact*. The effective contact is a contact sufficient for transmission to occur, which encompasses both contact type (needle sharing, sexual, casual) and timing of contact. For

instance, repeated or prolonged close contact is usually needed to transmit *M. tuberculosis*, but transmission of surface-stable pathogens such as Norovirus may not require two people to be in the same space at the same time. Respiratory infections pass from person to person with simple social contact as the effective contact to spread the infection. Thus, proximity + time proximate sets the risk, possibly increasing the importance of social clustering if more distal (weaker) social ties are at risk through direct transmission. For example: similar people are on the bus at the same time because they all work in the same place and go home to the same part of town on the same schedule: in the case of SARS-CoV-2, even incidental contacts may pass infection. Weaker social ties may also be at risk if the risk and behavioral patterns of social contacts mirror those of the seed case through homophily.

Assessment of community mixing patterns can provide insight into transmission or transmission risk in two ways. First, it provides clues as to what constitutes a network community in the local area and the properties of the weak tie bridges between communities. This understanding can be leveraged to get a sense of continued transmission vs peaks within each micro-population/community. Further, understanding the weak ties between communities provides a sense of how an infection might travel through a community for the purposes of intervening to slow transmission in the overall community and/or forecasting resources. Second, heterophilous (out-group) ties can increase risk in one group by virtue of being linked to a higher-prevalence group [9]. This phenomenon has been observed when having older partners who have had a longer time in which to acquire HIV leads to high risk among younger men who have sex with men in San Francisco [10] and young women in South Africa [11].

For an ongoing or longer-term epidemic, identifying the social network communities (the groups of people with dense sets of strong ties) is a key part of reducing onward transmission. First, incorrectly aggregating groups who appear to have similar traits or risk factors may obscure the actual overall risk to smaller network communities that have circulating uncontrolled infection and thus higher prevalence and incidence. For example, Latinos in North Carolina are frequently grouped together for HIV prevention efforts, although foreign-born Latinos with HIV tended to mix with other foreign-born Latinos and be part of heterosexual networks, whereas US-born Latinos with HIV tended to mix with other men who have sex with men and were less likely to be in a putative transmission cluster with another Latino [12]. Second, failing to identify which subgroups have the highest prevalence (and thus risk of new infection for group members) can mean that an intervention is applied in the wrong community, as with HIV pre-exposure prophylaxis where the people most at risk may not have access, or as with the earliest SARS-CoV-2 testing sites, which were frequently set up near communities at lower risk due to the ability to work and learn from home.

Once each subpopulation is then identified, an examination of the percent positivity and secondary attack rate can provide information about infection spread within the community. Percent positivity within each group can be compared to overall testing rates, although there may be bias if there is a disincentive to test within certain communities due to lack of access, lack of support if diagnosed, or economic harm due to having to leave work or care for a family member upon receipt of a diagnosis. However, a comparison of percent positivity rates between strong and weak ties may provide clues as to whether transmission is happening among close contacts (if higher among strong ties who are possibly part of the same transmission chain) or higher among weaker or incidental contacts (possibly indicating high community prevalence). For the purposes of a protocol designed to interrupt transmission, following contacts with the highest percent positivity or most consistent positive results may lead to opportunities to interrupt transmission through isolation of cases or other management practices.

A network analysis can be *sociometric* or *egocentric*. A sociometric (whole network) analysis is valuable for understanding the risk of transmission among the relationships along which an infection might diffuse. An egocentric (person-centered) analysis permits analysis of social mixing patterns, which may be of particular importance for a casually-transmitted infection that may pass not between identified contacts but instead among homophilous groups that happen to have something in common that puts them in the same place at the same time, increasing the importance of homophily as the driver of a latent network or increasing the importance of the weak ties between people for transmission.

Respondent-Driven Sampling

Respondent-driven sampling (RDS) is a link-tracing design that relies on cohort members to recruit their contacts to participate [13,14]. These chain referral/peer-recruitment methods can be especially effective in cases when the population at risk is not well known because either the population or the outcome is hidden. RDS rests on the assumption that people know their own networks better than a researcher, clinician, or public health professional. Additional benefits include that it is passive with respect to the tracing design; it allows community members to participate as much or as little as they choose; and especially among groups with a historical mistrust of public health or medicine, it leverages trust between people. This latter feature may enable a person recruited to the study to recruit a new study member more readily than a research or public health organization could. In some link tracing designs, the cohort can be weighted post recruitment to make inferences about the target population [14].

In link-tracing designs such as RDS, “seed” participants receive “coupons” to give to their contacts. The seed case can then recruit a number of contacts, from zero up to the number of coupons received, to participate. Anyone recruited by another study member is called a “peer” or “wave” participant. In a research setting, there are dual incentives for participation

and recruitment: one incentive is what is traditionally provided for sharing information with the researchers and participating in the study. The second is an additional incentive for the study participant to recruit other people into the study, which is often on the basis of each person successfully recruited. Peer participants who are successfully recruited also participate in the study and may receive coupons to distribute as well. This continues outward until the sample size is reached. No one participates more than once.

In traditional RDS, researchers lack control of the sampling process, although in many cases there are restrictions on eligibility. In the Snowball Study, we exerted some control over the process because we increased the incentive in some situations for cohort members to recruit peers who were members of groups that were under-represented in our cohort compared with the target population (Durham County).

Snowball sampling [15,16] is one type of link-tracing design in which a seed case is recruited, and that seed case collects their contacts for enrollment, who then collect *their* contacts for enrollment, moving outward the same way that you would roll up a snowball. This method not only mirrors how infections are transmitted from one person to another, but also is very similar to the process used by public health contact tracing. A separate set of eligibility criteria may be applied to the peer recruits, depending on the goal of the cohort. For instance, HIV contact tracing does not have a criterion for contacts about HIV status or location of residence (they can be tested locally or designated for tracing out-of-jurisdiction), whereas the seeds are generally newly-diagnosed HIV+ index cases located in the jurisdiction of the contact tracer.

As with contact tracing, participants in a link-tracing protocol can describe either strong vs weak ties. While contact tracers ultimately make the decision about which contacts will be investigated, in a link-tracing design, the participant decides whom to recruit. The definition of strong vs weak tie can be amended from its traditional definition based on what is being studied.

Snowball Study Population and Findings

We conducted an egocentric analysis of mixing patterns in the network among the 509 Snowball Study respondents (384 seed cases; 125 peers). These 509 respondents described 2,199 contacts (842 cohabitants; 1,357 other contacts).

The Snowball Study restricted enrollment to people aged 18 years and older (though contacts described could be any age). Seed cases were required to be residents of Durham County who were newly diagnosed with SARS-CoV-2 via PCR test at a Duke University Health System site; who did not opt out of participating in research; and who had checked their SARS-CoV-2 test result in their electronic medical record. Peer participants needed to live close enough to Duke University that they could be sampled.

We focused on age, ethnicity/race, and primary language spoken at home when seeking to enroll cohort members. The potentially eligible seed cases available to us did not match the target population for gender and some ethnicity/race categories (**Figure 1**).

Figure 1. Demographic Proportions of Target Population and Cohort. The dark blue bar on the left shows the demographic proportion of the target population (Durham County) and the gray bar on the right shows the demographic proportion of the enrolled Snowball Study cohort (seed and peer cases). The second blue bar ("Invited") and the third teal bar ("Consented") are the demographic proportions of the people available to the study who were invited to participate as seed cases and those who consented to join the study, respectively. Participation was restricted to 18+ years of age, so the age category proportions skew toward older ages for the study. Overall, the final Snowball Study sample tracks the invited seeds (the eligible set), except on ethnicity/race where Snowball participants with completed surveys were too few among Black or African American and too many among White participants when compared to who was invited to participate.

Among Snowball Study respondents, the egocentric network analysis revealed strong in-group mixing by ethnicity/race: 85% of contacts described by both White and Black respondents were the same race as the respondent. Homogeneity by ethnicity/race was somewhat lower for Asian (57% in-group) and Hispanic/Latinx (48% in-group) respondents, however Durham County (the target population) is composed of 5% Asian and 14% Hispanic/Latinx residents, so there was still a strong preference for in-group ties (**Figure 2**).

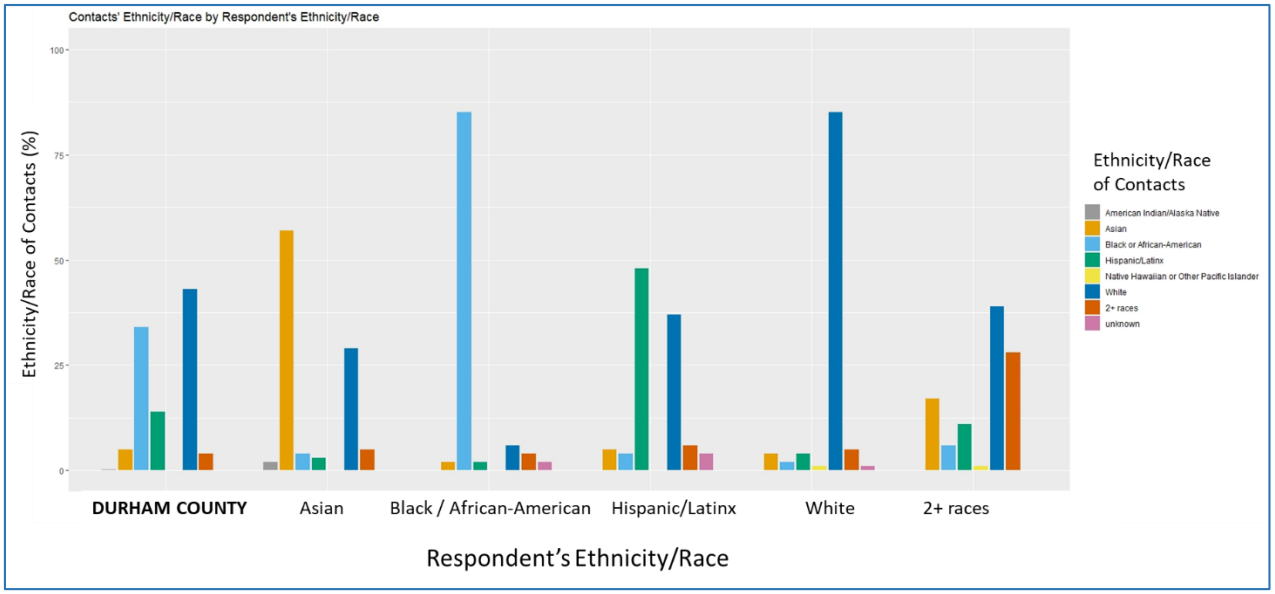

Figure 2. Contacts' Ethnicity/Race by Snowball Study Respondent's Ethnicity/Race. Under the assumption of no assortativity by ethnicity/race, the same pattern of ethnicity/race proportions as in Durham County (the target population) should have been repeated across all proportions of contacts' ethnicity/race (the y-axis) for each category of respondent's ethnicity/race (the x-axis). Instead, high "in-group" mixing is observed for all ethnic/racial categories of respondents.

Black or African American respondents reported contacts who were significantly more dissimilar in age, indicating high diversity by age when looking at each egocentric network component centered around a Black or African American respondent. Respondents aged 20-29 years had the most similarity in ages among their egocentric components, whereas dissimilarity increased up to respondents aged 50-59 years. This is unsurprising, considering that people aged 50-59 years in the United States are frequently in the work force and/or are caring for people in different generations (**Figure 3**).

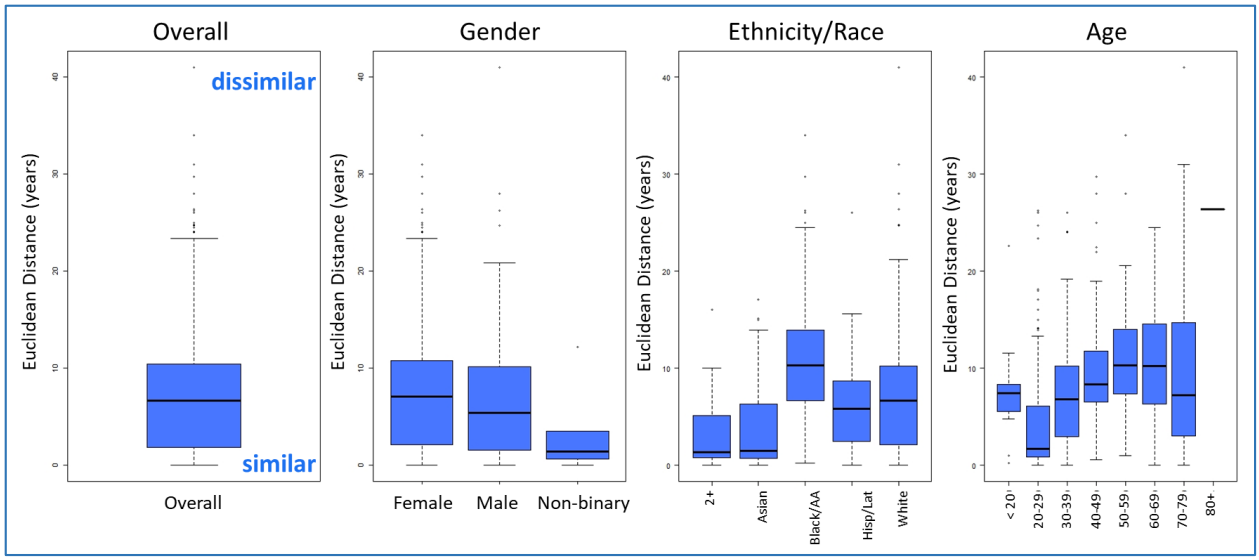

Figure 3. Box Plots of Euclidean Distance across Egocentric Network Components by Snowball Study Respondent's Traits. Euclidean Distance, a measure of *similarity* for a continuous variable across an entire egocentric network component, is shown for age among Snowball Study respondents by selected traits. A lower score indicates more similarity and a higher score is more dissimilarity; these ego-level scores are presented in box plots showing the age similarity scores for respondents in different demographic categories. Note that the trait selected for the respondent (ego) may not match the trait of the contact; this measure is for difference in age among an egocentric network component by the trait of the respondent only to assess whether some groups have a higher spread of ages in their egocentric networks. This graph shows that non-binary respondents had less age spread across their egocentric networks than respondents who identified as female or male: the 25th percentile and median lines are close to 0, indicating that half of non-binary respondents had little variability in age across the ages of themselves and the contacts they described. A higher rate of dissimilarity is observed among Black or African American people compared to other ethnicities and races; shown by the higher level of the levels of the median and 1st and 3rd quartiles. Age displays a u-shaped curve of age, with the exception of respondents younger than 20 years or aged 80 years or older who have more dissimilarity, both largely due to describing family or caregivers in other generations.

Mixing patterns are valuable for determining both where infections might concentrate in a network—one of the communities that may have a burst—and for understanding how an infection might reach a vulnerable community from another. For example, slowing transmission among Black or African American respondents might benefit not only that community as a whole, but slowing transmission among middle-aged community members might may also lead to protecting younger and older community members, as Black or African American respondents were more likely to report younger and older contacts.

Additionally, these linkages between communities theoretically can be leveraged for enrollment. The Snowball Study did not enroll as many Hispanic/Latinx or Black or African American respondents as desired to match the target population (**Figure 1**). If we were to leverage the contacts described, enrolling people who reported two or more races might be a pathway to reach Hispanic/Latinx or Black or African American participants (as seen in **Figure 2** as the relatively higher proportion of contacts of these ethnicities/races reported).

We used percent positivity as the metric for whether we were able to reach people who were positive but unaware of their diagnosis, combined with a comparison of the demographics of the seed cases and the peers. Seed cases implicitly had access to care and the means to get tested by virtue of having been tested in this academic medical center, and peers were given coupons for free testing. The Snowball Study percent positivity rates by month are presented below (**Figures 4-6**) and described in greater depth in the Eighth Quarter Progress Report.

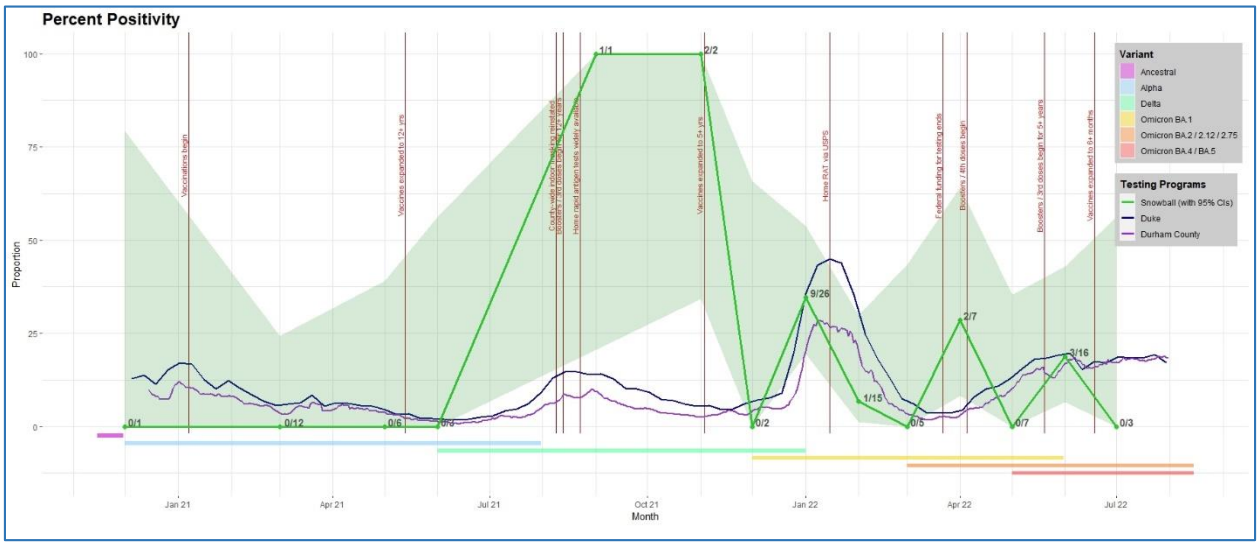

Figure 4. Percent Positivity Among Snowball Study Peers. In this graph, the Snowball Study percent positivity (green line) has a 95% confidence interval for each month of study recruitment. The target population (Durham County) percent positivity is shown in purple. A local academic medical center (Duke University) percent positivity is shown in dark blue. Below the horizontal 0 line, locally predominant variants are shown. Vertical red lines are policy and infrastructure changes which may affect testing availability and/or uptake.

We compared the demographics of seed cases to those of peers to assess whether we were reaching community members who may not have been engaged in a healthcare setting. We compared the demographic proportions of our seed cases to the peer participants who were not in a congregate living setting (**Table 1**). It seemed as if the RDS chains were functioning as intended, reaching further into the community: peers were less likely to be affiliated with Duke than seeds (approximately 1/3 vs approximately 1/2, respectively). Among peer participants, we increased the proportion of males, getting closer to the composition of the target population. Among peer participants, we increased the proportion of people aged 50-79 years (especially for those aged 70-79 years), indicating that there is a good pathway to enrolling this group via other cohort members.

Table 1. Comparison of Cohort Seeds and Cohort Peers. Demographic proportions of seed cases and peer participants are compared to assess whether peer recruitment can result in a cohort that more closely matches the target population.

	Total		Seeds		Peers		*not in shelter*		*in shelter*	
N	509		384		125		107		18	
Age										
0-19	10	2.0%	9	2.3%	1	0.8%	1	0.9%	0	0.0%
20-29	175	34.4%	138	35.9%	37	29.6%	34	31.8%	3	16.7%
30-39	139	27.3%	112	29.2%	27	21.6%	26	24.3%	1	5.6%
40-49	74	14.5%	56	14.6%	18	14.4%	12	11.2%	6	33.3%
50-59	54	10.6%	35	9.1%	19	15.2%	13	12.1%	6	33.3%
60-69	37	7.3%	25	6.5%	12	9.6%	11	10.3%	1	5.6%
70-79	19	3.7%	8	2.1%	11	8.8%	10	9.3%	1	5.6%
80+	1	0.2%	1	0.3%	0	0.0%	0	0.0%	0	0.0%
under 18	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Gender										
Male	174	34.2%	119	31.0%	55	44.0%	37	34.6%	18	100.0%
Female	326	64.0%	261	68.0%	65	52.0%	65	60.7%	0	0.0%
Non-Binary	9	1.8%	4	1.0%	5	4.0%	5	4.7%	0	0.0%
Other	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Race										
Amer Indian	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Asian	58	11.4%	45	11.7%	13	10.4%	13	12.1%	0	0.0%
Black or AA	105	20.6%	76	19.8%	29	23.2%	19	17.8%	10	55.6%
Hispanic/Latinx	37	7.3%	31	8.1%	6	4.8%	6	5.6%	0	0.0%
Nat Hawaiian	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
White	285	56.0%	211	54.9%	74	59.2%	66	61.7%	8	44.4%
2 or more	24	4.7%	21	5.5%	3	2.4%	3	2.8%	0	0.0%
unknown	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Duke	237	46.6%	200	52.1%	37	29.6%	37	34.6%	0	0.0%

Note: Some (n=18) of the peers were recruited from a congregate living setting. Some of the recruitment within that location was assisted by the directors, so these peers were excluded from some of the cohort analyses.

We also created sender-receiver matrices (**Tables 2** and **3**) to look at recruitment patterns among the demographic traits of interest to assess the comfort level with recruitment among different groups.

The matrices show the attribute of the referrer on the rows, with the attribute of the peer who was successfully recruited into the study and who completed the survey as the column. Study cohort members who recruited multiple people appear multiple times on their row, and the total matrix size is the number of peers successfully recruited into the study, consented, and with a complete survey. Numbers along the diagonal (bolded) reflect in-group mixing, where the referrer or coupon distributor successfully recruited someone with the same attribute value. These patterns can be leveraged for recruitment.

In-group recruitment was common by ethnicity/race (**Table 2**). Asian, Black or African American, and White peers tended to be successfully recruited by someone of the same race. However, Hispanic/Latinx participants stand out: none of the coupons distributed by Hispanic participants went to another Hispanic participant, or none of the Hispanic peers were recruited by a prior Hispanic participant. This finding is of interest because we did not enroll as many Hispanic participants as we had hoped to. We also see that Black or African American participants were able to successfully recruit within race, which could have been a way to increase participation among a group that has been historically excluded from research.

Table 2. Sender-Receiver Matrix by Ethnicity/Race. The referring participant is represented in the rows and the recruited participant is represented in the columns.

Coupon recipients

Coupon distributors: RACE	Asian	Black or AA	Hisp	White	2 or more	
Asian	**8**	1		5		*14*
Black or AA		**23**		9		*32*
Hisp	2	1		3	1	*7*
White	3	4	5	**57**	1	*70*
2 or more			1		**1**	*2*
	13	*29*	*6*	*74*	*3*	*125*

Recruitment tended to occur either within the same age band, or one age band above or below (**Table 3**). Participants aged 20-29 years who managed to successfully recruit another participant mostly did so within their own age group, although they also recruited 30-39 and 50-59 year old participants. Many of the recruitments more than two age bands apart are family members. Participants aged 30-39 years recruited people across every age band except 80 years or older; in fact, no one that was recruited as a peer was 80 years of age or older. All people in the survey who were older than 80 were seed cases (the peer column for 80+ years is empty). We can see that 30-39 year olds distributed to nearly every age band, but if we look down the column for 30-39 year olds we can see that they were mostly recruited by their own age group and one band younger.

On the other hand, participants aged 50-59 years only successfully recruited two people aged 40-49 years and recruited five people in their own age band (50-59 years).

However, 50-59 year olds were recruited by people across the spectrum, from the group younger than 20 years all the way through 60-69 year olds. Those aged 50-59 years were not as effective at recruiting other participants, but they appear amenable to recruitment by other members, many of whom were family members.

Table 3. Sender-Receiver Matrix by Age. The referring participant is represented in the rows and the recruited participant is represented in the columns.

Coupon distributors \ Coupon recipients	0-19	20-29	30-39	40-49	50-59	60-69	70-79	80+	
0-19					2				*2*
20-29		**27**	10		2				*39*
30-39	1	6	**14**	4	2	4	3		*34*
40-49			2	**8**	6	1	1		*18*
50-59				2	**5**				*7*
60-69		4	1	4	2	**3**	2		*16*
70-79						3	**5**		*8*
80+						1			*1*
	1	*37*	*27*	*18*	*19*	*12*	*11*	*0*	*125*

For a further analysis of percent positivity, we separated Snowball tests administered by whether the peer was a *strong* or *weak* tie. We calculated this in two ways. In the first graph, cohabitants and local family members were strong ties and all others were weak ties. In the second graph, strong ties were those peers identified as a contact in the coupon source's survey and peers not identified were weak ties. For both graphs, the Snowball proportions have a 95% confidence interval and each point is labeled with the number of positive PCR tests over the total number of PCR tests.

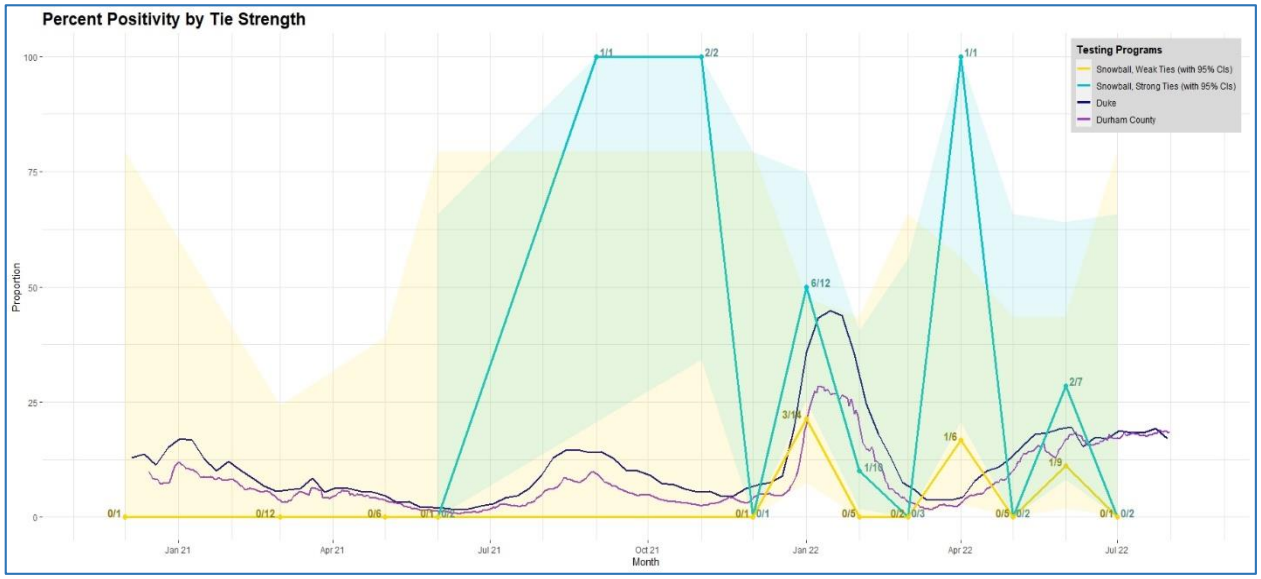

Figure 5. Snowball Study Percent Positivity by Strong and Weak Ties. Strong ties (blue) are household members and local family members. Weak ties (yellow) are all others.

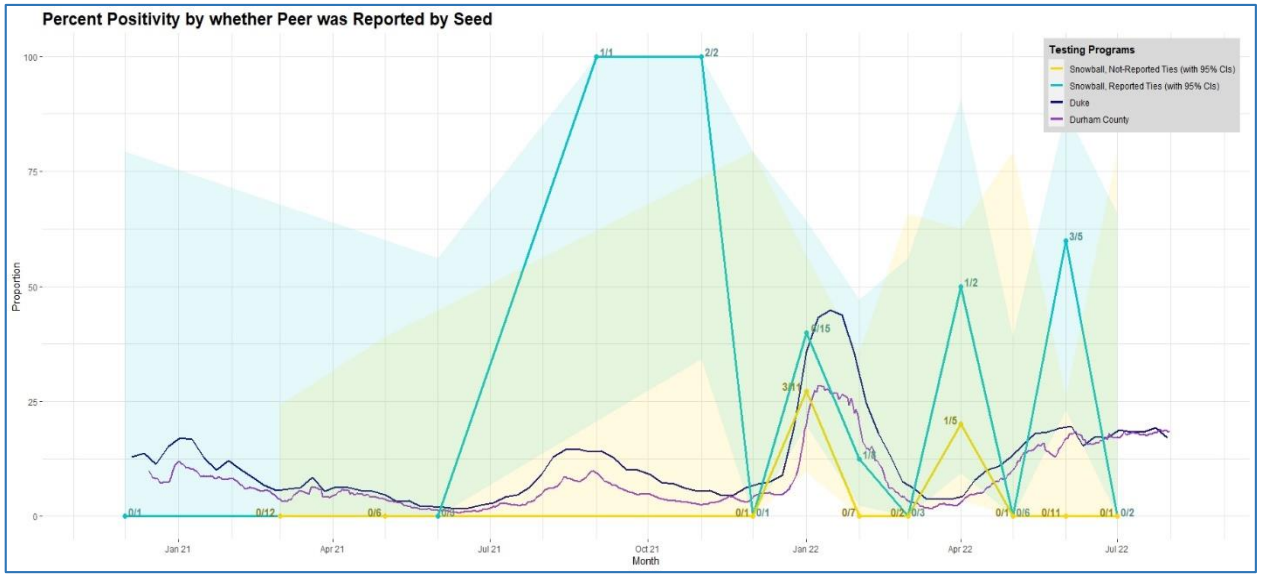

Figure 6. Snowball Study Percent Positivity by whether Peer was Reported by Source. Strong ties (blue) were described in the respondent's survey and were recruited for study participation. Weak ties (yellow) are those who were not described in the respondent's survey, but were recruited for study participation anyway.

We also extended this among strong ties by calculating the secondary attack rates within the household, then compared the household secondary attack rates by selected characteristics of the household or seed case to determine whether some factors were associated with higher or lower secondary attack rates (**Table 4**).

Among 384 seeds with complete surveys, 302 (79%) had at least one cohabitant who was not also a contemporaneous seed case, and who also let us know the general location where they believed that they were infected. Of these, one-quarter (74; 25%) believed that they caught SARS-CoV-2 within their own home. For these 74, we calculated the number of likely secondary infections within the household based on the cohabitants who displayed symptoms or were diagnosed after the initial household member brought the infection into the home, out of the total number of household members minus the initially infected person. For the other 228 (75%) who believed they were infected outside their home, we calculated the number of likely secondary infections from the seed within the household based on the number of cohabitants who displayed symptoms or were diagnosed 3-12 days after the seed case's symptom onset, out of the total number of household members besides the Snowball seed. We present these results below, as a total secondary attack rate and then by the selected characteristics.

Table 4. Secondary Attack Rates Within Households by Selected Characteristics.

	n	Secondary Attack Rate (%) (95%CI)
Overall	302	21 (19-23)
Children in household*		
Yes	106	23 (20-25)
No	196	19 (12-19)
Cohabitants composition*		
Some or all familial relations	250	23 (20-24)
No familial relations	52	12 (7-19)
Duke affiliation		
Yes; affiliated	154	22 (20-26)
No; not affiliated	148	21 (17-22)
Household social distancing prior 2 weeks*		
Always, Most of the time	134	28 (25-31)
Sometimes, Rarely, Never	168	15 (11-16)
Household mask wearing prior 2 weeks*		
Always, Most of the Time, Didn't leave house	250	23 (20-24)
Sometimes, Rarely, Never	52	12 (12-21)
Seed's daily contacts		
Above median (>15)	144	20 (17-23)
At or below median (0-15)	158	23 (19-25)
Predominant circulating variant†		
Delta	63	26 (24-33)
Omicron BA.1	138	22 (16-22)
Omicron BA.2 / 2.12 / 2.75	84	19 (12-20)
Omicron BA.4 / BA.5	15	12 (6-20)

* Difference in groups significant at $\alpha \leq 0.05$

† 2 participants recruited during Alpha wave were dropped from this analysis due to small cell size

We expected to see a lower secondary attack rate in households affiliated with Duke, as these households are likely better resourced than other households in the area and had excellent access to testing, presumably leading to earlier detection and hopefully less onward transmission. We were also surprised that households that reported higher levels of mitigating social behaviors (masking, social distancing) at the time of the household's infection had significantly higher secondary attack rates than households that were less likely to report engaging in these behaviors. We cannot explain this result from the data that we have. We did expect the higher secondary attack rates among households with children and households where family members resided together that we observed.

Delta appeared to have a higher secondary attack rate than the other variants among this cohort. This is not in line with estimated effective reproduction numbers for the variants, in which Omicron and its sublineages appear to be more transmissible. It is possible that higher vaccination rates and less severe symptoms led to fewer Omicron infections being detected within households, and could indicate that a lower proportion of cases were or are being counted in the communities.

Methods for Social Mixing Findings

Cohort and Target Population

One goal of the Snowball project was to reach people using RDS methodology who might not otherwise have been tested. We did this by 1) checking how well our cohort matched the target population; 2) examining how the seeds differed from the peers, under the assumption that the chains allow us to reach people who might not have been able themselves to access care, as well as looking at in- vs out-group patterns of recruitment; and 3) by examining the percentage of peers who tested positive for COVID, as they represent a group who may not have been tested elsewhere.

We periodically checked the demographic proportions of our cohort against the target population (a table of proportions for each variable of interest). We used this to prioritize seed cases and incentivized peer recruitment among people who were members of groups who were not well represented in the study cohort and de-emphasized participation among members of groups who were over-represented in the cohort. This is the check of representativeness, and success in the first step of guiding enrollment to reach either a representative or an under-represented set of people for the cohort.

We also compared the demographic proportions of the seed cases to the demographic proportions of the peer participants to assess whether we were reaching a different set of community members by creating a table of seed demographic proportions compared to peer demographic proportions (a table of proportions for each variable of interest). The sender-receiver matrices were restricted to successfully recruited participants (for the Snowball Study, this was defined as any participant who was recruited, provided informed consent to participate, and completed the survey). The demographic category of the recruiting participant is the row and the demographic category for the person who was recruited is the column. The matrix cells sum to the total number of people who were successfully recruited into the study by another participant. Any participant who recruited more than one person into the study would be represented that many times in the matrix, but each person successfully recruited would be represented only once. For the sender-receiver matrices, we created a dataset with one row for each successfully recruited peer with the peer's demographic traits as well as the demographic traits of the referring participant (even if the referring participant was repeated across multiple rows). We then ran a cross-tabulation of the referring participant's demographic traits by the recruited participant's demographic traits. These are the second checks of the success of the peer recruitment method in collecting a representative cohort.

Finally, we calculated the percent positivity overall and split by whether the peer recruited was strongly or weakly tied to the recruiting participant.

Percent Positivity

To calculate the percent positivity, we aggregated the number of tests conducted for the Snowball Study to the month in which sampling occurred and calculated the proportion of tests that were positive for SARS-CoV-2. We performed this calculation for all tests conducted, and then split by whether the person who was sampled was recruited along a strong or a weak network tie to compare percent positivity. We calculated a confidence interval for proportions for the Snowball Study samples, and compared this to the percent positivity (positive tests divided by total tests administered) for local testing programs – in our case, the county testing program (abstracted from CDC's website [17,18]) and the programs at two local academic medical centers (Duke University and the University of North Carolina at Chapel Hill [data not publicly available]). R code for this analysis is included in Appendix SCM-A: Snowball Study Social Mixing Analysis – Example Code.

Mixing Patterns

To check for assortativity (preference for homophily/in-group mixing) among categorical variables such as combined ethnicity/race or gender, we calculated the proportion of contacts per demographic trait category by the respondent's demographic trait category for categorical variables and compared those proportions to the background/target population. If there is not homophilous mixing on that trait, the proportions of contacts should follow the proportions of the target population. Instead, Figure 2 shows high ethnic/racial homophily among Snowball respondents, as the proportions of same-race contacts are observed to be well above those expected based on the background/target population.

To check for similarity between egos and their contacts among a continuous variable (age; Figure 3), we calculated Pearson's Phi statistic, which provides an average Euclidean distance measurement among the ego and contacts in each egocentric network component [19]; a lower score indicates more similarity and a higher score indicates greater dissimilarity. Pearson's Phi compares ages across the entire egocentric network component (not just from ego to contact) and accounts for number of contacts, so the measure does not increase simply due to having a greater number of contacts.

Ordinal variables (e.g., socioeconomic status) can be changed to a binary variable above or below a stated threshold and treated as categorical or assigned a weight based on number of strata crossed for a continuous assessment, although some information is lost in both transformations.

It is important to note that if these analyses are conducted for the social ties described by the respondent, the analysis is likely to be relevant to the mixing patterns among strong rather than weak ties. This has implications for the drivers influencing the formation of the network communities in which cases could rise quickly (leading to one of the bursts in the

epidemic). To better understand whether risk differs by some traits, both the plots of in-group vs outgroup mixing for categorical variables and Pearson's Phi can be calculated for different traits of interest (subgroup analysis). For example, in-group vs out-group mixing by ethnicity/race among sexual partners may differ by gender and/or sexual orientation, and Pearson's Phi for age of social contacts may differ according to whether the respondent is affiliated with a university.

R code for these analyses is included in Appendix SCM-A: Snowball Study Social Mixing Analysis – Example Code.

Secondary Attack Rates

For this analysis, we calculated the secondary attack rate within households, using survey data from seeds with at least one other household member. The secondary attack rate was the number of people who were documented or likely to be positive based on test results and/or symptoms, over total household members minus the respondent. For houses in which we enrolled multiple people, we used only the first person enrolled.

For seeds who believed that they were infected at home, we calculated the number of other household members who were also diagnosed or symptomatic at the time of the seed case's survey completion, over the total number of household members minus the initially-infected person, as the likely secondary infections within the household based on the cohabitants who displayed symptoms or were diagnosed after the initial household member brought the infection into the home.

For seeds who believed that they were infected outside of the home, we calculated the proportion of cohabitants who had symptoms or were diagnosed 3-12 days after the seed case's symptom onset or until time of survey completion (if less than 12 days after symptom onset) as an indicator of likely onward (secondary) transmission from the seed.

We calculated the overall household secondary attack rate and then compared rates among:

- Households with children vs not;
- Households comprising family members or significant others vs roommates:
- Households affiliated with Duke;
- By seed's reported household social-distancing;
- By seed's reported mask-wearing;
- By seed's number of daily contacts outside the home being above or at/below this analysis subset's median; and
- The predominant circulating variant when the seed enrolled in the study.

We calculated 95% confidence intervals for the proportions for the comparisons.

R code for these analyses is included in Appendix SCM-A: Snowball Study Social Mixing Analysis – Example Code.

R Program Code for Analyses

Example R code is supplied as Appendix SCM-A: Snowball Study Social Mixing Analysis – Example Code for the comparison of the target population and cohort, the assessment of mixing patterns, the percent positivity graphs, and the secondary attack rates.

References for Social Contact Models

1. McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annu Rev Sociol, 27*, 415-44. Retrieved from http://www.jstor.org/stable/2678628 Return to text
2. Centers for Disease Control and Prevention. (2008). Recommendations for Partner Services Programs for HIV Infection, Syphilis, Gonorrhea, and Chlamydial Infection. *MMWR, 57*(RR-9). Retrieved from http://www.cdc.gov/mmwr/pdf/rr/rr5709.pdf Return to text
3. Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology, 78*(6), 1360-80. Return to text
4. Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory, 1*, 201-33. Return to text
5. Heckathorn, D. D., Broadhead, R. S., Anthony, D. L., & Weakliem, D. L. (1999). AIDS and Social Networks: HIV Prevention through Network Mobilization. *Sociological Focus, 32*(2), 159-78. Retrieved from internal-pdf://sociological_focus-0569955075/sociological_focus.pdf Return to text
6. Friedkin, N. E. (1980). A Test of Structural Features of Granovetter's Strength of Weak Ties Theory. *Social Networks, 2*, 411-22. Return to text
7. Garnett, G. P., Hughes, J. P., Anderson, R. M., Stoner, B. P., Aral, S. O., Whittington, W. L., Handsfield, H. H., & Holmes, K. K. (1996). Sexual Mixing Patterns of Patients Attending Sexually Transmitted Diseases Clinics. *Sex Transm Dis, 23*(3), 248-57. Return to text
8. Shu, P., Tang, M., Gong, K., & Liu, Y. (2012). Effects of Weak Ties on Epidemic Predictability in Community Networks. *ArXiv*. Retrieved from https://arxiv.org/pdf/1207.0931.pdf. Return to text
9. Anderson, R. M., Gupta, S., & Ng, W. (1990). The Significance of Sexual Partner Contact Networks for the Transmission Dynamics of HIV. *J Acquir Immune Defic Syndr, 3*(4), 417-29. Return to text
10. Service, S. K., & Blower, S. M. (1995). HIV Transmission in Sexual Networks: An Empirical Analysis. *Proc Biol Sci, 260*(1359), 237-44. doi:10.1098/rspb.1995.0086 Return to text
11. Pettifor, A. E., Rees, H. V., Kleinschmidt, I., Steffenson, A. E., MacPhail, C., Hlongwa-Madikizela, L., Vermaak, K., & Padian, N. S. (2005). Young People's Sexual Health in South Africa: HIV Prevalence and Sexual Behaviors from a Nationally Representative Household Survey. *AIDS, 19*(14), 1525-34. Return to text
12. Dennis, A. M., Hue, S., Pasquale, D., Napravnik, S., Sebastian, J., Miller, W. C., & Eron, J. J. (2015). HIV Transmission Patterns among Immigrant Latinos Illuminated by the

Integration of Phylogenetic and Migration Data. *AIDS Res Hum Retroviruses, 31*(10), 973-80. doi:10.1089/AID.2015.0089 *Return to text*

13. Heckathorn, D. D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems, 44*(2), 174-99. *Return to text*

14. Heckathorn, D. D. (2002). Respondent-Driven Sampling Ii: Deriving Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems, 19*(1), 11-34. *Return to text*

15. Heckathorn, D. D. (2011). Snowball Versus Respondent-Driven Sampling. *Sociol Methodol, 41*(1), 355-66. doi:10.1111/j.1467-9531.2011.01244.x *Return to text*

16. Heckathorn, D. D., & Cameron, C. J. (2017). Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling. *Annual Review of Sociology, 43*(1), 101-19. doi:10.1146/annurev-soc-060116-053556 *Return to text*

17. CDC COVID-19 Response Team. (2022). United States COVID-19 County Level of Community Transmission Historical Changes - Archived Retrieved 14-September-2022, from CDC COVID-19 Response, https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-County-Level-of-Community-T/nra9-vzzn. *Return to text*

18. CDC COVID-19 Response Team. (2022). Weekly COVID-19 County Level of Community Transmission as Originally Posted Retrieved 14-September-2022, from CDC COVID-19 Response, https://data.cdc.gov/Public-Health-Surveillance/Weekly-COVID-19-County-Level-of-Community-Transmis/dt66-w6m6. *Return to text*

19. Perry, B. L., Pescosolido, B. A., & Borgatti, S. P. (2018). *Egocentric Network Analysis: Foundations, Methods, and Models*. New York, N.Y.: Cambridge University Press.

Return to text

Back to top

Appendix SCM-A: Snowball Study Social Mixing Analysis – Example Code and Files

The Snowball Toolkit has been made publicly available on the website:

https://sites.duke.edu/dnac/resources/snowballtoolkit/.

The site, which briefly introduces the Snowball Toolkit, includes the R code and example input files for the Snowball Study social mixing analyses.

Three zip files are linked on the site for these analyses. The first includes example input files (.xlsx and .csv) and R code for the main analyses used to understand social mixing patterns and assess the Snowball Study outcomes, which included its effectiveness in enrolling a representative cohort.

- The social_mixing_analysis.R file includes the code to compare the study cohort to the target population; compare mixing patterns among cohort members to what would be expected if people mixed in the community without any assortativity along both categorical (i.e., ethnicity/race) or continuous (i.e., age) attributes of the people; the graphs to compare percent positivity of the study or sampling program being tested to other local testing programs; and the calculation of secondary attack rates within a household or defined setting, both overall and by selected traits.
- contacts.xlsx is the input attributes file that is the basis of the social mixing analyses. It has the attributes of everyone represented in the network: the enrolled participants and the contacts described by that person. For each enrolled participant (SnoID) who took the survey (the respondent), this long-format dataset has one row for the respondent and a row for each contact described by that respondent. Each row has columns to track the referring (CASE), household (HH), and familial (FAM) connections; relationship type between the respondent and cohabitant (Recommend, cohab, cont, fam); enrollment status of each person (Seed, Peer); demographic attributes of the contact being described (gender, age, ethnicity/race, employment status, education level, marital status); and the respondent's assessment of the contact's SARS-CoV-2 status and reason for that assessment. This is test data.
- demographics.csv is the demographic proportions of the background / target population. The file has columns for gender/sex, age band, and combined ethnicity and race. This is real data for Durham County, NC, as estimated for July 2019.
- household_data.csv contains the pertinent variables for the secondary attack rate calculations. This wide-format dataset has one row for each enrolled participant (sno_id) who had at least one other cohabitant (up to 7 cohabitants). The dataset

includes a diagnosis date for the referring participant (target_date); indicators used for the stratified secondary attack rate calculations (household_members:predom.var); and a set of columns for each cohabitant (series of indicator columns for demographics and SARS-CoV-2 status, including a date used for the days between infections [hierarchical symptom onset to diagnosis]); household indicators used for the stratified analyses (household_response, mask_practice); and the days between the respondent's and cohabitant's SARS-CoV-2 dates (cohab_covid_daysdiff_1: cohab_covid_daysdiff_7). This is test data.

- PercentPositive.xlsx has the information used for the percent positivity comparisons. The file has the ID number of the person referred to the study for testing, the dates of sampling, and the results. This is test data.
- strongWeak.xlsx contains the indicators for whether the contact described was referred by a strong or weak contact. This file has the ID number of the contact described (Record ID) and 3 different examples of ways to stratify strong vs weak contacts (as the *a priori* measure commonly used in the social sciences; by coworker status; and whether the contact was described as a contact in the referring participant's survey). This is test data.

The last two zip files contain the data made available by CDC disclosing number of SARS-CoV-2 tests and positive results by FIPS code by date. These were the basis of our comparison of the Snowball Study's percent positivity against the target population, which was Durham County (FIPS code 37063).

All study data provided in the input files are test data and do not represent actual study data collected.

Back to top

Predictive Model

Predictive Model from Biometric Data that Maximizes Sensitivity to Recommend/Prescribe Diagnostic Testing

Because diagnostic testing occurs at a single point in time and must be optimally timed to detect infection, the Snowball study integrated with the CovIdentify study (Duke IRB #2020-0412), which uses biometric monitoring from common consumer wearable devices to detect digital biomarkers that can be used to detect infection early in the course of illness. Participants who enrolled in Snowball were also invited to join the CovIdentify study. Study participants were given the option of sharing their data from a wearable device they already owned, or were offered a smartwatch for biometric data collection of measures such as heart rate, physical movement, sleep data, and blood oxygen saturation. Forty Snowball participants co-enrolled in CovIdentify via this referral pathway and were integrated into the larger CovIdentify cohort that was used to develop the predictive model from the biometric cohort.

The Intelligent Testing Allocation (ITA) method used to develop the predictive model and the results and performance of the model were made available as a preprint manuscript in April of 2022; the final, peer-reviewed version was published in the journal *NPJ Digital Medicine* in September of 2022. The citation is below and the final published paper itself is attached to this Toolkit as Appendix PM-A.

> *Shandhi MMH, Cho PJ, Roghanizad AR, Singh K, Wang W, Enache OM, Stern A, Sbahi R, Tatar B, Fiscus S, Khoo QX, Kuo Y, Lu X, Hsieh J, Kalodzitsa A, Bahmani A, Alavi A, Ray U, Snyder MP, Ginsburg GS, Pasquale DK, Woods CW, Shaw RJ, Dunn JP. A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19. NPJ Digit Med. 2022 Sep 1;5(1):130. doi: 10.1038/s41746-022-00672-z. PMID: 36050372; PMCID: PMC9434073*

Lastly, from the outset of the CovIdentify project, the study team has been committed to making any algorithms developed for the study publicly available. To that end, the de-identified CovIdentify dataset generated and/or analyzed during the current study will be submitted 1 year from the publication date of the manuscript to the Digital Health Data Repository (DHDR) repository:

https://github.com/DigitalBiomarkerDiscoveryPipeline/Digital_Health_Data_Repository

under the title *BigIdeasLab_CovIdentify*. The ITA model development code used for this manuscript is available at the digital biomarker discovery pipeline (DBDP) GitHub repository (https://github.com/DigitalBiomarkerDiscoveryPipeline/CovIdentify).

Back to top

Appendix PM-A: Shandi et al., 2022

A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19.

npj Digital Medicine (2022) 5:130 ; https://doi.org/10.1038/s41746-022-00672-z

Back to top

npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19

Md Mobashir Hasan Shandhi[1,11], Peter J. Cho[1,11], Ali R. Roghanizad[1,11], Karnika Singh[1], Will Wang[1], Oana M. Enache[2], Amanda Stern[1], Rami Sbahi[1], Bilge Tatar[1], Sean Fiscus[1], Qi Xuan Khoo[1], Yvonne Kuo[1], Xiao Lu[1], Joseph Hsieh[1], Alena Kalodzitsa[1], Amir Bahmani[3], Arash Alavi[3], Utsab Ray[3], Michael P. Snyder[3], Geoffrey S. Ginsburg[4], Dana K. Pasquale[5,6], Christopher W. Woods[7,8], Ryan J. Shaw[9,10] and Jessilyn P. Dunn[1,2]✉

Mass surveillance testing can help control outbreaks of infectious diseases such as COVID-19. However, diagnostic test shortages are prevalent globally and continue to occur in the US with the onset of new COVID-19 variants and emerging diseases like monkeypox, demonstrating an unprecedented need for improving our current methods for mass surveillance testing. By targeting surveillance testing toward individuals who are most likely to be infected and, thus, increasing the testing positivity rate (i.e., percent positive in the surveillance group), fewer tests are needed to capture the same number of positive cases. Here, we developed an Intelligent Testing Allocation (ITA) method by leveraging data from the CovIdentify study (6765 participants) and the MyPHD study (8580 participants), including smartwatch data from 1265 individuals of whom 126 tested positive for COVID-19. Our rigorous model and parameter search uncovered the optimal time periods and aggregate metrics for monitoring continuous digital biomarkers to increase the positivity rate of COVID-19 diagnostic testing. We found that resting heart rate (RHR) features distinguished between COVID-19-positive and -negative cases earlier in the course of the infection than steps features, as early as 10 and 5 days prior to the diagnostic test, respectively. We also found that including steps features increased the area under the receiver operating characteristic curve (AUC-ROC) by 7–11% when compared with RHR features alone, while including RHR features improved the AUC of the ITA model's precision-recall curve (AUC-PR) by 38–50% when compared with steps features alone. The best AUC-ROC (0.73 ± 0.14 and 0.77 on the cross-validated training set and independent test set, respectively) and AUC-PR (0.55 ± 0.21 and 0.24) were achieved by using data from a single device type (Fitbit) with high-resolution (minute-level) data. Finally, we show that ITA generates up to a 6.5-fold increase in the positivity rate in the cross-validated training set and up to a 4.5-fold increase in the positivity rate in the independent test set, including both symptomatic and asymptomatic (up to 27%) individuals. Our findings suggest that, if deployed on a large scale and without needing self-reported symptoms, the ITA method could improve the allocation of diagnostic testing resources and reduce the burden of test shortages.

npj Digital Medicine (2022)5:130 ; https://doi.org/10.1038/s41746-022-00672-z

INTRODUCTION

The COVID-19 pandemic has severely impacted our world, with more than 562 million COVID-19 cases and 6.37 million deaths estimated worldwide[1]. In the US alone, there have been more than 90 million cases and 1 million deaths at the time of writing[2]. Mass surveillance testing has been identified as the most effective tool to monitor the spread of infectious diseases including COVID-19[3]. However, a combination of cost, availability, and impracticality of frequent and widespread testing impedes the mass epidemiologic surveillance needed to curb new disease outbreaks. To date, COVID-19 diagnostic test shortages are still prevalent globally, and shortages continue to occur in the US with the onset of new variants (e.g, Delta, Omicron)[4–6]. For example, when the Delta variant emerged in July 2021, daily demand for tests across the US surged from 250k to 1.5 million in the span of 2 months[7]. A similar circumstance occurred with the Omicron variant, where testing capacity failed to meet the sudden demand[8–10]. Inefficient diagnostic testing is also exacerbating the emerging threat of monkeypox in the US[11,12]. Furthermore, rural-urban disparities in testing access have worsened existing inequities resulting in further harm to underserved communities[13,14]. In June 2020, it was estimated that 64% of counties in the United States, predominantly rural, did not have access to COVID-19 testing[15]. Such circumstances lead to underreporting of COVID-19 incidence and may lead to a premature sense of security and unwarranted changes in public health measures[14]. Thus, there is an unprecedented need to improve our current and future methods for mass COVID-19 surveillance testing, especially as stronger testing capacity has been associated with reduced mortality and greater pandemic control[16].

By targeting surveillance testing toward individuals who are more likely to be infected with the disease, more positive cases

[1]Department of Biomedical Engineering, Duke University, Durham, NC, USA. [2]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA. [3]Department of Genetics, Stanford University, Stanford, CA, USA. [4]All of Us Research Program, National Institutes of Health, Bethesda, MD, USA. [5]Department of Sociology, Duke University, Durham, NC, USA. [6]Department of Population Health Sciences, School of Medicine, Duke University, Durham, NC, USA. [7]Division of Infectious Diseases, Duke University Medical Center, Durham, NC, USA. [8]Durham VA Medical Center, Durham, NC, USA. [9]School of Nursing, Duke University, Durham, NC, USA. [10]Duke Mobile App Gateway, Clinical and Translational Science Institute, Duke University, Durham, NC, USA. [11]These authors contributed equally: Md Mobashir Hasan Shandhi, Peter J. Cho, Ali R. Roghanizad. ✉email: jessilyn.dunn@duke.edu

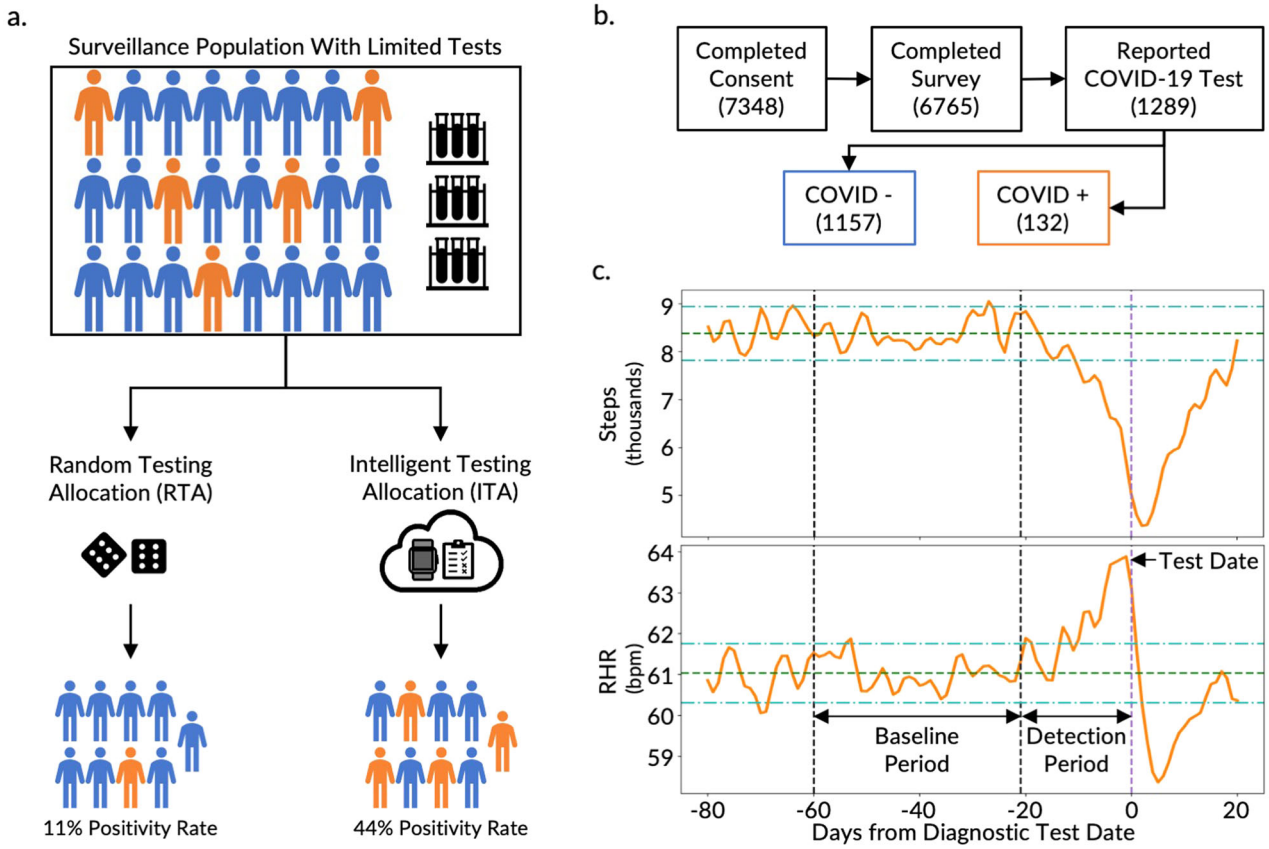

Fig. 1 Overview of the Intelligent Testing Allocation (ITA) model, the CovIdentify cohort, and data. a Overview of the ITA model in comparison to a Random Testing Allocation (RTA) model that demonstrates the benefit of using the ITA model over existing RTA methods to improve the positivity rate of diagnostic testing in resource-limited settings. Human symbols with orange and blue colors represent individuals with and without COVID-19 infection, respectively. **b** A total of 7348 participants were recruited following informed consent in the CovIdentify study, out of whom 1289 participants reported COVID-19 diagnostic tests (1157 diagnosed as negative for COVID-19 and 132 diagnosed as positive for COVID-19). **c** The top panel shows the time-averaged step count and the bottom panel shows the time-averaged resting heart rate (RHR) of all participants ($n = 50$) in the training set (Supplementary Fig. 3, blue) who were tested positive for COVID-19 with the pre-defined baseline (between −60 and −22 days from the diagnostic test) and detection (between −21 and −1 days from the diagnostic test) periods marked with vertical black dashed lines. The dark green dashed lines and the light green dash-dotted lines display the baseline period mean and ± 2 standard deviations from the baseline mean, respectively. The light purple dashed vertical line shows the diagnostic test date.

can be captured with the same number of tests, increasing the positivity rate of the tested population (Fig. 1a)[4]. The positivity rate (i.e., percent positive rate or percent positive) is the percentage of all diagnostic tests performed that are positive. The likelihood of disease presence prior to a diagnostic test, or the pretest probability, is dependent on disease prevalence in the population under surveillance. By filtering the broader surveillance population to a subpopulation with a higher likelihood of infection, the allocation and utility of tests can be improved (Fig. 1a). In other words, more positive cases can be captured with the same number of tests and, thus, the testing positivity rate is increased. The development of tools to increase the testing positivity rate is not only crucial in the early phase of an infectious disease outbreak when the available clinical diagnostic testing tools are inadequate to meet the existing demand, but also throughout pandemics in remote locations, underserved communities, and low- and middle-income countries where testing is known to be particularly scarce[17].

The rapid adoption of commercial wearable devices such as smartwatches and activity trackers brings forth opportunities to apply artificial intelligence methods toward the development of novel tools to support an intelligent disease detection infrastructure. Methods such as reinforcement learning or graph neural networks have already been proposed to aid contact tracing and surveillance testing[18,19]. Multiple studies suggest the utility of digital biomarkers, objective and quantifiable digitally collected physiological and behavioral data (e.g., resting heart rate (RHR), step count, sleep duration, and respiratory rate), collected by consumer devices along with patient-reported symptoms to monitor the progression of respiratory and influenza-like illnesses[20–27]. These studies emphasize the utility of wearables data as compared with symptom surveys or known exposure to COVID-19 as a result of its accessibility and scalability.

To determine who to test in settings where there are a limited number of diagnostic tests available (i.e., limited testing capacity), we explored whether information from wearables could help rank individuals by their likelihood of a current COVID-19 infection. To achieve this, we developed an Intelligent Testing Allocation (ITA) model that leverages differences in digital biomarkers to distinguish individuals who are likely positive or negative for COVID-19 in order to improve current methods of diagnostic test allocation and increase testing positivity rates.

RESULTS

We developed the CovIdentify platform in April 2020 to integrate commercial wearable device data and electronic symptom surveys to calculate an individual's real-time risk of being infected with

COVID-19. A total of 7348 participants e-consented to the CovIdentify study between April 2, 2020, and May 25, 2021, through the secure Research Electronic Data Capture (REDCap) system (Fig. 1b)[28]. Of those who consented, 6765 participants enrolled in the study (Supplementary Table 1) by completing an enrollment survey consisting of 37–61 questions that followed branching logic (Supplementary Note 1)[28]. Of those enrolled, 2887 participants connected their smartwatches to the CovIdentify platform, including 1689 Garmin, 1091 Fitbit, and 107 Apple smartwatches. Throughout the course of the study, 362,108 daily surveys were completed by 5859 unique participants, with a mean of 62 and a median of 37 daily surveys completed per individual. Of all CovIdentify participants, 1289 participants reported at least one diagnostic test result for COVID-19 (132 positive and 1157 negative) (Fig. 1b). All survey and device data collected through CovIdentify were transferred securely to a protected cloud environment for further analysis. Out of the 1289 participants with self-reported diagnostic test results, 136 participants (16 positive and 120 negative) had smartwatch data available during the time periods needed for analysis. These 136 participants had 151 ± 165 days of wearable data before the corresponding diagnostic test date. The relatively small number of participants with available smartwatch data out of the larger population is consistent with similar bring-your-own-device studies aimed at COVID-19 infection prediction from personal devices[22,23,27].

Development of the Intelligent Testing Allocation (ITA) model

A diagnostic testing decision support model was designed to leverage real-world data to intelligently allocate diagnostic tests in a surveillance population where there are insufficient tests available to test all people in the surveillance group (Fig. 1a, top). To increase the study population size, we augmented our dataset with data from the MyPHD study. Similar to CovIdentify, MyPHD collected simultaneous smartwatch, symptom, and diagnostic testing data during the COVID-19 pandemic[23,27]. The wearables and diagnostic testing data were publicly available[23,27] while symptom data were added for this work. From the MyPHD study, smartwatch, symptom, and diagnostic testing data from an additional 1129 participants (110 positive and 1019 negative) were included in this analysis, including 53 ± 52 days of wearable data before corresponding diagnostic test dates.

Differences in resting heart rate (RHR) and steps measured by smartwatches well before and immediately prior to a COVID-19 diagnostic test

To compare digital biomarkers between healthy and infected states, data were segmented into two time periods: a baseline period (22–60 days prior to the diagnostic test date) and a detection period (21 days prior to the diagnostic test date). We chose this window for the detection period to encompass the COVID-19 incubation period (2–14 days) reported by the CDC as well as the common delay between symptom onset and diagnostic testing. Consistent with prior literature[20,24], daily RHR increased significantly during the detection period from baseline for those who were COVID-19 positive, with an average difference (±SD) of 1.65 ± 4.63 bpm ($n = 117$, p value <0.001, paired t-test) over the entire time periods. On average, daily RHR values more than two standard deviations from the baseline mean were present as early as 13 days prior to the positive test, with an increasing trend that peaked at 1 day prior to the test date (Fig. 1c, bottom). Conversely, the step count during the detection period decreased significantly from baseline, with a difference of −854 ± 2386 steps/day ($n = 125$, p value <0.0001, paired t-test). On average, step counts less than two standard deviations from the baseline mean were present as early as 10 days prior to the positive test and reached the minimum value 2 days after the test date (Fig. 1c, top). For the subset of participants in our dataset with available symptom onset dates, daily RHR and step count that differed beyond two standard deviations from the baseline mean occurred as early as 5 days before the symptom onset date (Supplementary Fig. 1). Timelines for this and other real-world infection studies should be considered as rough estimates because exact dates of exposure and symptom onset are unknown, unlike in controlled infection studies[26,29]. Our findings, however, are consistent with the 2–14-day COVID-19 incubation period reported by the CDC[30].

There was also a significant difference in digital biomarkers between the baseline and detection periods of participants who tested negative, but it was less pronounced than for those who tested positive. Specifically, the daily RHR difference was 0.58 ± 4.78 bpm ($n = 1094$, p value <0.05, paired t-test) and the step count difference was −281 ± 2013 steps/day ($n = 1136$, p value <0.0001, paired t-test). We hypothesized that the digital biomarker differences in the COVID-19-negative group were because a subset of the negative group may have experienced a health anomaly other than COVID-19 (e.g., influenza) that resulted in physiological differences between the baseline and detection periods. Another recent study also observed RHR elevation and activity reduction in individuals who were COVID-19 negative but flu positive, and the magnitudes of these differences were lower than in individuals who were COVID-19 positive[22]. To explore the possibility that our COVID-19-negative group contains false negatives due to test inaccuracies or physiological differences due to a health anomaly besides COVID-19, we performed hierarchical clustering on the symptom data from individuals who reported negative tests and found a trend toward multiple subgroups (Supplementary Fig. 2). This finding supports the existence of COVID-19-negative subgroups. It should also be noted that the highly significant p value for the digital biomarker differences in the COVID-19-negative group is likely attributable to the higher number of participants (9-fold higher) compared with the COVID-19-positive group.

Cohort definition

For the ITA model development, we only included subjects with sufficient wearable data (≥50% days with a device-specific minimum amount of data availability during periods of sleep for participants with high-frequency wearable data or ≥50% days with device-reported daily values for participants without high-frequency wearable data) in each of the baseline and detection periods. Sleep periods were defined as epochs of inactivity that occurred between midnight and 7 AM on a given day[27]. Consequently, 83 participants from CovIdentify (9 COVID-19 positive and 74 COVID-19 negative) and 437 participants from MyPHD (54 COVID-19 positive and 383 COVID-19 negative) were included in the ITA model development process (Table 1). Of the 63 COVID-19-positive cases, 24 had a clinically documented diagnosis, while the remainder were self-reported. Of the 520 participants with sufficient wearable data, 469 had high-frequency minute-level wearable data (280 from Fitbits) from which we calculated daily RHR and step counts. Device-reported daily values

Table 1. Summary of the cohorts.

Cohort	Total N (Test N)	Total COVID + (test)	Total COVID– (test)
All-Frequency (AF)	520 (105)	63 (13)	457 (92)
All-High-Frequency (AHF)	469 (97)	54 (11)	415 (86)
Fitbit-High-Frequency (FHF)	280 (63)	40 (7)	240 (56)

Total refers to training + test data.

Table 2. Features extracted from the digital biomarkers (DBs) for the development of ITA algorithm.

Metric	Definition	Equation
Deviation metrics		
Delta (Δ)	Deviation in digital biomarker from baseline median value	$DB_{Detection} - DB_{Baseline,\ Median}$
Delta_Normalized	Delta normalized by baseline median value	$((DB_{Detection} - DB_{Baseline,\ Median}) / DB_{Baseline,\ Median})$
Delta_Standardized	Delta standardized by baseline median and interquartile range (IQR)	$((DB_{Detection} - DB_{Baseline,\ Median}) / DB_{Baseline,\ IQR})$
Z-score	Deviation in digital biomarker from baseline mean, standardized by baseline standard deviation (SD)	$((DB_{Detection} - DB_{Baseline,\ Mean}) / DB_{Baseline,\ SD})$
Summary statistics (features)		
Average	Average of interday deviation metrics	
Median	Median of interday deviation metrics	
Maximum	Maximum of interday deviation metrics	
Minimum	Minimum of interday deviation metrics	
Range	Range of interday deviation metrics	

were available for the remaining 51 participants. To explore whether high-frequency wearable data or high-frequency wearable data from a single device type could improve the performance of digital biomarkers for ITA, we developed and validated our ITA model using three cohorts, which we refer to as (1) the All-Frequency (AF) cohort: participants with both high-frequency and device-reported daily values, (2) the All-High-Frequency (AHF) cohort: participants with high-frequency data only, and (3) the Fitbit-High-Frequency (FHF) cohort: participants with high-frequency Fitbit data only (Supplementary Fig. 3 and Supplementary Table 2). We analyzed these three cohorts separately in the subsequent analysis and compared the resulting ITA model performance. We divided each cohort into an 80% train and 20% test split, with FHF as a subset of AHF, which itself is a subset of AF to ensure that no observations in the training set of one cohort existed in the test set of another (Supplementary Fig. 3).

To explore differences in digital biomarkers (median or mean) between the detection and baseline periods that may be useful for the development of ITA model features, we designed four deviation metrics including (1) Δ (detection − baseline), (2) normalized Δ, (3) standardized Δ, and (4) *Z*-score ((detection − baseline mean) / baseline standard deviation) (Table 2). Each of the four deviation metrics was calculated on the training data by digital biomarkers (RHR and step count), day in the detection period, and cohort (examples in Supplementary Figs. 4 and 5), resulting in four calculated metrics per cohort per biomarker. These training data deviation metrics were used as inputs into the subsequent statistical analysis for feature extraction and the ITA model training. We extracted the same resultant features from the independent test set for subsequent ITA model evaluation.

On average, step count decreased (ΔSteps) significantly from baseline to the detection period in COVID-19-positive versus -negative participants (574 vs. 179, 479 vs. 234, and 601 vs. 216 steps per day for the AF, AHF, and FHF training data, respectively; p value <0.05, unpaired t-tests) (Fig. 2a and Supplementary Figs. 6a and 7a, top plots). Conversely, RHR increased (ΔRHR) significantly from baseline to the detection period in COVID-19-positive versus -negative participants (1.8 vs. 0.7, 1.9 vs. 0.8, and 1.8 vs. 0.7 bpm for the AF, AHF, and FHF training data, respectively; p value <0.05, unpaired t-test) (Fig. 2a and Supplementary Figs. 6a and 7a, bottom plots). The 95% confidence intervals of the mean ΔSteps and the mean ΔRHR overlap considerably between positive and negative participants for the initial phase of the detection period (approximately 21–5 days prior to the test date). However, closer to the diagnostic test date (approximately 4–1 day prior to the test date) the 95% confidence intervals of mean ΔSteps largely do not overlap, and the 95% confidence intervals of mean ΔRHR do not overlap at all (Fig. 2a). The fact that the 95% confidence intervals of mean ΔSteps and mean ΔRHR do not overlap later in the detection period is consistent with prior literature[31] and suggests that it is possible to aggregate data into summary statistics to develop a decision boundary that effectively separates COVID-19-positive and -negative cases. However, the overlap in estimated mean values prior to day 5 suggests that separation between positive and negative cases may be more challenging prior to that point in time. Although the 95% confidence intervals closer to the test date were non-overlapping, there was overlap in the variance of the digital biomarkers between the two groups during that time period (Supplementary Fig. 8), which may hinder model performance as separation of the 95% confidence intervals does not necessarily imply significant differences between the groups[32]. Similar estimates of variability have not been reported prior, so we were unable to compare our mean statistics variability to prior literature.

To maximize the separability of the COVID-19-positive and -negative groups in the training set, we performed statistical analysis to explore how different lengths and start times of the detection window, parametrized respectively by two variables (the detection end date, defined by days prior to the diagnostic test date, and the detection window length defined by number of days), would affect the separation between these two groups. We performed a combinatorial analysis across these two parameters (detection end date and detection window length) to calculate five summary statistics (mean, median, maximum, minimum, and range) of the four deviation metrics (Table 2) to be used as features for model building. This resulted in 40 total summary statistics (20 each from steps and RHR), which we refer to as steps and RHR features, respectively. Statistical comparison of the steps and RHR features between the COVID-19-positive and COVID-19-negative groups was performed on the training data for the AF, AHF, and FHF cohorts separately to uncover the statistically significant features (unpaired t-tests; Benjamini–Hochberg corrected p value <0.05).

A systematic grid search to optimize the detection end date and detection window length demonstrated that the closer the detection period is to the diagnostic test date, the larger the number of features that are significantly different between the COVID-19-positive and -negative groups (Fig. 2b and Supplementary Figs. 6b and 7b). Across all evaluated detection end dates, the day prior to the diagnostic test date (detection end date = −1) generated the largest number of significant features for all

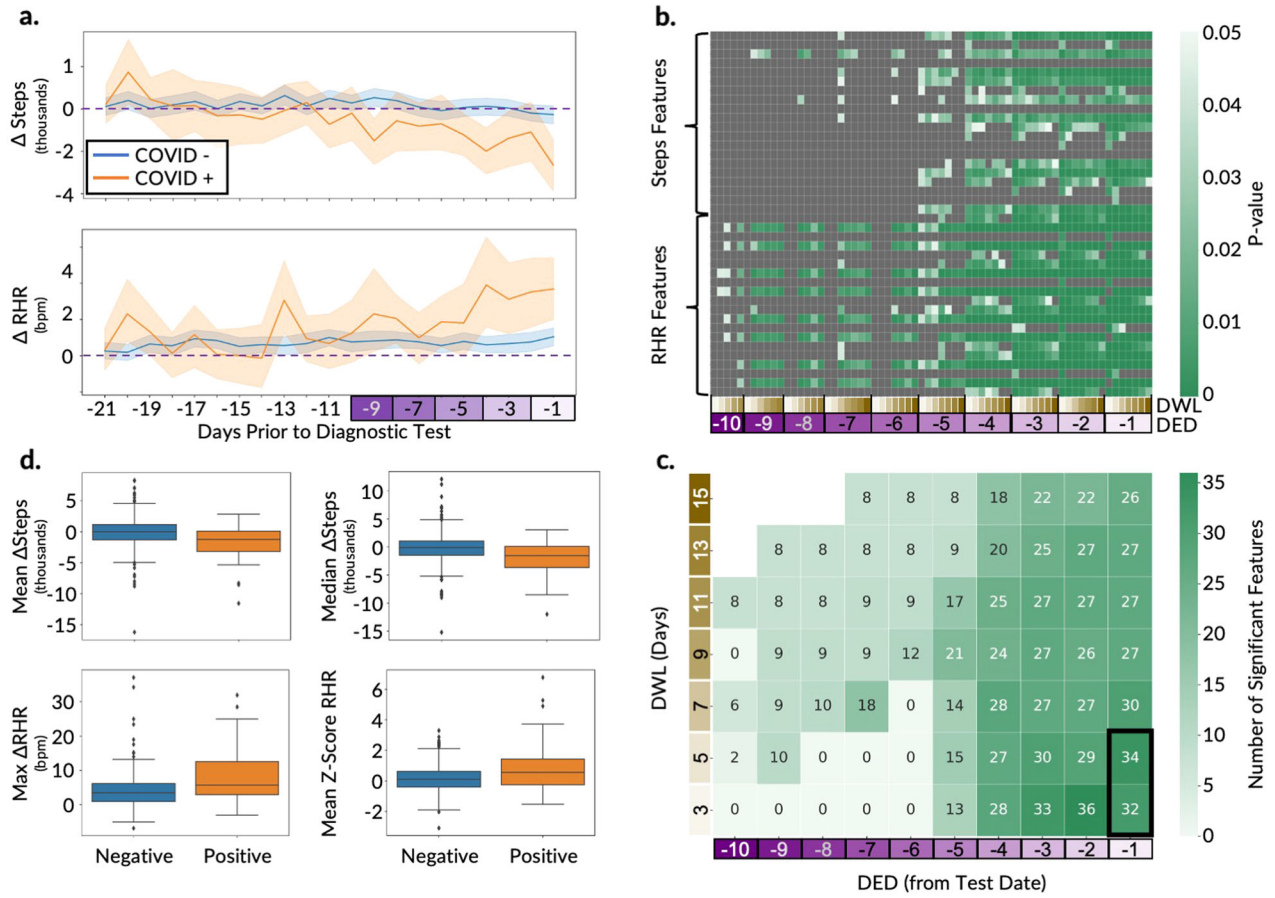

Fig. 2 Overview of digital biomarker exploration and feature engineering for the ITA model development on the AF cohort. a Time-series plot of the deviation in digital biomarkers (ΔSteps and ΔRHR) in the detection window compared to baseline periods, between the participants diagnosed as COVID-19 positive and negative. The horizontal dashed line displays the baseline median and the confidence bounds show the 95% confidence intervals. **b** Heatmaps of steps and RHR features that are statistically significantly different (p value <0.05; unpaired t-tests) in a grid search with a different detection end date (DED) and detection window length (DWL) combinations, with green boxes showing p values <0.05 and gray boxes showing p values ≥0.05. The p values are adjusted with the Benjamini–Hochberg method for multiple hypothesis correction. **c** Summary of the significant features (p value <0.05; unpaired t-tests) from **b**, with each box showing the number of statistically significant features for the different combinations of DED and DWL. The intersection of the significant features across DWL of 3 and 5 days with a common DED of 1 day prior to the test date (as shown using the black rectangle) was used for the ITA model development. **d** Box plots comparing the distribution of the two most significant steps and RHR features between the participants diagnosed as COVID-19 positive and negative. The centerlines denote feature medians, bounds of boxes represent 25th and 75th percentiles, whiskers denote nonoutlier data range and the diamonds denote outlier values.

cohorts. Also, across all cohorts, there were more significant RHR features than steps features (Fig. 2b and Supplementary Figs. 6b and 7b). Additionally, RHR features became significant earlier in the detection period than steps features (detection end date as early as −10 vs. −5 days, respectively), which indicates that changes in RHR occur earlier than steps during the course of infection. Comparison across the three cohorts revealed AF generated the highest number of significant features compared with the AHF and FHF cohorts, which may be attributable to the larger population size of AF. This demonstrates the tradeoff in wearables studies between high-frequency data, which is less common but contains more information, and larger population data, which contains data at a variety of sampling frequencies but overall more data to train the models. Across detection window length values, 3 and 5 days generated the largest number of significant features for all cohorts (Fig. 2c and Supplementary Figs. 6c and 7c), while 5 days also corresponded to the date of the maximum divergence between ΔSteps and ΔRHR (Fig. 2a). Ultimately, this systematic analysis pointed to an optimal detection end date of 1 day prior to the diagnostic test date and an optimal detection window length of 5 days for the detection window duration, both of which were used to generate features for the ITA model.

When implementing the detection end date timepoint and detection window length duration that best separated the COVID-19-positive and -negative groups, there were 28–31 significant features (p value <0.05; unpaired t-tests with Benjamini–Hochberg multiple hypothesis correction) that overlapped across the three cohorts, indicating their robustness to differences in data resolution and device types (Supplementary Table 3). The top 7–9 features, ranked in order of significance, originated from the RHR digital biomarker. To gain a more mechanistic understanding of the RHR and step digital biomarkers, we explored the top two most significantly different (lowest p value) features for each digital biomarker between those who were COVID-19-positive or -negative in the AF cohort (Fig. 2d). The decrease in steps during the detection period as compared to baseline was greater in those with COVID-19, with a 2054 vs. 99 median decrease in steps (median ΔSteps) and a 1775 vs. 64 mean decrease in steps for those who were COVID-19 positive vs. those who were COVID-19 negative, respectively (p values <0.0001; unpaired t-tests with Benjamini–Hochberg multiple hypothesis correction). Conversely,

the increase in maximum deviation in RHR in the detection period as compared to baseline (maximum ΔRHR) and the increase in mean of *Z*-scores in the detection period as compared to baseline (mean of *Z*-score RHR) were both significantly higher for COVID-19-positive participants compared to COVID-19-negative participants (8.4 vs. 4.3 bpm for maximum ΔRHR and 0.9 vs. 0.2 for the mean of *Z*-score-RHR; p values <0.0001; unpaired t-tests with Benjamini–Hochberg multiple hypothesis correction). Consistent across all three cohorts, the median and mean ΔSteps were the most significant (lowest p value) steps features (Supplementary Figs. 6d and 7d). However, the top two RHR features differed, which were median and mean *Z*-score-RHR, and maximum ΔRHR and maximum of normalized ΔRHR for the AHF and FHF cohorts, respectively (Supplementary Figs. 6d and 7d and Supplementary Table 3). The observation of the same top two steps features given the differences in the top two RHR features across the three cohorts may originate from the resolution and device-reported digital biomarkers. For example, the definition of a step and the calculation of the daily step count may be more similar across different device types, while the RHR definition and available HR data resolution may vary more substantially across device types. Although these top features are significantly different between those who are COVID-19 positive and negative, their distributions do overlap, even though the tailedness varies in direction and extent (Fig. 2d and Supplementary Figs. 6d, 7d, and 9), which points to broader challenges surrounding predictive modeling efforts using standard consumer wearable device data for COVID-19 infection detection.

To achieve our broader goal of determining who should receive a diagnostic test under circumstances where there are limited tests available, we aimed to design a model that outputs the probability of a person being infected. However, because our ground truth information is binary (positive or negative for COVID-19), we designed this model as a binary classifier that enabled a straightforward evaluation of its performance. We used the features that were significantly different in the training data between those who were COVID-19 positive and negative (29 features for AF, 28 for AHF, and 31 for FHF) as inputs into five machine learning classification models: logistic regression, k-nearest neighbors, support vector machine, random forest, and extreme gradient boosting (Supplementary Table 4). We chose these five well-established classification models to explore how increasing model complexity and the addition of non-linearity impact the model performance. We trained these classification models on the training data using nested cross-validation (CV) with an inner loop for hyperparameter tuning and an outer loop for model selection. We chose recall as our preferred scoring metric for model selection and evaluation to emphasize the relative impact/cost of false negatives compared to false positives, as an individual who is truly positive for COVID-19 and is wrongly classified as negative (or healthy) would further spread disease.

Following training, we evaluated the performance of the trained model on the independent test set and used two well-established reporting metrics, including the most commonly reported metric for studies of this kind (the area under the curve for the receiver operating characteristic curve (AUC-ROC))[24,33–37], and the metric that is most appropriate for this classification task (AUC for the precision-recall curve (AUC-PR))[38] (Supplementary Table 3, Figs. 3 and 4, and Supplementary Fig. 10). AUC-PR is more appropriate with class-imbalanced data[38,39], which is the case here (12–15% COVID-19 positive and 85–88% negative in each of the three cohorts). The class imbalance in our dataset was not correctable through resampling methods—we have observed that distributions of features overlap between the COVID-19-positive and -negative participants, as demonstrated in the individual feature comparison (Fig. 2d and Supplementary Figs. 6d and 7d), as well as in the low dimensional representation (using principal component analysis and t-stochastic neighbor embedding) of all the features in the training set of the AF cohort (Supplementary Fig. 11).

Of the five models tested, logistic regression outperformed all other models based on the training AUC-PR for all three cohorts and was also the best performing model based on the training AUC-ROC for the AF and FHF cohorts. The superior performance of the logistic regression among other (more complex and nonlinear) models may be attributed to the tendency of more complex and nonlinear models to overfit the training data[40], which comes to light with our CV methods. The superior performance of the logistic regression also points to the potential to develop explainable machine learning predictive models for the ITA model that enables rapid translation from bench to bedside. Overall, the classifier performed best in the FHF cohort (Supplementary Table 3, Fig. 3c, f, and Supplementary Fig. 10c, f), followed by the AHF cohort, (Fig. 3b, e and Supplementary Fig. 10b, e) and finally the AF cohort (Fig. 3a, d and Supplementary Fig. 10a, d). These performance differences indicate that device-related and data resolution differences may confound disease-related physiological differences captured by digital biomarkers. Therefore, building models using a single device type and with higher resolution data improves performance. For the FHF cohort, the logistic regression model resulted in an AUC-ROC of 0.73 ± 0.12 and AUC-PR of 0.55 ± 0.21 on the cross-validated training set (Fig. 3c, f), and AUC-ROC of 0.77 and AUC-PR of 0.24 on the test set (Supplementary Fig. 10c, f). The AUC-ROC from the models were similar to those reported in recent similar studies[24,34,37].

However, the performance of the models based only on AUC-ROC in the context of imbalanced data can be misleading, as a large change in the number of false positives may have a small effect on the false-positive rate[39]. The precision metric, which integrates both true positives and false positives, can mitigate the effect of an imbalanced dataset (e.g., the higher proportion of negatives seen in this type of data) on a model's performance. Our precision-recall analysis (Fig. 3d–f and Supplementary Fig. 10d–f) demonstrates that we can improve the recall (minimizing false negatives) at the expense of precision. In an extreme example, we were able to achieve 100% recall with a precision of 0.4 on the cross-validated training set of the FHF cohort, whereas, a dummy classifier with random chance (i.e., Random Testing Allocation (RTA)) can achieve a precision of 0.15 on this dataset. It is also important to note that we are not considering resource-limited settings in the ROC and PR analysis; instead, it is assumed that there are a sufficient number of diagnostic tests available for the entire surveillance group. In a resource-limited setting, 100% recall may not be achievable due to the shortage of diagnostic testing.

To understand the relative contribution of the steps and RHR digital biomarkers to the ITA model performance, we developed two separate sets of models using features based only on either steps or RHR using the training set data with logistic regression, and later validated on the test set. Consistent with previous literature[24,34] the models using steps-based features alone had a higher AUC-ROC than models using RHR-based features alone (cross-validated AUC-ROC of 0.67 vs. 0.64, 0.69 vs. 0.63, and 0.72 vs. 0.68 for steps vs. RHR features for the AF, AHF, and FHF training sets, respectively) (Fig. 3). Interestingly, when using the AUC-PR as the performance metric, models using features based on RHR digital biomarkers outperformed models using features based on steps digital biomarkers, a finding that has not been previously reported (cross-validated AUC-PR of 0.30 vs. 0.38, 0.28 vs. 0.37, and 0.40 vs. 0.49 for steps and RHR features for the AF, AHF, and FHF training datasets, respectively) (Fig. 3). The validation on the test sets also demonstrated similar results (AUC-ROC of 0.61 vs. 0.60, 0.66 vs. 0.58, and 0.71 vs. 0.70 and AUC-PR of 0.16 vs. 0.18, 0.17 vs. 0.17, and 0.18 vs. 0.22 for steps vs. RHR features for the AF, AHF, and FHF test sets, respectively) (Supplementary Fig. 10). Overall, the addition of steps features increased the AUC-ROC of the ITA

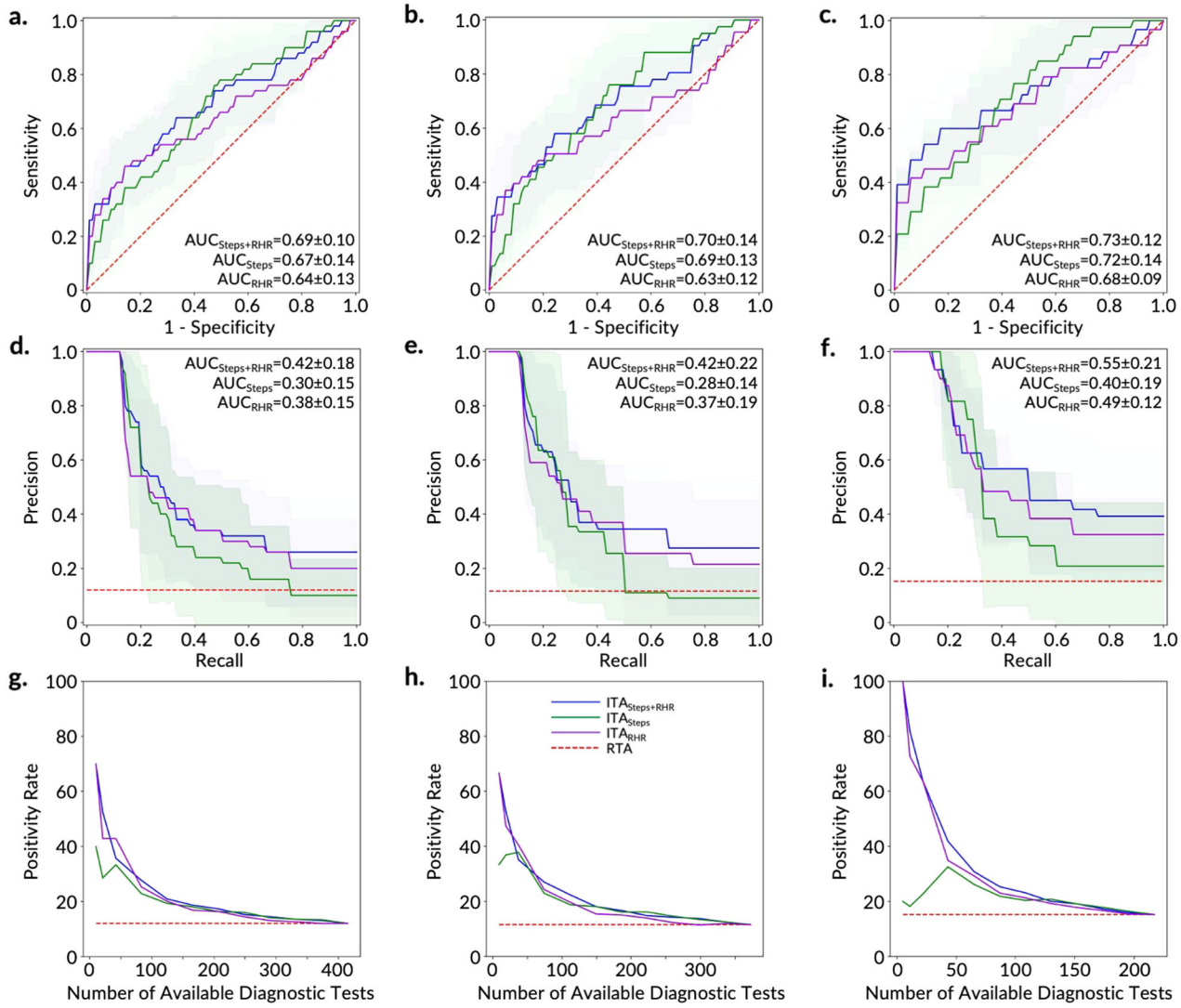

Fig. 3 Prediction and ranking results of the ITA models on the training sets for the AF (a, d, and g), AHF (b, e, and h), and FHF (c, f, and i) cohorts using features from a combination of Steps and RHR (blue), Steps (green), and RHR (violet) digital biomarkers. a–c Receiver operating characteristics curves (ROCs) and **d–f** precision-recall curves (PRCs) for the discrimination between COVID-19-positive participants and -negative participants in the training set. The light blue, light green, and light violet areas show one standard deviation from the mean of the ROCs/PRCs generated from 10-fold nested cross-validation on the training set and the red dashed line shows the results based on a Random Testing Allocation (RTA) model (the null model). **g–i** The positivity rate of the diagnostic testing subpopulation as determined by ITA given a specific number of available diagnostic tests. The red dashed line displays the positivity rate/pretest probability of an RTA (null) model.

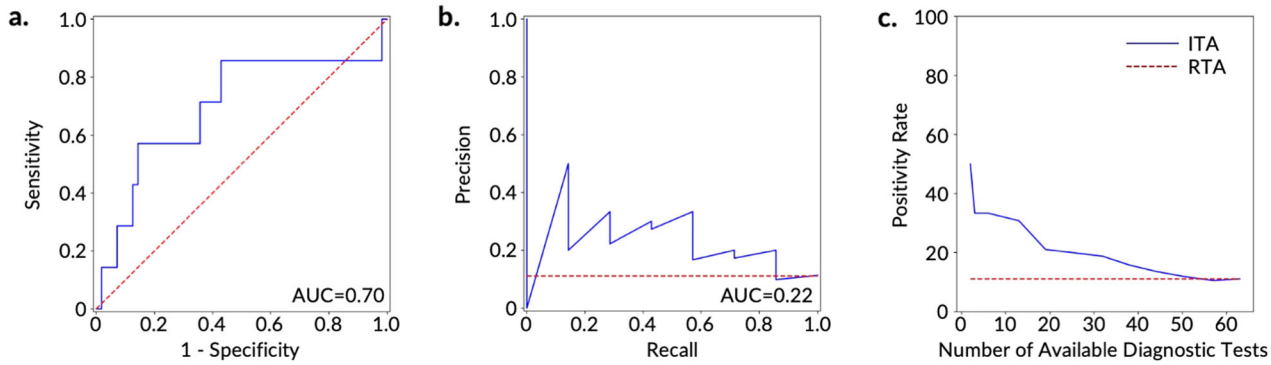

Fig. 4 Prediction and ranking results of the ITA models on the test set of the FHF cohort using RHR digital biomarkers. a ROC and **b** PRC for the discrimination between COVID-19-positive participants ($n = 7$) and -negative participants ($n = 56$). The red dashed line shows the results based on an RTA model. **c** Positivity rate of the diagnostic testing subpopulation as determined by ITA given a specific number of available diagnostic tests. The red dashed line shows the positivity rate of an RTA (null) model.

model by 7–11% compared with RHR features alone, while RHR features improved the AUC-PR of the ITA model by 38–50% compared with steps features alone on the training set. In other words, the exclusion of each steps and RHR features individually decreased the AUC-ROC of the ITA model by 7–10% and 1–3% for the training set (5–11% and 2–9% for the test set), respectively, compared to the ITA model with both steps and RHR features (Fig. 3a–f and Supplementary Fig. 10a–f). On the other hand, the exclusion of each steps and RHR features individually decreased the AUC-PR of the ITA model by 10–12% and 19–27% for the training set (5–15% and 5–25% for the test set) compared to the ITA model with both steps and RHR features. These results suggest that, while steps features provide more salient information on the tradeoff between the true-positive rate and false-positive rate, RHR features provide more salient information on the tradeoff between the true-positive rate and the precision (positive predictive value). In other words, while steps features improved the specificity of the predictive model, RHR features improved the precision.

In addition to comparing the performance of ITA models with steps and RHR features alone to ITA models with both steps and RHR features on both training and test set, we also compared the relative feature importance in the logistic regression model using both steps and RHR features on the training set. Our results demonstrated that two, one, and four of the top five features originated from RHR in the AF, AHF, and FHF cohorts, respectively, with the remaining features originating from steps (Supplementary Fig. 12). In all three cohorts, median ΔSteps and mean ΔSteps were the two most important steps features, which was consistent with our earlier statistical analysis. Maximum ΔRHR was the most important RHR feature for the AF and AHF cohorts and the second most important RHR feature for the FHF cohort, and was also one of the top two most significant features in our earlier statistical analysis for the AF and FHF cohorts.

Improvement in positivity rate for COVID-19 diagnostic testing using the ITA method

We next evaluated how the ITA model can improve the current standard of practice for COVID-19 infection surveillance. Under current surveillance testing methods in the US, while some tests are taken due to symptoms or possible exposure, many are taken as precautionary measures for traveling or for surveillance in schools and workplaces[30]. While such forms of widespread RTA surveillance are beneficial, the positivity rate of widespread diagnostic testing is typically low and, thus, requires sufficient testing capacity in order to prevent testing shortages (e.g., sold out at-home testing kits). Applying an equivalent RTA surveillance approach to our study population results in a 12% positivity rate in both our AF-training (50 COVID-19-positive participants out of 365 participants in total) and AF-test (13 COVID-19-positive participants out of 92 participants in total) datasets. It is important to note that the 12% positivity rate is consistent for all levels of diagnostic testing capacity (0–100% of population). When employing ITA with both steps and RHR features, and adding the constraint of limited diagnostic testing capacity (10–30% of population), the testing positivity rate of the cross-validated model increased 2–3 fold (21–36% positivity rate) for the training dataset (Fig. 3g) and 1.5–2.5 fold (19–29% positivity rate) for the testing dataset (Supplementary Fig. 10g).

A comparison of the three cohorts demonstrated that the best performing ITA model with both steps and RHR features stemmed from the FHF cohort and was followed by the AHF cohort (Fig. 3h, i and Supplementary Fig. 10h, i). By utilizing ITA and assuming a diagnostic testing capacity at 10–30% of the population, the positivity rate of the FHF and AHF cross-validated training datasets increased by 4 fold (64% positivity rate) and 3 fold (35% positivity rate) when compared to the RTA positivity rates of 15% and 12% for FHF and AHF cohorts, respectively. For the FHF cohort, the positivity rate further increased up to 6.5 fold (100% positivity rate) in the cross-validated training dataset when the diagnostic testing capacity was reduced to 2.5–5% of the population (5–11 diagnostic tests to be allocated to individuals in the training dataset) (Fig. 3i). Using the independent test dataset with both steps and RHR features, the positivity rate of the FHF and AHF cohorts increased by 1.5–3 fold (17–31% positivity rate) and 2–3 fold (21–32% positivity rate), respectively, compared to the RTA positivity rate of 11%, when the diagnostic testing capacity was 10–30% of the population. These results indicate the potential of the ITA model to target diagnostic testing resources toward individuals who have a higher likelihood of testing positive (i.e., increasing the positivity rate of diagnostic testing) and enable more efficient allocation of testing capacity. When we compared the ITA model performance in terms of improving the positivity rate of the diagnostic testing in a resource-limited setting among models with steps and RHR features separately and together, the results demonstrated that ITA models using only RHR features often achieved similar performance (similar positivity rate) on the training set and similar and in some cases even better performance (further improved positivity rate) on the test set in comparison with the models that used both steps and RHR features together (Fig. 3g–i and Supplementary Fig. 6g–i). For example, the ITA model using only RHR features improved the positivity rate up to 4.5 fold (positivity rate of 50%) compared to the RTA positivity rate of 11% on the test set of FHF cohort (Supplementary Fig. 10i). The superior performance of the ITA model using RHR-only features over the ITA model using steps-only and the ITA model using both steps and RHR features may be attributed to the nonspecific nature of the steps features, which can experience changes unrelated to COVID-19 (other diseases, quarantine, stress, etc.). These results demonstrate the potential to develop an ITA system to allocate diagnostic testing in limited resource settings only using physiological digital biomarkers without relying on potentially nonspecific activity digital biomarkers, which is a key finding from our work.

We further explored how the ITA model performs in symptomatic versus asymptomatic COVID-19-positive individuals in each cohort. We considered participants to be symptomatic who reported any symptoms in the detection period or on the diagnostic test date. Assuming a diagnostic testing capacity of 30%, ITA indicates testing for 19 of 29 symptomatic and 7 of 21 asymptomatic COVID-19-positive individuals in the cross-validated model using both steps and RHR features, and 5 of 8 symptomatic and 1 of 5 asymptomatic COVID-19-positive individuals in the independent test set of the AF cohort. In other words, 7 of 26 (27%) and 1 of 6 (17%) COVID-19-positive individuals were asymptomatic in the ITA-determined subpopulation for the cross-validated training set and an independent test set of the AF cohort, respectively. Results were similar for the AHF and FHF cohorts (Supplementary Table 5). These findings indicate that the ITA model can not only target diagnostic testing resources toward individuals with symptoms, but also those without any reported symptoms, further increasing the utility of this method.

DISCUSSION

The COVID-19 pandemic revealed the fragility of our existing healthcare infrastructure to detect the virus and prevent its spread. One key tool for reducing disease spread is bringing diagnostic testing to the right people at the right time and ensuring appropriate interpretation of the diagnostic testing results based on the prevalence of the disease in the population[4]. In light of this need, in April 2020 we developed CovIdentify to integrate commercial wearable device data and electronic symptom surveys to assess the real-time risk of being infected with COVID-19. We envisioned two possible scenarios where

CovIdentify would be useful for informing intelligent testing decisions, including (1) ranking individuals in a group by likelihood of current infection with COVID-19 to determine who to test, and (2) tracking a single individual over time for evidence of new infection onset to determine when to test. In our initial development of the ITA model, we focused on the first question, and ultimately improved the positivity rate of COVID-19 diagnostic testing up to 6.5 fold when compared against RTA. These results indicate that if deployed on a large scale, the ITA model could potentially be used to better allocate diagnostic testing resources. To test the real-world efficacy of the ITA model, a simple approach may be to compare the positivity rate of ITA recommended diagnostic testing versus traditional surveillance testing in cohorts of school teachers in the same jurisdiction or school (i.e., similar prevalence rate). This method is likely applicable to other diagnostic areas as well, where digital biomarkers can be used to indicate the likelihood of disease.

In this work, we demonstrated that wearable device data can be used to strategically target the allocation of diagnostic tests to where they are most useful. This approach not only increases testing efficiency and allocation but also reduces the costs and supply chain burden of surveillance testing which is an ongoing challenge. Our results further demonstrate that the ITA method is able to filter a surveillance population to generate a subpopulation with a higher density of true positives, regardless of the prevalence and pretest probability of COVID-19 infection in the population under surveillance for the disease, and, thus, increases testing positivity rates. We note that, although there is a possibility that our COVID-19-negative group may contain other illnesses (e.g., flu) which also reflects a more realistic setting, we were still able to improve resource allocation by over 450% in the independent test set. Another key contribution of our work is the utility of the ITA model using only physiological digital biomarkers (RHR). As steps (and other physical activity) may be reduced due to other reasons than COVID-19 infection, steps may result in nonspecific models, as we have observed from our results on the independent test set. For that reason, an ITA model using more specific digital biomarkers (e.g., RHR) demonstrates the potential of solely relying on physiological data from wearables to develop such an ITA model in a resource-limited setting. We also demonstrate the utility of the ITA to filter individuals for allocating diagnostic tests not only in cases of symptomatic individuals but also for asymptomatic individuals who may not be tested and diagnosed otherwise. While the sensitivity and specificity of diagnostic tests are not affected by ITA, this more efficient testing allocation approach identifies more cases in less time and with fewer resources[41–44].

The basis of the ITA method is the detection of physiological changes associated with infection onset, which are well established to be detectable by biometric sensors[20–24,26,34–37]. Consistent with prior literature, we demonstrate here that digital biomarkers derived from heart rate and physical activity are indicative of infection onset. A unique contribution of our work is the demonstration of differences in digital biomarker significance with respect to time prior to the diagnostic test date; specifically, we show that differences in RHR features were significant between COVID-19-positive and -negative groups as early as 10 days prior to the diagnostic test date whereas differences in most steps features were not significant until 5 days prior to the diagnostic test date. One steps feature, minimum ΔSteps, was significant up to 9 days prior to the diagnostic test date, potentially demonstrating a link between activity levels (and perhaps noncompliance with lockdown measures) and COVID-19 exposure. Furthermore, RHR begins to deviate from baseline earlier than steps (as early as 13 vs. 10 days prior to the diagnostic test date, respectively), and the peak effect (maximum deviation from baseline) of infection also occurs earlier in RHR than steps (1 day prior vs. 2 days after the diagnostic test date, respectively) for those who were COVID-19 positive. These results indicate that changes in physiology (RHR) occur earlier in the infection period, while symptoms and reduced physical activity (steps) transpire later in the infection period, when people may limit their movement due either to illness or mandatory quarantine. A recent COVID-19 study assessing prolonged physiological and behavioral changes using wearables also observed that COVID-19-positive individuals took more time to return to their RHR baseline values compared to their step and sleep baseline values following the acute COVID-19 infection period[31]; however, this work explored the post-infection period of the data whereas here we explore the pre-infection period as well as the acute infection period using a systematic grid search approach. Another recent study[34] that developed machine learning models to passively detect COVID-19 using wearable data noted relative changes in feature importance when including data post-diagnosis. However, to our knowledge, we are the first to demonstrate and establish the dynamics of feature importance over time prior to the diagnostic test date, indicating which features should be weighted more heavily in prediction models and when.

Another important contribution of our work is demonstrating the utility of RHR and steps features in the tradeoff between the true-positive rate and false-positive rate (ROC analysis) and the tradeoff between the true-positive rate and the positive predictive value (PR analysis). Specifically, we show that while steps features provide more salient information on the tradeoff between the true-positive rate and false-positive rate, RHR features provide more salient information on the tradeoff between the true-positive rate and the precision (positive predictive value). To our knowledge, this is the first demonstration of this tradeoff in predictive model development for COVID-19 infection detection. The ITA model, in addition to using features of RHR and steps, can likely be further extended and improved with features from other digital biomarkers such as skin temperature, respiratory rate, blood oxygen saturation, and sleep duration[25,26,35,36]. It is anticipated that each of these distinct digital biomarkers would capture a physiological response to infection at different times during the detection period, thus improving the robustness and overall performance of the ITA approach.

One of the important observations from our work was the clear separation of the 95% confidence intervals of the means of digital biomarkers between COVID-19-positive and -negative populations as early as 5 days prior to the test date (Fig. 2a and Supplementary Figs. 6a and 7a), while the variances of the groups have overlapping distributions in the same time window (Supplementary Fig. 8). Notably, a lack of overlap in 95% confidence intervals does not necessarily imply significant differences between the groups[32] as standard deviation is a valuable descriptive measure of the data that should be considered as well. There are many possible sources of variance in studies involving wearable data, including the inclusion of different device types and technologies, contexts of measurement (e.g., time of day, activity type, etc.), differences in physiological response to infection, etc. We mitigated this issue by segmenting by device type and data resolution, as well as by utilizing measurements during resting periods only for the RHR calculation. In the future, larger datasets can enable segmentation by demographics (e.g., age, sex, weight, etc.) that would likely further reduce the variance. Sharing datasets between studies, as demonstrated here, can also augment the study population and further reduce the variance. An open question is whether the resolution of current photoplethysmography-based wearable heart rate technologies is high enough to adequately detect signals above the population variance.

Here, we did not deploy the ITA method in real-time and, thus, its performance in practice still remains to be tested. Both the CovIdentify and MyPHD studies were primarily bring-your-own-device study designs, in which people who already own smart

devices are recruited to participate. The bring-your-own-device design presents two major challenges: (1) participants must own a smart device, which limits eligibility to those who can afford devices, and (2) many different types of devices are used, introducing an additional source of noise in the analysis. We mitigated the first challenge by developing and implementing the Demographic Improvement Guideline, which resulted in a 250% increase in the representation of black and African American participants and a 49% increase in the Latinx and Hispanic population within 4 months of the implementation of the guideline[45]. The second challenge by dividing our overall dataset into cohorts with homogeneous sampling frequencies and/or device types. Although we recognize that certain factors decrease the likelihood of wearable device ownership, such as lower income or living in a rural area[46–48], the precipitously decreasing cost of wearable technology is rapidly increasing the equitable distribution of these technologies[49].

Another limitation of the study is the data missingness and its impact on the deviation of the digital biomarkers, as the source of missingness may confound the disease-related physiological variation. For example, we observed that some participants in our study did not wear their devices when they were feeling sick and/or during sleep, as observed in other studies[23], which resulted in a reduction in data availability as a result of our rigorous data inclusion criteria. For that reason, it can be a challenge to isolate the effects of physiological and behavioral changes on the digital biomarkers. Furthermore, some devices require more frequent charging (e.g., Apple Watch), which results in more missing data that may also impact model performance. We mitigated this challenge by further developing our model on a single device and homogeneous sampling frequency (FHF) cohort.

Another limitation of the study is the self-reported diagnostic testing results from the majority of our study participants. While we acknowledge that self-reported COVID-19 testing results can be less reliable than clinically documented results, similar COVID-19 digital health studies[23,24,27] utilized self-reported diagnostic testing results for their algorithm development. To instill further confidence in this approach, it is worth noting that if any inaccuracies do exist in the reported testing, which is to be expected in a real-world setting where inaccurate diagnostic testing can occur regularly, our study population was sufficiently large to be powered to handle such noise and variance as demonstrated by the strength of the results.

The recent body of work on COVID-19 detection using smartwatches uses AUC-ROC to evaluate model performance[24,34–37], which is only an appropriate metric for class-balanced data, and is otherwise misleading[38,39]. In these large-scale studies conducted on a convenience sample of the population for a disease with low prevalence, there exists an inherent challenge of class imbalance because most of the study population does not contract the disease. This was a challenge that we faced in our study, and, further complicating matters, many of the COVID-19-positive participants did not wear their wearable devices at the start of their infection, exacerbating the class imbalance. While less frequently reported than AUC-ROC, the AUC-PR is the correct evaluation metric for evaluating a classifier on imbalanced data[38], which is what we report here. We show that even with a strong AUC-ROC, the AUC-PR demonstrates the limitations of performance. Methods to resolve class imbalance, especially when working with wearable device data, can be further investigated for future studies. Furthermore, more advanced artificial intelligence methods such as reinforcement learning or graph neural networks may further enhance the performance of the ITA model and is a topic that will be further explored in future studies.

While our study focused on improving testing allocation for COVID-19, the methods developed herein are extensible to other types of infections and could be used to fortify our future pandemic preparedness. Using ITA to improve disease surveillance could be especially important in underserved communities that may benefit from the fact that the ITA method is useful even with only steps digital biomarkers which may be obtained from smartphones which are owned by 85% of the population in the US[50] and up to 76% globally[51]. By targeting diagnostic testing toward individuals who are more likely to truly be infected with a disease, we can improve the allocation and utility of diagnostic tests, ultimately reducing mortality and increasing our ability to control current and future pandemics.

METHODS

Participant recruitment and data collection

The CovIdentify study launched on April 2, 2020 (Duke University Institutional Review Board #2020-0412). Eligibility criteria included age over 18 years and internet access. Social networks and social media advertising were used to recruit participants. By May 25, 2021, a total of 7348 participants were recruited and e-consented through the REDCap system[28]. During enrollment, participants were given the option to donate 12 months of retrospective wearable data and 12 months of prospective wearable data. Wearable data was collected via the CovIdentify iOS app for devices connected to the Apple Health kit (e.g., Apple Watch) or via Application Programming Interfaces for other devices (e.g., Garmin and Fitbit devices). The participants were also asked to complete an onboarding (enrollment) survey and daily surveys. The surveys were in English or Spanish and included questions on symptoms, social distancing, diagnostic testing results, and related information (Supplementary Note 1). Surveys were collected using the CovIdentify iOS app, text messaging, and/or emails. All wearable data and survey results were stored in a secured Microsoft Azure data platform and later analyzed in the Microsoft Azure Machine Learning environment. Soon after CovIdentify was launched, exploratory data analysis (EDA) revealed major differences between CovIdentify demographics and the demographics of COVID-19-positive cases and deaths in the U.S., as well as overall U.S. demographics based on the 2020 U.S. Census[52,53]. We sought to mitigate the imbalance throughout the duration of the study by providing wearable devices to under-represented populations[45]. COVID-19 vaccine reporting was added to the daily surveys in February 2021, where we asked questions regarding the vaccination date, vaccine brand, vaccine-related symptoms, and dose number.

Wearable data processing and analysis

Participants were asked to fill out an enrollment survey following the informed e-consent. Daily symptom surveys and wearable data from the participants were analyzed both separately and together. For the overall analysis, we only included participants with self-reported diagnostic test results for COVID-19. These participants were further divided into two categories based on the self-reported diagnostic test results: COVID-19 positive and COVID-19 negative.

In addition to the data collected via CovIdentify, we augmented our analysis by including data from the MyPHD study, as reported in the two recent publications by Mishra et al.[23] and Alavi et al.[27]. The data from Mishra et al. included heart rate, step count, and sleep data for 27 COVID-19-positive cases. It also included metadata of symptom onset and test dates. The data from Alavi et al. included heart rate and step count data for 83 COVID-19-positive cases and 1019 COVID-19-negative cases as well as metadata including symptom onset and test dates.

For wearable data analysis, we only included days of wearable data when both heart rate and step count were available. Out of the 1239 participants (113 from CovIdentify and 1126 from MyPhD study) who had both heart rate and step count data available, we had device-reported daily values of RHR and step count for 67 participants, and high-frequency (second or minute level, depending on device types) wearable data for 1172 participants. For participants with high-frequency heart rate data, we calculated daily RHR from the heart rate data points recorded between midnight and 7 AM, when there were no steps recorded. For those participants with available high-frequency wearable data, we chose a data-driven threshold (i.e., a minimum number of heart rate data points between midnight and 7 AM with zero recorded steps) to include our calculated RHR data from that day in the subsequent analysis. As the sampling rate varies by device type (Fitbit, Garmin, and Apple Watch), we generated separate data distributions of the datasets for these three

device types and selected the first quartile of heart rate data points per device as the data-driven threshold, which resulted in a threshold of 2630, 19, and 1389 heart rate data points for Fitbit, Apple Watch, and Garmin devices, respectively. In other words, on a given day, a participant with Fitbit wearable data required at least 2630 heart rate data points between midnight and 7 AM with zero recorded steps to be included in the subsequent analysis. Following this intraday data point threshold, we used an interday data threshold: a minimum number of days with available wearable data to be included in the analysis (50% in the baseline period and 50% between 9 days and 1 day prior to the diagnostic test date in the detection period). We explored different minimum number of days of available wearable data in the baseline and detection periods and selected these two thresholds to maximize the number of participants while keeping the performance of the ITA model on the training dataset consistent, defined as less than 10% variation of the performance metrics (AUC-ROC and AUC-PR)).

Cohort definition

The wearable data availability thresholds (both intraday and interday) resulted in an AF cohort of 520 participants (83 from CovIdentify and 437 from MyPHD) with sufficient wearable data (63 COVID-19 positive and 457 COVID-19 negative). 24 of the 63 COVID-19 positive cases had clinical documentation for their diagnosis while the others were self reported. We then created two more subsets from this cohort (Supplementary Fig. 3): (1) AHF cohort: participants with high-frequency wearable data (469 participants, 54 COVID-19 positive and 415 COVID-19 negative), and (2) FHF cohort: participants with high-frequency wearable data from a single source (Fitbit) (280 participants, 40 COVID-19 positive and 240 COVID-19 negative) to explore the impact of utilizing wearable data from different sources and resolutions on the ITA model development. We employed these three cohorts separately for the ITA model development and compared the resulting models' performance in the corresponding training and test datasets of these cohorts. We divided each cohort into an 80% train and 20% test split, with FHF as a subgroup of AHF (which itself is a subset of AF) to ensure that no observations in the training dataset of one cohort existed in the test dataset of another (Supplementary Fig. 3).

Digital biomarker definition

Given the use of datasets with different device types, a consistent RHR definition was used in order to harmonize the cohorts with high-frequency wearable data. We calculated the daily RHR digital biomarker by aggregating the high-frequency heart rate data points available between midnight and 7 AM, when there were no steps recorded. Step count was calculated by summing all recorded step values during a 24-h period in order to produce a daily step count digital biomarker.

Feature engineering and extraction

Following the creation of three cohorts (AF, AHF, and FHF) and their corresponding training and test sets, we performed EDA and extracted features from the time-series digital biomarkers (RHR and step count). For the EDA on the time-series digital biomarkers, we explored the difference in trajectories of digital biomarkers between COVID-19-positive and COVID-19-negative participants (Fig. 2a and Supplementary Figs. 6a and 7a). Following the EDA, we extracted the features mentioned in Table 2 from the raw digital biomarkers. We first calculated four deviation metrics, which capture the deviation in digital biomarkers from participants' baseline during the detection phase. Following the deviation metrics calculation, we calculated summary statistics of these four deviation metrics which we refer as to features for this manuscript. We extracted the same features from the training and test datasets. Following the feature extraction, we performed statistical analysis on the features from the training datasets of the three cohorts to see which features are statistically different between the two groups and how their significance levels vary with different detection period combinations (detection end date and detection window length) using a systematic grid search to optimize detection end date and detection window length (Fig. 2b and Supplementary Figs. 6b and 7b). We utilized multiple hypothesis testing with Benjamini–Hochberg adjusted p values for this statistical analysis. Following the statistical analysis and systematic grid search to obtain the optimal detection period to extract the features, we only utilized the intersection of the statistically significant features (p value <0.05; unpaired t-tests with Benjamini–Hochberg multiple hypothesis correction) extracted from digital biomarkers recorded between 5 days and 1 day and 3 days and 1 day prior to the diagnostic test date for the development of the ITA model.

ITA model development

Following feature extraction, we developed predictive models to classify COVID-19-positive and -negative participants in the training dataset of each cohort (AF, AHF, and FHF) using nested CV and later validated the models on corresponding independent test datasets. We chose five state-of-the-art machine learning models (logistic regression, K-nearest neighbor, support vector machine, random forest, and extreme gradient boosting[54,55]) for the development of the ITA models to explore how increasing model complexity and adding non-linearity would impact the model performance. We trained these classification models on the training dataset using nested CV with an inner CV loop for hyperparameter tuning and an outer CV loop for model selection. For model training, we selected recall as our preferred scoring metric for model selection to emphasize the relative impact/cost of false negatives compared to false positives, as an individual who is truly positive for COVID-19 and is wrongly classified as negative (or healthy) would further spread disease. For model performance evaluation, we used two well-established reporting metrics, including the most commonly reported metric for studies of this kind (AUC-ROC)[24,33–37], and the metric that is most appropriate for this classification task (AUC-PR)[38] (Supplementary Table 3, Figs. 3 and 4, and Supplementary Fig. 10). AUC-PR is more appropriate with class-imbalanced data[38,39], which is the case here (12–15% COVID-19 positive and 85–88% negative for each of the three cohorts). The results reported for the training dataset (Supplementary Table 3 and Fig. 3a–f) were generated from the validation on the held-out dataset (fold) from each iteration of the outer CV loop which was not used in the model training. Based on the CV results of the five machine learning models on the training dataset, we chose the logistic regression model to further evaluate performance on the independent testing dataset (Supplementary Fig. 10a–f). For validation on the independent test dataset, we trained the logistic regression model on the entire training dataset using a grid search with five stratified folds for hyperparameter tuning and selected the best model (with tuned hyperparameters) to validate on the test dataset.

Nested cross-validation

For model development with the training dataset, we utilized nested CV over traditional CV, which is a common approach in similar studies[24,34,36,37], because it uses the same data for hyperparameter tuning and model performance evaluation[56]. In nested CV (also called double CV), the hyperparameter tuning procedure is nested (inner loop) under the model selection procedure (outer loop) and the inner loop is used for optimizing the hyperparameters of the model with inner CV, and the outer loop is used to compute the error of the optimized model with outer CV[57]. For the nested CV, we divided the training set into ten stratified folds (keeping the ratio of COVID-19-positive and -negative participants the same across each fold) for the outer loop. For each iteration of the outer loop, the model was trained on data from nine folds by optimizing the hyperparameters of the model with inner CV, and validating on the left-out fold, a process which was repeated nine more times. In each iteration of the outer loop, the outer training data (from nine folds) were further divided into five stratified folds (inner loop) to tune hyperparameters using a grid search. Out of the five iterations with the grid search in the inner loop, the best model (including hyperparameters) was selected, and this model was used in the model performance evaluation in the outer loop. This way of model development using two CV steps separates hyperparameter tuning and model selection in order to reduce bias in model performance.

Feature importance ranking

To calculate the feature importance ranking, we trained the logistic regression model using a grid search with five stratified folds for hyperparameter tuning and selected the best model (with optimized hyperparameters) to train on the entire training set of each cohort, and extracted the coefficients for each feature used in the optimized model. We reported the absolute value of each coefficient as the relative importance of the features (Supplementary Fig. 12).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The de-identified CovIdentify dataset generated and/or analyzed during the current study will be submitted 1 year from the publication date to the Digital Health Data Repository (DHDR) repository (https://github.com/DigitalBiomarkerDiscoveryPipeline/Digital_Health_Data_Repository) under the title BigIdeasLab_CovIdentify. The de-identified MyPHD dataset used in Alavi et al. (*Nature Medicine* 2021) study can be downloaded at the following publicly available link: https://storage.googleapis.com/gbsc-gcp-project-ipop_public/COVID-19/COVID-19-Wearables.zip and the dataset used in Mishra et al. (*Nature Biomedical Engineering* 2020) study can be downloaded at the following publicly available link: https://storage.googleapis.com/gbsc-gcp-project-ipop_public/COVID-19-Phase2/COVID-19-Phase2-Wearables.zip.

CODE AVAILABILITY

ITA model development code used for this manuscript is available on the digital biomarker discovery pipeline (DBDP) GitHub repository (https://github.com/DigitalBiomarkerDiscoveryPipeline/CovIdentify).

Received: 12 July 2022; Accepted: 3 August 2022;
Published online: 01 September 2022

REFERENCES

1. COVID Live. *Coronavirus Statistics – Worldometer*. https://www.worldometers.info/coronavirus/ (2022).
2. United States COVID. *Coronavirus Statistics – Worldometer*. https://www.worldometers.info/coronavirus/country/us/ (2022).
3. Moghadas, S. M. et al. The implications of silent transmission for the control of COVID-19 outbreaks. *Proc. Natl Acad. Sci. USA* **117**, 17513–17515 (2020).
4. Why pretest and posttest probability matter in the time of COVID-19. *ASM.org* https://asm.org/Articles/2020/June/Why-Pretest-and-Posttest-Probability-Matter-in-the (2020).
5. Stolberg, S. G. & LaFraniere, S. With Omicron, U.S. testing capacity faces intense pressure. *The New York Times* (2021).
6. Reports of price gouging amid shortages of COVID-19 tests. https://www.cbsnews.com/news/at-home-test-covid-price-gouging/ (2022).
7. O'donnell, C. U.S. COVID-19 tests again in short supply as infections soar, schools reopen. *Reuters* (2021).
8. Omicron testing shortages and delays are making results useless—and deepening COVID inequality. *Fortune* https://fortune.com/2022/01/10/omicron-testing-shortages-delays-covid-inequality/ (2022).
9. Heilweil, R. How omicron broke COVID-19 testing. *Vox* https://www.vox.com/recode/2021/12/21/22848286/omicron-rapid-test-covid-19-antigen (2021).
10. More coronavirus tests will be available next month, Fauci says, as U.S. struggles with shortage. *Washington Post* (2021).
11. Huang, P. There has been a shortage of testing and vaccines for Monkeypox. *NPR* (2022).
12. Perspective | Testing failures helped covid spread. We must do better with monkeypox. *Washington Post* (2022).
13. Rader, B. et al. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *J. Travel Med.* **27**, taaa076 (2020).
14. Souch, J. M. & Cossman, J. S. A commentary on rural-urban disparities in COVID-19 testing rates per 100,000 and risk factors. *J. Rural Health* **37**, 188–190 (2021).
15. New surgo analysis identifies highly vulnerable rural communities as COVID-19 testing deserts. *Surgo Ventures* https://surgoventures.org/portfolio/action-areas/new-surgo-analysis-identifies-highly-vulnerable-rural-communities-as-covid-19-testing-deserts (2022).
16. Kannoth, S., Kandula, S. & Shaman, J. The association between early country-level COVID-19 testing capacity and later COVID-19 mortality outcomes. *Influenza Other Respir. Viruses* **16**, 56–62 (2022).
17. CDC. Community, work, and school. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/racial-ethnic-disparities/increased-risk-exposure.html (2020).
18. Meirom, E., Maron, H., Mannor, S. & Chechik, G. Controlling graph dynamics with reinforcement learning and graph neural networks. in *Proceedings of the 38th International Conference on Machine Learning* 7565–7577 (PMLR, 2021).
19. Du, J. et al. Optimal diagnostic test allocation strategy during the COVID-19 pandemic and beyond. *Stat. Med.* **41**, 310–327 (2022).
20. Radin, J. M., Wineinger, N. E., Topol, E. J. & Steinhubl, S. R. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *Lancet Digit. Health* **2**, e85–e93 (2020).
21. Li, X. et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biol.* **15**, e2001402 (2017).
22. Shapiro, A. et al. Characterizing COVID-19 and influenza illnesses in the real world via person-generated health data. *Patterns* **2**, 100188 (2021).
23. Mishra, T. et al. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat. Biomed. Eng.* **4**, 1208–1220 (2020).
24. Quer, G. et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nat. Med.* **27**, 73–77 (2021).
25. Miller, D. J. et al. Analyzing changes in respiratory rate to predict the risk of COVID-19 infection. *PloS One* **15**, e0243693 (2020).
26. Grzesiak, E. et al. Assessment of the feasibility of using noninvasive wearable biometric monitoring sensors to detect influenza and the common cold before symptom onset. *JAMA Netw. Open* **4**, e2128534 (2021).
27. Alavi, A. et al. Real-time alerting system for COVID-19 and other stress events using wearable data. *Nat. Med.* **28**, 175–184. https://doi.org/10.1038/s41591-021-01593-2 (2022).
28. Harris, P. A. et al. Research Electronic Data Capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
29. Rapeport, G. et al. SARS-CoV-2 human challenge studies—establishing the model during an evolving pandemic. *N. Engl. J. Med.* **385**, 961–964 (2021).
30. CDC. COVID-19 and your health. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html (2020).
31. Radin, J. M. et al. Assessment of prolonged physiological and behavioral changes associated with COVID-19 infection. *JAMA Netw. Open* **4**, e2115959 (2021).
32. Krzywinski, M. & Altman, N. Error bars. *Nat. Methods* **10**, 921–922 (2013).
33. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
34. Gadaleta, M. et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. *Npj Digit. Med.* **4**, 1–10 (2021).
35. Mason, A. E. et al. Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict Study. *Sci. Rep.* **12**, 3463 (2022).
36. Conroy, B. et al. Real-time infection prediction with wearable physiological monitoring and AI to aid military workforce readiness during COVID-19. *Sci. Rep.* **12**, 3797 (2022).
37. Natarajan, A., Su, H.-W. & Heneghan, C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *Npj Digit. Med.* **3**, 1–8 (2020).
38. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).
39. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. in *Proceedings of the 23rd international conference on Machine learning – ICML '06* 233–240 (ACM Press, 2006). https://doi.org/10.1145/1143844.1143874.
40. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. *Nat. Methods* **13**, 703–704 (2016).
41. Institute of Medicine (US) Council on Health Care Technology; Sox, H., Stern, S., Owens, D. & Abrams, H. L. in *The Use of Diagnostic Tests: A Probabilistic Approach*. *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions: Monograph of the Council on Health Care Technology* (National Academies Press (US), 1989).
42. Frequently Asked Questions (FAQs). *grants.nih.gov*. https://grants.nih.gov/faqs#/inclusion-basic-on-sex-gender-and-race-ethnicity.htm (2020).
43. Watson, J., Whiting, P. F. & Brush, J. E. Interpreting a COVID-19 test result. *BMJ* **369**, m1808 (2020).
44. Gubbay, J. B. et al. Impact of coronavirus disease 2019 (COVID-19) pre-test probability on positive predictive value of high cycle threshold severe acute respiratory coronavirus virus 2 (SARS-CoV-2) real-time reverse transcription polymerase chain reaction (RT-PCR) test results. *Infect. Control Hosp. Epidemiol.* 1–5 https://doi.org/10.1017/ice.2021.369 (2021).
45. Cho, P. J. et al. Demographic Imbalances Resulting From the Bring-Your-Own-Device Study Design. *JMIR Mhealth Uhealth* **10**, e29510 (2022).
46. Vogels, E. A. Digital divide persists even as Americans with lower incomes make gains in tech adoption. *Pew Research Center* https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/ (2021).
47. Vogels, E. A. Some digital divides persist between rural, urban and suburban America. *Pew Research Center* https://www.pewresearch.org/fact-tank/2021/08/19/some-digital-divides-persist-between-rural-urban-and-suburban-america/ (2021).

48. Macridis, S., Johnston, N., Johnson, S. & Vallance, J. K. Consumer physical activity tracking device ownership and use among a population-based sample of adults. *PLoS One* **13**, e0189298 (2018).
49. Guk, K. et al. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials* **9**, 813 (2019).
50. Demographics of mobile device ownership and adoption in the United States. *Pew Research Center: Internet, Science & Tech* https://www.pewresearch.org/internet/fact-sheet/mobile/ (2021).
51. Silver, L. Smartphone ownership is growing rapidly around the world, but not always equally. *Pew Research Center's Global Attitudes Project* https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ (2019).
52. U.S. Census Bureau QuickFacts: United States. https://www.census.gov/quickfacts/fact/table/US/RHI125219 (2020).
53. CDC. COVIDView, Key Updates for Week 23. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html (2020).
54. Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **19**, 281 (2019).
55. Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
56. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res.* **11**, 2079–2107 (2010).
57. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinforma.* **7**, 91 (2006).

ACKNOWLEDGEMENTS

This work was supported in part by Duke OIT, Duke Bass Connections Fellowship, Duke Margolis Center for Health Policy, Duke MEDx, Microsoft AI in Health, Duke CTSI (UL1TR002553), and NC Biotech (2020-FLG-3884). This article was prepared while G.S.G. was employed at Duke University. D.K.P. was funded by NIH/NICHD (R25HD079352) and CDC (BAA 75D301-20-R-68024). The opinions expressed in this article are the authors' own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government. We would like to also thank Leatrice Martin, Veronica Palacios-Grandez, John Owens, Julie Ekstrand, Cecilia Plez, Hugh Thomas, Philabian Lindo, Richard Outten, Shellene Walker, Tracey Futhey, Jimmy Dorff, Rob Carter, Sean Dilda, Vanessa Simmons, Andy Ingham, Charley Kneifel, Andrew Olson, Whitney Welsh, Jonathan McCall, Margaret Pendzich, Marialuisa Solis-Guzman, Erich Huang, Victoria Christian, Marissa Stroo, Ceci Chamorro, Camila Reyes, and Ashanti Ballard for their contributions to the infrastructure and their dedication to this project. This publication was made possible in part by BAA 75D301-20-R-68024 from the US Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the US CDC.

AUTHOR CONTRIBUTIONS

Study conception and design: J.P.D., M.M.H.S., P.J.C. and A.R.R. Project supervision: J.P.D. and R.J.S. IRB review and participant recruitment and coordination: P.J.C., K.S., W.W., D.K.P., G.S.G., C.W.W., R.J.S. and J.P.D. E-consent system (REDCap) and participant guidance: P.J.C., K.S. Wearable and survey data collection and processing: M.M.H.S, P.J.C., A.R.R., O.M.E., S.F., Q.X.K., Y.K., X.L., J.H., A.K., A.B., A.A. and U.R. Software engineering: P.J.C., A.S., R.S., and B.T. Algorithm development and data analysis: M.M.H.S., P.J.C., A.R.R., and J.P.D. Manuscript preparation: M.M.H.S., P.J.C., A.R.R. and J.P.D. Manuscript review and editing: all co-authors. Funding: J.P.D. and R.J.S.

COMPETING INTERESTS

J.P.D. is an Associate Editor of *npj Digital Medicine*. M.M.H.S is an Editorial Board Member of *npj Digital Medicine*. J.P.D. is on the scientific advisory board of Human Engineering Health Oy and is a consultant for ACI Gold Track. C.W.W. is a founder of Predigen that is merged with Biomeme. He is also on the scientific advisory board of Biomeme/Predigen, FHI Clinical, IDbyDNA, Regeneron, and Roche Molecular Sciences. He is also a consultant for bioMerieux/Biofire, Domus, Karius, Steradian, and SeLux Diagnostics. He is also on the data and safety monitoring board of Bavarian Nordic and Janssen. M.P.S. is a co-founder and member of the scientific advisory board of Personalis, Qbio, January, SensOmics, Protos, Mirvie, NiMo, Onza, and Oralome. He is also on the scientific advisory board of Danaher, Genapsys, and Jupiter. Other authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00672-z.

Correspondence and requests for materials should be addressed to Jessilyn P. Dunn.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.