Journal of Clinical and Translational Science

www.cambridge.org/cts

Translational Research, Design and Analysis Special Communication

Cite this article: Verma SS, Chung WK, Dudek S, Williamson JL, Verma A, Robinson S, Rader DJ, Reilly MP, Sengupta S, FitzGerald GA, Kiryluk K, and Ritchie MD. Research on COVID-19 through patient-reported data: a survey for observational studies in the COVID-19 pandemic. *Journal of Clinical and Translational Science*, page 1 of 5. doi: 10.1017/cts.2020.509

Received: 22 May 2020
Revised: 30 June 2020
Accepted: 1 July 2020

Keywords:
COVID-19 data collection; infection rates; patient survey; pre-existing conditions; co-morbidities; population health; lifestyle factors

Address for correspondence:
M. D. Ritchie, PhD, A301 Richards Building, 3700 Hamilton Walk, University of Pennsylvania, Philadelphia, PA 19104, USA.
Email: marylyn@pennmedicine.upenn.edu

Authors contributed equally as first and last # authors respectively.

Research on COVID-19 through patient-reported data: a survey for observational studies in the COVID-19 pandemic

Shefali Setia Verma[1,*], Wendy K. Chung[2,3,*], Scott Dudek[1], Jennifer L. Williamson[4], Anurag Verma[1], Scott Robinson[3], Daniel J. Rader[1,5,6], Muredach P. Reilly[7,8], Soumitra Sengupta[9], Garret A. FitzGerald[5,6], Krzysztof Kiryluk[2,10,#] and Marylyn D. Ritchie[1,5,#]

[1]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; [2]Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA; [3]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY, USA; [4]Vagelos College of Physicians and Surgeons, Office for Research, Columbia University, New York, NY, USA; [5]The Institute for Translational Medicine and Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; [6]Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; [7]Cardiology Division, Department of Medicine, Columbia University, New York City, NY, USA; [8]The Irving Institute for Clinical and Translational Research, Columbia University, New York City, NY, USA; [9]Department of Biomedical Informatics, Columbia University, New York, NY, USA and [10]Division of Nephrology, Columbia University, New York, NY, USA

Abstract

Understanding the clinical risk factors for COVID-19 disease severity and outcomes requires a combination of data from electronic health records and patient reports. To facilitate the collection of patient-reported data, as well as accelerate and standardize the collection of data about host factors, we have constructed a COVID-19 survey. This survey is freely available to the scientific community to send electronically for patients to complete online. This patient survey is designed to be comprehensive, yet not overly burdensome, to gather data useful for a range of clinical investigations, and to accommodate a wide variety of implementation settings including at a COVID-19 testing site, at home during infection or after recovery, and/or for individuals while they are hospitalized. A widely adopted standardized survey that can be implemented online with minimal resources can serve as a critical tool for combining and comparing data across studies to improve our understanding of COVID-19 disease.

Introduction

A novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes COVID-19, began in Wuhan, China in late 2019 [1,2] but now has led to a global pandemic [3,4]. This disease has spread to over 100 countries, and many are experiencing their first wave of exposure to this novel coronavirus. As of June 30, 2020, more than 10.1 million individuals worldwide have tested positive for SARS-CoV-2 infection and over 503,000 people have died due to COVID-19 [5]. Severe infection causes acute respiratory illness as well as other severe complications including acute renal failure, hepatitis, myocarditis, arrhythmias, thrombosis, heart attack, and stroke [6–9]. One of the greatest challenges in managing patients with COVID-19 is the variability in host response to the virus and the wide range of clinical outcomes [10]. While severely ill patients require hospitalization and intensive care, most others have milder symptoms such as fever, sore throat, cough, anosmia, and myalgias; and estimates of asymptomatic infection are up to 50% of those infected [11,12]. Understanding the factors that influence spread of the virus and differential response to infection is critical. These host factors potentially include age, sex, occupation, community characteristics, social determinants of health, comorbidities, medications, and host genetics.

To accelerate and standardize the collection of data about host factors, we have constructed a COVID-19 survey. Some of the data needed for prediction modeling are available in electronic health records (EHRs). However, participant-reported surveys can serve as a valuable supplementary tool to target different questions that are not captured in an EHR [13–16]. Additionally, such data can be gathered from individuals who have not been hospitalized. This is particularly relevant in this pandemic since many infected patients are either asymptomatic or mildly symptomatic and remain at home. Adoption of standardized variables is essential for interoperability, aggregation, replication, and validation of findings across multiple large studies [17,18]. Standardized instruments will facilitate maximally effective use of data contributed by

participants. Other groups have developed surveys to address different research questions. For example, daily health surveys about symptoms and how individuals feel have been implemented by several groups, including HowWeFeel [19], Apollo [20], and COVID-19 Watcher [21]. The COVID symptom tracker [22] helps to identify regional hotspots, prevalence of disease in different regions, as well identifying early symptoms. Workplace surveys, such as those targeting health care workers and other essential workers, are focused on evaluating the risk of COVID-19 exposure given current conditions in the workplace [23,24]. Additional surveys are available from the CDC, Johns Hopkins, and the Global Alliance for Genomics and Health; all of these can be accessed at the CD2H website [25]. Additionally, the PhenX Toolkit along with NLM and NIEHS have created a site with many of these surveys catalogued [26]. While there are many surveys available currently, the survey that we developed was intended to serve both in-patient and outpatient COVID-19 positive individuals, asymptomatic individuals, and COVID-19 negative individuals as well. We recognize that it is important to include COVID-19 negative individuals in the data collection process so that in our statistical analyses, we have a comparison group, especially for looking at comorbid conditions.

At the time when this survey was created, no single survey captured all of the information that we were interested in collecting in our patients. It is possible that at this point, several surveys address a similar set of questions. Nonetheless, these two surveys are freely available and are easy to implement in REDCap, so likely useful for organizations who have not yet implemented COVID-19 surveys. We selected questions based on feedback from patients, community advisory council members, clinicians, scientists, and epidemiologists at both of our institutions. We looked at how some questions were asked in other surveys and followed recent studies to add questions while focusing on brevity and content. The idea was to capture information potentially relevant to COVID-19 disease severity and outcomes that may be missing from the EHR.

Here, we present a standardized participant-reported online survey (http://covidhealthquest.com) applicable to: (1) individuals at high risk of exposure, such as health care workers; (2) individuals who have tested positive for SARS-COV-2 as outpatients, with or without symptoms; (3) infected individuals who have been admitted to the hospital; (4) individuals who have recovered from COVID-19; (5) individuals with symptoms who have not been tested; and (6) currently healthy individuals who may be at risk for future infection. We encourage broad deployment of this survey based on self-reported data to capture a large, diverse population to create a standardized dataset to address hypotheses about COVID-19 infection. At both Columbia University and the University of Pennsylvania, an electronic version of the survey is being administered automatically to all individuals providing an electronic consent to participate in the Columbia COVID-19 Biobank or Penn Medicine COVID-19 Biobank, respectively. The survey completion rates exceed 83% among those 1617 patients who agree to participate in the Biobank at Columbia and 91% of the 900 patients who consented at Penn. This survey tool has also been disseminated widely and internationally. For example, the translated version of the survey is being administered during biobank recruitment at the University of Warsaw in Poland and by the Gaslini Institute in Genova, Italy. Additionally, this survey was added to the collection of phenotype tools compiled by the COVID-19 Host Genetics Initiative of the International Common Disease Alliance. We believe that this survey, integrated with EHR data and biospecimens, could identify clinical and environmental covariates important for prediction of the clinical course after SARS-CoV-2 infection. Such information will allow for more accurate risk stratification early in the course of infection and facilitate the design of randomized trials to test therapies that might prevent progression to severe phenotypes of the disease.

COVID-19 Patient Survey

This survey is designed to be comprehensive, yet not overly burdensome, to gather data useful for a range of clinical investigations and to accommodate a wide variety of implementation settings including at a COVID-19 testing site, at home during infection or after recovery, and/or for individuals while they are hospitalized. The information collected in the survey is divided into four main themes for ease of self-report and organization:

1. **Personal profile:** The goal of this section is to collect basic demographic, socioeconomic, and household members' information.
2. **COVID-19-related questions:** The questions in this section are focused on gathering information on COVID-19 symptoms and disease course from infected individuals. Questions regarding symptoms, travel history, social interactions, self-reported quarantine compliance, and known exposures to individuals with COVID-19 are asked in this section.
3. **General health questions:** This section addresses past and present medical history and medication-related questions. Participants report this information to supplement EHR data (when available) to facilitate quick data capture. These questions are designed to address the contribution of underlying preexisting health conditions and medication usage to clinical outcomes.
4. **Lifestyle and environmental exposure questions:** This last section addresses questions about education, occupation, use of recreational drugs, smoking, vaping, and physical activity.

This survey was developed through the collaboration of two sites holding Clinical and Translational Science Awards (CTSA) from the NIH with the goal of deployment across multiple health systems for future large-scale collaborations [25]. Clinician, scientist, patient, community advisory council, and epidemiologist feedback were gathered to ensure that questions were clear and concise and that response choices included the full range of answers. Two versions of the survey have been developed to allow for two different survey strategies to be deployed. The two formats allow for the trade-off between the time it takes to complete the survey and the level of detail of the questions asked. The goal here was to strike a balance between patient burden and depth of information. The majority of questions in the two surveys are overlapping with the intention of replicability and validation of the findings across study sites. A visual comparison of the two surveys is shown in Fig. 1.

Both surveys are freely available to be used by the global community at http://covidhealthquest.com to collect these measures on COVID-19 disease in both confirmed and potentially affected individuals.

Implementation of the Surveys

Both surveys were implemented in REDCap and approved by local institutional review boards. Answers from participants are evaluated as they are entered, using REDCap adaptive logic to determine if additional questions are relevant. This strategy simplifies the procedure, eliminating questions irrelevant to the participant. This also increases the time to complete the survey. Most questions were designed to be answered with check boxes or in drop-down

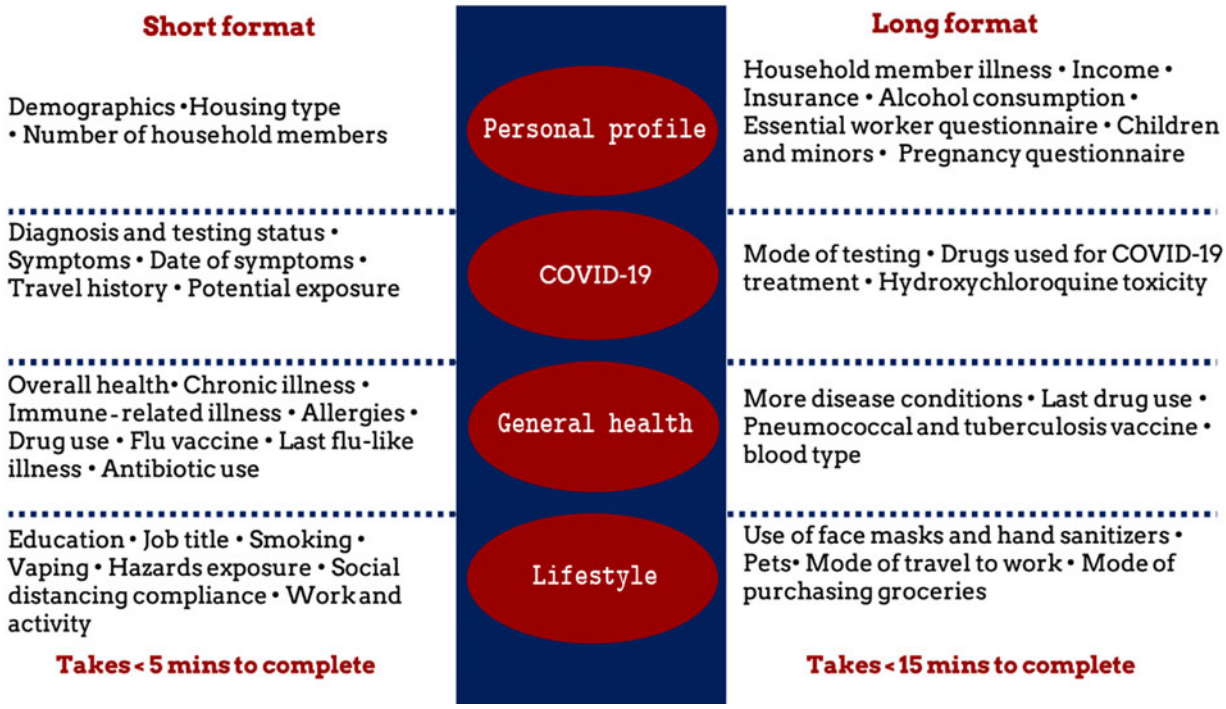

Fig. 1. Description of questions collected in short format of survey (in left panel) and additional questions asked in long format of survey (in right panel). The survey questions are divided into four major themes as listed in the center panel.

menus so as to avoid free text entries; this facilitates greater capability for standardization, validation, and replication across sites. Additionally, both versions have been translated into eight other languages, including Spanish, Italian, German, French, Polish, Japanese, Chinese, and Korean. All of these translations are available from the GitHub repositories linked to the website (http://covidhealthquest.com).

University of Pennsylvania has gathered survey data from approximately 900 patients. Patients who were tested for COVID-19 were contacted along with subset of individuals who were part of existing biobank (Penn Medicine BioBank). The REDCap implementation of the survey was distributed via email or text. From the set of 930 patients, there are 479 females (51.6%) and 449 males (48.3%), and 2 other (0.2%). The ethnicity/ancestry distribution is 77.0% White/European ancestry, 14.6% Black/African ancestry, 3.66% Asian ancestry, and 5.49% Hispanic. Nine hundred and one participants completed all questions, out of which 455 participants were diagnosed with COVID-19 and 446 were not diagnosed with COVID-19. Fig. 2 shows the distribution of preexisting conditions for these 901 individuals (dark blue bars).

Columbia University has collected survey data from 1454 patients, including 488 (66%) females, 965 (34%) males, and 1 (0.07%) other. The ancestry distribution is 38% White/European, 13% African-American, 35% Hispanic, 5% East Asian, and 9% Multiple ancestry or other. Fig. 2 shows the distribution of preexisting conditions for these individuals (light blue bars).

Once we have collected survey data from a large number of patients such that we have statistical power for a data freeze and thorough statistical analysis, we will merge the survey data with EHR data for each patient. This will allow for (1) verification of the clinical information including clinical conditions and medications and (2) integration with other clinical information that was not requested in the survey such as hospitalization or outcomes from their COVID-19 infection. We also plan to collect biospecimens from all of these individuals, such that downstream multi-omics analyses can be integrated with these surveys and EHR data. The pre-existing conditions shown in Fig. 2 are preliminary based on the first group of patients recruited into each biobank and thus, it is too early to draw any conclusions about these results.

Once these data have been collected at the University of Pennsylvania and Columbia University, it is anticipated that the results will be shared through the National COVID Cohort Collaborative (N3C) [27], of which both Penn and Columbia are members. This consortium is aggregating clinical data across the country to create a large, comprehensive dataset to ask questions regarding COVID-19. We anticipate that the survey data collected would be shared along with the EHR data through this initiative.

Summary

The goal of these surveys is to collect participant-reported data in a simple and standardized digital approach. Adoption of standardized variables is essential for interoperability and data aggregation and facilitates replication and validation of findings across multiple large studies. A widely adopted standardized survey that can be implemented online with minimal resource can serve as a critical tool for combining and comparing data across studies. Such harmonized survey data can also facilitate validation and replication. Though the surveys presented here are balanced in terms of number of questions posed versus participant engagement and convenience, the primary aim is to engage as many participants as possible in sharing the data and to gather the information not available through EHRs. Integration of survey data with biosamples and phenotypes extracted from EHRs may elucidate factors that predict disease progression, improve the design of clinical trials, and aid in refining strategies for prevention, quarantine, and social distancing.

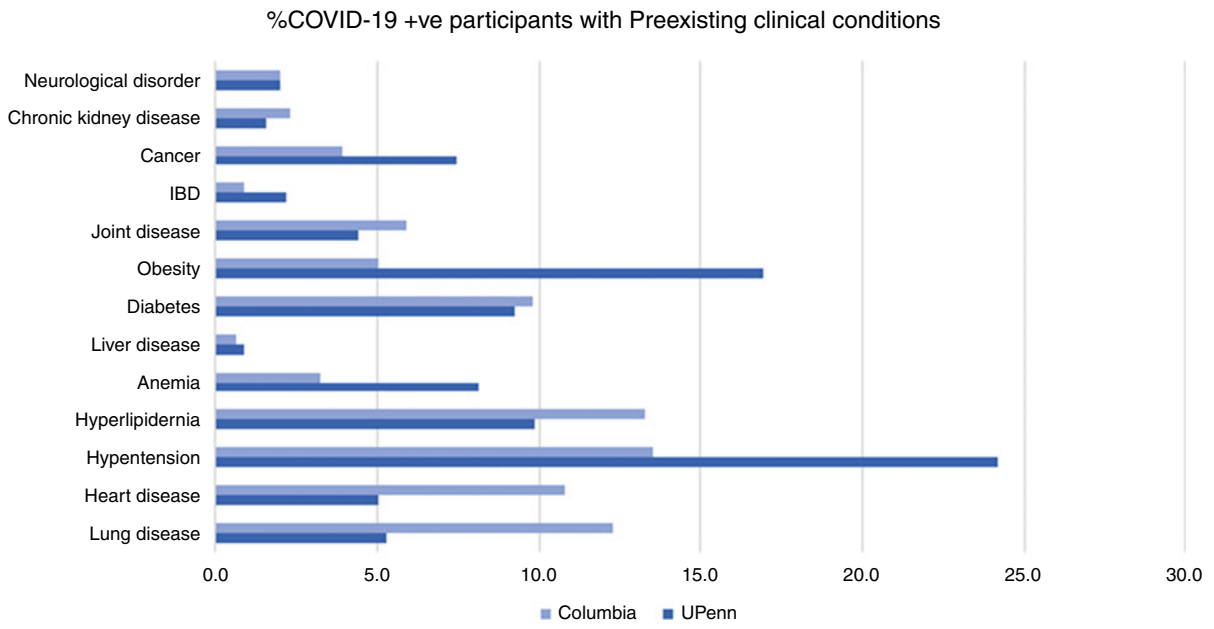

Fig. 2. Counts of preexisting clinical conditions among COVID-19 positive (+ve) participants in the first set of survey respondents at University of Pennsylvania (dark blue) Columbia University (light blue). IBD, inflammatory bowel disease.

Though we describe several important features and utilities of our survey, we understand that deployment of surveys and analysis of data along with the feedback from participants, clinicians, as well as policymakers would help identify biases and heterogeneity in implementation. This feedback will be used constructively for iterative improvements in follow-up surveys. We have translated the survey in eight additional languages to facilitate global use by health institutions and clinical trials sites for future collaborations and replication of findings across studies. The initial analyses from first respondents of the survey provide insights into the comorbidities, but further investigation would help identify if the comorbidities merely reflect the burden of aging or are truly indicative of susceptibility to severe COVID-19 outcomes. A larger sample size is needed for better inference.

Acknowledgments. We would like to thank the native speakers who performed translations of the survey forms into their respective languages, including Perla Mendoza, Karla Mehl, Miguel Hernandez, and Graciela Gonzalez-Hernandez (Spanish), Francesca Lugani and Giorgio Sirugo (Italian), Natalia Krata, Bartosz Foroncewicz, and Marta Byrska Bishop (Polish), Pierre Cohen and Marie Laure Rowland (French), Jan Lanzer and Daniel Albert (German), Jingyuan Xie, Binglan Li, and Xinyuan Zhang (Chinese), Hioshi Suzuki and Yuki Bradford (Japanese), and Hajeaong Lee, Dokyoon Kim, and Garam Lee (Korean). This work was supported by the Precision Medicine Resource of the Columbia CTSA, grant number UL1TR001873 from the National Center for Advancing Translational Sciences (NCATS), and the University of Pennsylvania CTSA (grant number 1U54TR001623). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health (NIH).

Code availability. Codebooks and REDCap data dictionaries for both survey tools are available on from our website http://covidhealthquest.com. This site links to the GitHub repositories for the most up-to-date versions of the surveys as an open source project for broad adoption by the scientific community. In addition to English, we provide additional questionnaire versions translated into Spanish, Italian, German, French, Polish, Japanese, Chinese, and Korean. The translations were conducted by native speakers with clinical background for each language.

Disclosures. These authors declare that there is no conflict of interest: SSV, SD, JLW, AV, SR, DJR, MPR, SS, KK. WKC reports Scientific Advisory Board, Regeneron Genetics Center. GAF is a senior advisor to Calico Laboratories. MDR is on the scientific advisory board for Cipherome and Goldfinch Bio.

References

1. **Wu F, Zhao S, Yu B, *et al.*** A new coronavirus associated with human respiratory disease in China. *Nature* 2020; **579**(7798): 265–269.
2. **Zhou P, Yang X-L, Wang X-G, *et al.*** A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; **579**(7798): 270–273.
3. **Zheng J.** SARS-CoV-2: an emerging coronavirus that causes a global threat. *International Journal of Biological Sciences* 2020; **16**(10): 1678–1685.
4. **Gates B.** Responding to covid-19 — a once-in-a-century pandemic? *The New England Journal of Medicine* 2020; **382**: 1677–1679. doi: 10.1056/NEJMp2003762
5. **WHO.** Coronavirus disease (COVID-19) pandemic [Internet], 2020. (https://www.who.int/emergencies/diseases/novel-coronavirus-2019)
6. **Young BE, Ong SWX, Kalimuddin S, *et al.*** Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. *JAMA* 2020; **323**: 1488–1494.
7. **Zhang C, Shi L, Wang F-S.** Liver injury in COVID-19: management and challenges. *The Lancet Gastroenterology & Hepatology* 2020; **5**(5): 428–430.
8. **Boettler T, Newsome PN, Mondelli MU, *et al.*** Care of patients with liver disease during the COVID-19 pandemic: EASL-ESCMID position paper. *JHEP Reports* 2020; **2**(3): 100113.
9. **Zhang J-J, Dong X, Cao Y-Y, *et al.*** Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy* 2020; **75**: 1730–1741.
10. **CDC. Coronavirus Disease 2019 (COVID-19).** Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19) [Internet], 2020. (https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html)
11. **Day M.** Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ* 2020; **369**: m1375.
12. **Bai Y, Yao L, Wei T, *et al.*** Presumed asymptomatic carrier transmission of COVID-19. *JAMA* 2020; **323**(14): 1406.
13. **Estabrooks PA, Boyle M, Emmons KM, *et al.*** Harmonized patient-reported data elements in the electronic health record: supporting meaningful use by primary care action on health behaviors and key psychosocial

factors. *Journal of the American Medical Informatics Association* 2012; **19**(4): 575–582.
14. **Owen-Smith A, Mayhew M, Leo MC, *et al.*** Automating collection of pain-related patient-reported outcomes to enhance clinical care and research. *Journal of General Internal Medicine* 2018; **33**(Suppl 1): 31–37.
15. **Snyder C, Wu AW.** Users' Guide to Integrating Patient-Reported Outcomes in Electronic Health Records [Internet]. Johns Hopkins University; 2017. (https://www.pcori.org/sites/default/files/PCORI-JHU-Users-Guide-To-Integrating-Patient-Reported-Outcomes-in-Electronic-Health-Records.pdf)
16. **Kukafka R, Ancker JS, Chan C, *et al.*** Redesigning electronic health record systems to support public health. *Journal of Biomedical Informatics* 2007; **40**(4): 398–409.
17. **Gold M, Mclaughlin C.** Assessing HITECH implementation and lessons: 5 years later. *The Milbank Quarterly* 2016; **94**(3): 654–687.
18. **Hulsen T, Jamuar SS, Moody AR, *et al.*** From big data to precision medicine. *Frontiers in Medicine (Lausanne)*. 2019; **6**: 34.
19. HowWeFeel: An App Designed to Tackle Coronavirus Together [Internet]. Department of Biostatistics. 2020. (https://www.hsph.harvard.edu/biostatistics/2020/04/howwefeel-an-app-designed-to-tackle-coronavirus-together/)
20. **Spector R.** Stanford medicine team launches survey via smartphone app to assess impact of COVID-19. Stanford Medicine News Center [Internet], April 16, 2020. (http://med.stanford.edu/news/all-news/2020/04/stanford-team-launches-covid-19-survey.html)
21. **Columbia COVID-Watcher [Internet]**. Columbia University Irving Medical Center. (https://cumc.co1.qualtrics.com/jfe/form/SV_8xnQr57NLKW1XRb?s=sms)
22. COVID Symptom Tracker - Help slow the spread of COVID-19 [Internet]. (https://covid.joinzoe.com/us)
23. Survey for Local Healthcare Workers – Workforce Reserve for COVID-19 [Internet]. Saint Mary's County Health Department. 2020. (http://www.smchd.org/2020/04/survey-for-local-healthcare-workers-workforce-reserve-for-covid-19/)
24. COVID-19 Workplace survey [Internet]. (https://www.surveymonkey.com/r/COVID-19survey)
25. COVID Case Report Forms and Phenotyping Standards | CD2H COVID-19. (https://covid.cd2h.org/forms_and_standards)
26. NIH Disaster Research Response [Internet]. (https://dr2.nlm.nih.gov/)
27. National COVID Cohort Collaborative (N3C) [Internet]. National Center for Advancing Translational Sciences. 2020. (https://ncats.nih.gov/n3)