Evaluating Medical Devices Remotely: Current Methods and Potential Innovations

Anne Collins Mclaughlin

North Carolina State University

Patricia R. Delucia

Rice University

Frank A. Drews

University of Utah

Monifa Vaughn-Cooke

University of Maryland

Anil Kumar

San Jose State University

Robert R. Nesbitt

AbbVie

Kevin Cluff

BioWork Engineering, LLC

All authors are affiliated with the HFmedic consortium.

Keywords: Analysis and evaluation; Design strategies, tools; Qualitative methods; Remote usability testing and evaluation; Medical devices and technologies

Precis: With the long term impact of the COVID-19 pandemic comes the need to continue medical device research, but remotely. We provide a review of the opportunities and challenges of remote evaluations and conclude that in many cases, remote evaluations are viable options for medical devices.

Abstract

Objective: We present examples of laboratory and remote studies, with a focus on studies appropriate for medical device design and evaluation. From this review and description of extant options for remote testing, we provide methods and tools to achieve research goals remotely.

Background: The FDA mandates human factors evaluation of medical devices. Studies show similarities and differences in results collected in laboratories compared to data collected remotely in non-laboratory settings. Remote studies show promise, though many of these were behavioral studies related to cognitive or experimental psychology. Remote usability studies are rare but increasing as online technologies allow for synchronous and asynchronous data collection.

Method: We reviewed potential methods of remote evaluation of medical devices, from testing labels and instruction to usability testing and simulated use. Each method was coded for the attributes (e.g., supported media) that need to be considered when designing a usability study.

Results: We present examples of how published usability studies of medical devices could be moved to remote data collection. We also present novel systems for creating such tests, such as the use of 3D printed or virtual prototypes. Finally, we advise on targeted participant recruitment.

Conclusion: Remote testing will bring opportunities and challenges to the field of medical device testing. Current methods are adequate for most purposes, excepting the validation of Class III devices.

Application: The tools we provide enable remote evaluation of medical devices. Evaluations have specific research goals, and our framework of attributes helps to select or combine tools for valid testing of medical devices.

Evaluating Medical Devices Remotely: Current Methods and Potential Innovations

Due to the COVID-19 pandemic and the need for social distancing to reduce spread of the coronavirus, laboratory research has decreased in a wide range of disciplines (Servick, Cho, Guglielmi, Voge, & Couzin-Frankel, 2020), with termination of studies that involve in-person data collection from human participants (Clay, 2020). This affects not only academic institutions but industries that develop medical devices and must provide human factors validation to receive FDA approval. One alternative is to conduct human factors testing remotely.

We present an overview of the technologies and best practices for remote evaluations of medical devices, from observational studies to usability tests to controlled behavioral experiments. We combined searches of the literature using the Summon database with our own knowledge of tools used by user researchers in industry. Many of the tools most used in industry did not show up in the published literature, but we believed it was important to detail their features to best help those needing to user test medical devices remotely. Because remote testing is a cutting-edge field, we limited our literature search to the last fifteen years and emphasized work found from the last five years. We evaluated the match of these methods to FDA guidelines for medical device evaluation, the attributes of devices that can be tested, and other considerations such as cost and whether the platform was well-established. We focused on options that required little-to-no knowledge of programming or system administration.

The FDA outlines the expectations for a human factors evaluation of a medical device according to device class (FDA, 2016). Class I devices are considered low-risk, for example a surgical tool. Class II devices have some risk in their use, for example pregnancy test kits or infusion pumps. Class III devices are considered high risk, as they often sustain life such as ventilators and pacemakers. Only 10% of medical devices are Class III (FDA). Because most remote testing will be formative, it can apply to all classes of device. However, for summative assessment remote testing will be most difficult for Class III devices and, at times, impossible.

The data collected for medical device usability can vary from qualitative and contextual information gathered during formative testing to safety-related use errors in summative testing. The most commonly needed data include signs of difficulty, close calls, and use errors. Reference to instructions for use, need for assistance, and unsolicited comments are also often desired (Wiklund, Kendler, & Strochlic, 2016). Many tools and techniques transfer well to remote use, such as surveys, interviews, and expert evaluations. Others are more challenging, such as simulated use and recruiting representative users for validation testing. Items to be tested vary as well, from the usability of instructions and warnings to the operation of physical devices. The methods reviewed here are most useful for testing Class I and Class II medical devices, and for formative evaluation of Class III devices for pre-market review processes (FDA, 2016).

Remote summative testing is more of a challenge and has not been addressed in published literature. Summative testing focuses on safety-related use errors with the actual device, meaning that a production-level device must be in the hands of the user (often a three-dimensional object). When collecting data during a summative test, use errors must be recorded and cannot be missed. Further, a comprehensive set of representative tasks must be carried out by representative users under the conditions that would be expected in the field. As summative testing is essentially required for Class II and Class III medical devices, it will depend on the device and testing needs to determine if a remote test is possible. An exception is the summative usability testing of electronic health records which do not require specialized equipment to be sent to the participant for usability testing. Though EHRs are not regulated as medical devices by the FDA (21st Century Cures Act of 2016), usability testing is needed to meet the Safety-Enhanced Design requirement of the Office of the National Coordinator for Health Information Technology (ONC, 2015).

Although usability testing performance is often measured on the order of minutes, we note that delays due to network connectivity issues could be a limitation if performance must be measured on the order of seconds. Thus, network connection would be a limiting factor for

many summative tests - at the very least, making some participant data unusable. However, it is unlikely to affect performance measurements for formative tests. That said, a poor connection, with dropped audio and video throughout, is a barrier to communication and heightens frustration, making even some formative tests or interviews unusable. As mentioned in our later section regarding recruitment, users at home with lower socioeconomic status may be the most adversely affected, either through bandwidth or through the needs of many persons in a home to use the same internet connection.

Comparison of laboratory and remote testing environments

Remote testing has the advantage of collecting data from large, diverse populations, quickly at low cost (Woods, Velasco, Levitan, Wan, & Spence, 2015). However, the *sine qua non* of a remote test is whether online results replicate those from a controlled laboratory environment. The implication is that the same psychological processes were activated in the two testing environments despite their physical differences (psychological fidelity, Kantowitz, 1988).

Cognitive performance. The quality of results from online studies of cognitive performance often are comparable to laboratory studies (Woods et al., 2015; see also Germine et al., 2012). Replicated results included the Forward Digit Span task, Flanker task, and Face Memory task. The most challenging results to replicate are those with short display presentations, such as masked priming tasks. Other concerns included stimulus timing (onset and duration), response time measurement, lack of stimulus control (e.g., visual size, luminance, resolution, color; auditory volume) of participant equipment, participant environment, duplicate or random responders, and ethical concerns about maintaining participant anonymity and privacy.

Results of a problem-solving laboratory study that compared three learning conditions were replicated in an online format but with a higher participant dropout rate and lower performance accuracy in the online condition (Dandurand, Shultz, & Onishi, 2008). Similarly, comparable performance data were obtained for online and laboratory administrations of an

interruption task that was time-sensitive, long in duration, and required sustained concentration (Gould, Cox, Brumby, & Wiseman, 2015). This study showed that online tests can replicate results of laboratory conditions for tasks that are more complex and longer in duration than those typically examined in comparisons of laboratory and online tests.

Although controlled laboratory experiments are considered the gold standard they are potentially limited by a lack of external validity, which is important to consider for medical device use. For example, some usability problems are unlikely to appear in a laboratory or highly controlled setting, such as sociological issues or working conditions not anticipated by the study designer (Wiklund, Kendler, & Strochlic, 2016). This is one reason we included review of tools that offer contextual and qualitative information on use (Table 1).

Usability tests. Results of usability testing of a regional hospital website in Switzerland were compared between laboratory and two remote testing conditions, including asynchronous and synchronous administrations (Sauer, Sonderegger, Heyden, Biller, Klotz, & Uebelbacher, 2019). Task completion rate, time, and efficiency did not differ across the three testing conditions. Nor were there differences between perceived usability, perceived workload, or affect. When the usability of a (computer simulated) smart phone was measured with laboratory and asynchronous remote formats, the difference in task completion time and efficiency between testing conditions was not significant when the usability of the smartphone was good (Sauer et al., 2019). When usability was poor, task completion time and click frequency was higher in the laboratory. Perceived usability ratings were higher in the lab, and workload did not differ statistically between testing conditions, regardless of the quality of the smartphone's usability.

Other examples of laboratory to online comparisons included comparing usability of email software in various tests-- conventional lab test, remote synchronous test, and remote asynchronous test. Findings showed few differences in performance results (e.g., task completion time), but identification of more usability issues by the conventional lab and remote

synchronous testing conditions (Andreasen, Nielsen, Schrøder, & Stage, 2007). Similar results were found when comparing synchronous lab and asynchronous remote testing using critical incident reporting, forum discussions, and longitudinal reporting in user diaries (Bruun, Gull, Hofmeister, & Stage, 2009). Evaluation of a shopping website using a think-aloud protocol had similar results when conducted in a laboratory or online, though the sample size was small (Thompson, Rozanski, & Haake, 2004). Descriptive results suggested that remote users took more time and made more errors, but identified more usability issues than in-person lab participants.

In summary, remote testing obtained results comparable to laboratory settings. However, published comparisons were few and limited to tasks without specialized hardware or software (e.g., vibrotactile devices, motion sensors). No studies were found comparing laboratory and remote testing of *medical devices.*

A "Human Factors Toolbox" for Remote Usability Testing of Medical Devices

We collected potential remote usability tools, from those appropriate for scientific study and use of inferential statistics to those intended to gather qualitative data from a small number of users. Because there are a large and growing number of software solutions available, we present those that are either most established or that enable a unique methodology. Table 1 provides a summary of tools and their attributes.

INSERT TABLE 1 HERE

Medical devices are often physical, three dimensional, with moving parts critical to their operation, and may require other equipment to be used (e.g., patient simulator). Prototypes are often expensive and difficult to create or repair. Because of the scarcity of remote medical device studies, we included usability testing of products similar to medical devices. Pros and

cons of each tool are provided, with considerations for the collection of performance and observational data (Table 1).

Remote testing relies on software that can host the surveys, stimuli, and enable communication. Some platforms were for specialized use, such as eye-tracking, while others purported to provide everything from participant recruitment to study-building to analysis and reports. Because of the particular attributes of medical devices, we have organized these platforms into categories: 1) those appropriate for the evaluation of 2-D or 'flat' stimuli: web interfaces, labels, instructions, and 2) those appropriate for the evaluation of 3D stimuli: physical devices and packaging. In each of these categories, we review how the platform has been used or validated in the psychological literature.

Website or other flat interface evaluation

Flat interfaces include websites, warnings, and labels. The available platforms varied greatly in terms of price, features, functionality, and need for technical knowledge (Table 1; Behavioral Experiment Hosting Platforms). Many were free to use but usually involved the need for more programming knowledge and online resources, such as web servers or installing the open source software from a repository. Sauter, Draschkow, and Mack (2020) provided a review of extant solutions for online behavioral studies requiring high experimental control. Overall, these platforms were for collecting scientific data. They emphasized timing accuracy and supported the typical protocol of "display stimuli -> collect response." Although they mimicked well-established measures, they have not all been validated to show that remote results were the same as those collected in a laboratory. They often differed in the inputs for the measures (e.g., allowing use of a mobile device rather than a keyboard) or in other ways that changed the outcome (e.g., screen brightness). The published comparisons of online are promising in this regard, but we recommend caution in assuming a validated cognitive test will replicate on these platforms.

Evaluation of 3D objects

Many medical devices need to be assessed with simulated use methods in a 3D environment. There are several options for evaluation, and the choice depends on the type of measures needed. The options are 1) display a virtual 3D prototype on a flat screen that can be manipulated via an interface (e.g., using the mouse to rotate the prototype to view the other side), 2) display a virtual 3D prototype on a flat screen using virtual or augmented reality, where the user can have limited interactions with the device, or 3) send a prototype or device to the user and record interactions via teleconferencing software or contextual video diary. All of these methods except the last are most suited for formative evaluation and iterative design (the "design verification" stage). The third option may fulfill simulated use at the "design validation" stage (Mejía-Gutiérrez & Carvajal-Arango, 2017).

The ubiquity of mobile devices with recording capability makes it possible for users to provide contextual information for their needs and use of medical devices (Table 1; General Purpose/Qualitative Emphasis). On the low cost end, users can take photos or make videos using their own mobile devices. These can be prompted by questions about their environment, such as "Show us how you manage your medications in the morning" or "Please make a video showing how you test your blood sugar using your current device." The constraint of relying on a user's smartphone concerns data access: it may be challenging for users to 1) understand how to send video files and 2) be able to store or transfer large video files using their own device. Also, populations of interest may not own, or be comfortable with smartphones. Commercial tools have been developed to aid user researchers in collecting these data. For example, indeemo (indeemo.com) offers a platform for remote, asynchronous ethnography, where users can be invited to download the app, prompts for audio, photo, video, or diary entries are automated, and the data are accessible to the researcher. Difficulty in recruiting some populations still applies.

The literature on remote 3D testing was sparse and often limited to novel computing solutions not easily available and accessible. We were unable to find any studies of remote evaluations of 3D medical devices, perhaps because, thus far, such tests have not been necessary. We did find evidence of testing done on other 3D devices, such as usability testing of a 3D mobile phone prototype online that showed the benefit of remote data collection (Figure 1, Kuutti et al, 2001). Kuutti et al. recommended training on use of the 3D viewer before exposure to the product. The evaluation of a camera interface in 3D (Kanai, Higuchi, & Kikuta, 2009) provided similar conclusions regarding benefits and limitations of a 3D virtual prototype usability test.

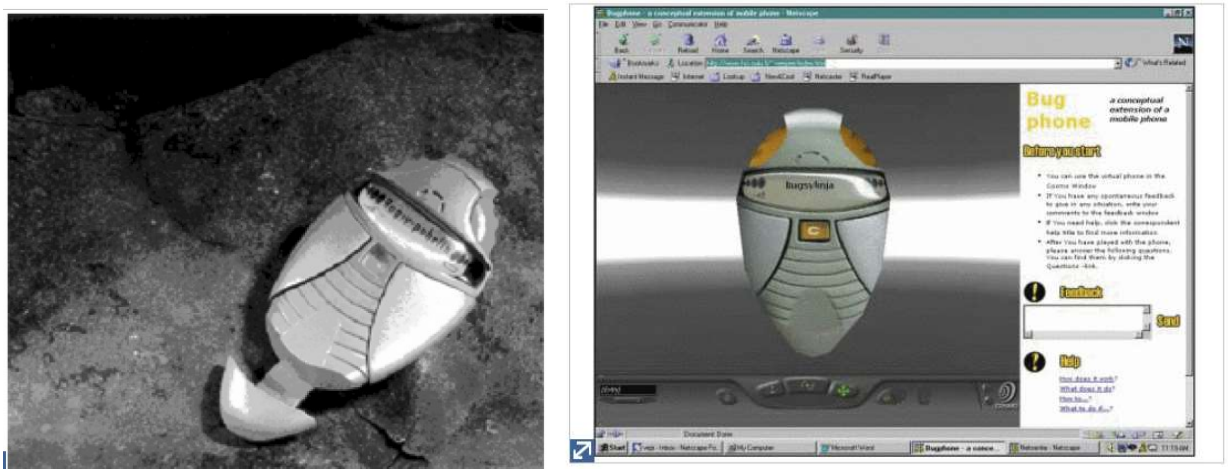

Figure 1. Virtual prototypes from Kuutti, 2001 as shown in usability tests.

Mixed reality (MR) was used for prototype testing, though not remotely. Mixed reality means a physical object was altered with virtual attributes. For example, in a study on the usability of a projector system, an abstracted physical form was created - a plastic block (Figure 2, Faust et al, 2019). When a fiducial marker was added to the form, participants saw an augmented reality image in the place of the block, where the block now appeared to be a fully functioning projector. Mixed reality thus allowed for physical interaction - the plastic block could be touched or lifted by a participant. Virtual buttons were shown on the block and participants could touch them to complete tasks with the projector. Performance and subjective

assessments were similar when compared to the same tasks with a real projector, making MR a promising option for 3D remote testing.

Some researchers developed head-mounted virtual and augmented reality displays using smartphones so that users could see objects in 3D, but these were not easily available (Rakkolainen, Raisamo, Turk, Höllerer, & Palovuori, 2016). Commercially available options included Google Cardboard (https://arvr.google.com/cardboard/), where a phone can be placed inside the cardboard viewer and held to the eyes to create an immersive virtual environment. Studies comparing in-lab VR systems to google cardboard systems found similar results (Mottelson & Hornbaek, 2017). Researchers can create virtual prototypes situated in a VR environment for remote testing. However, interactions with prototypes in VR are limited, making this method better for showing a design and collecting subjective data rather than performance data. No usability studies were found that employed this method for remote or in-person data collection.

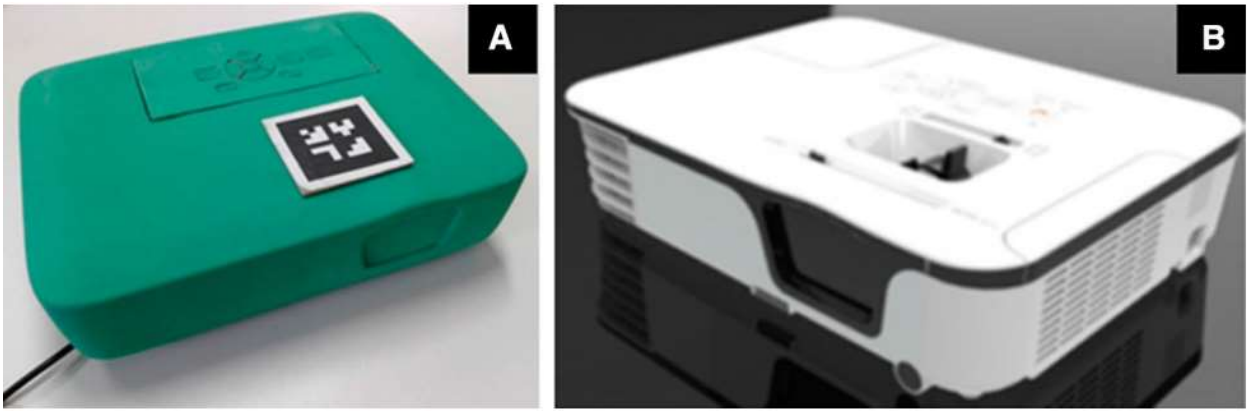

Figure 2. Reprinted stimuli from Faust et al, 2019. Left image shows the plastic model of the projector with no AR overlay to make it appear to be a projector. Right image shows the same model with AR overlay making it appear like a real projector, with a user interface appearing on the surface of the model. Buttons on the AR interface could be pressed and outcomes observed on the projection screen as though the plastic model were a functioning projector.

The mail system has been utilized in some user experience testing (Diamantidis et al., 2015). The product being tested was electronic and shown online (a medication inquiry system),

however the inputs to the test were pill bottles that were mailed to participants. This study was performed with participants low in health literacy. Two interfaces were tested, one on a mobile phone via text and the other on a PDA such as an iPod Touch. Participants entered information from the physical pill bottles into the electronic systems. Similar to this method, 3D Prototypes can be printed at low cost. Some services specialize in printing for the medical industry (e.g., stratasys.com). These prototypes can be mailed to users and paired with testing via videoconference or users filming themselves while carrying out the tasks. Data can include think-alouds and also provide insights on tactile interactions.

Eye-Tracking Software and Studies

For both flat and 3D interfaces/devices, eye-tracking is used by researchers studying medical devices (Koester, Brøsted, Jakobsen, Malmros, & Andreasen, 2017). Multiple online options exist, making data easy to collect provided the remote user has a webcam. A 2014 study showed similar results between webcam and traditional eye-tracking for “reasonably” sized images in the focal area and it is likely the technology has been improved and refined in the past six years (Burton, Albert, & Flynn, 2014). Unfortunately, the tracking is limited to the display, meaning that the medical device or interface must be shown in two dimensions. One of the earliest efforts took place in 2011, where the teleconferencing program Skype was paired with an eye tracking program to collect website usability data (Chynał & Szymański, 2011). Since then, remote eye tracking has exploded with commercial versions and academic or open-source versions (Table 1). Measures provided usually include videos of the gaze paths, heatmaps, and (less frequently) dwell time in areas of interest (AOIs).

Although use of online eye-tracking is a viable remote testing tool, use of wearable eye trackers will likely remain complicated. The cost of mobile systems, the difficulty of shipping them to enough participants (and receiving them back), sanitization during the pandemic, and the challenges for a participant to calibrate and record likely means their use would be reserved for testing devices with already highly trained and motivated experts (e.g., surgeons).

Recruitment Considerations for Patient-Facing Devices

The FDA encourages medical device manufacturers to include test participants who are "representative of the range of characteristics within their user group," with each group representing distinct user populations who will "perform different tasks or will have different knowledge, experience or expertise that could affect their interactions with elements of the user interface" (FDA, 2016). One advantage of remote usability testing is that individuals who cannot participate in laboratory testing due to high risk conditions preventing them from leaving their home can still be included in remote testing. Because of the importance of recruiting representative users for patient-facing devices, efforts put into finding and including these individuals should help to uncover usability issues that might otherwise have been missed. It is also easier for stakeholders to observe test sessions from distant geographic locations when testing is done remotely and to include more geographically diverse participant samples (Wiklund, Kendler, & Strochlic, 2016). Adhering to this guideline is critical for patient-facing devices, whose user population consists of highly heterogeneous chronic disease patients, dominated by high risk characteristics such as limited health literacy (Poureslami, Nimmon, Rootman, & Fitzgerald, 2017) and limited technological competence (Kruse, Karem, Shifflett, Vegi, Ravi, & Brooks, 2016). Also, many patient facing devices are *used* primarily "remotely" for disease self-management (e.g., glucometer) and must facilitate treatment in cases where direct physician supervision is not feasible (Greenwood, Gee, Fatkin, & Peeples, 2017). Unfortunately, patient recruitment and proportional representation in the design process is typically difficult and expensive due to population heterogeneity and recruitment barriers (Marquard & Zayas-Cabán, 2012). These barriers may be exacerbated when moving studies online.

Medical mistrust. Lower levels of trust in the medical system is well-documented, particularly among marginalized and socioeconomically disadvantaged populations (Benkert, et. al., 2019), who comprise a large portion of the chronic disease population. This impacts participation rates in studies, which may decrease when conducting studies in an unfamiliar

online format. In addition, recruitment efforts from a company or organization with whom participants are not familiar may fail. However, actively including trusted parties in the recruitment process (e.g., Primary Care Providers) may help to alleviate existing trust issues. The single most important factor affecting accrual is whether the patient's healthcare provider recommends that the patient participate in a particular study (Albrecht et al., 2008).

Health literacy. Health literacy refers to skills such as reading, writing, numeracy, communication, and the use of electronic technology (Güner & Ekmekci, 2019) that are necessary to make appropriate health decisions and navigate the healthcare system. To ensure representation of major user groups as required by FDA, it is recommended that patients are stratified based on expected health literacy, often assessed via an AHRQ health literacy survey tool (AHRQ, 2020), such as the Short Assessment of Health Literacy (Lee, Stucky, Lee, Rozier, & Bender, 2010) or Rapid Assessment of Adult Literacy in Medicine (Arozullah, et.a., 2007). Alternatively, patients with Medicare, Medicaid and no insurance are shown to have lower health literacy levels (NCES, 2006). Therefore, these groups can be recruited to target the low health literacy strata. Transferring usability studies that traditionally involved in-person interactions to online means the patient is responsible for adhering to study protocols, at times outside of the supervision of the study moderator, and will be more of a challenge than in-person studies.

Presenting literacy-level appropriate information that is linguistically and idiomatically aligned with the patient's needs is critical (Lopez, Sanchez, Killian, & Eghaneyan, 2018). It is widely recommended that printed information should not exceed a 7th or 8th grade reading level (Asiedu et al., 2020). These recommendations become even more critical in the context of remote usability studies. Other recommendations are to avoid medical jargon, use smaller and more manageable concrete steps to break down instructions, and assess comprehension (Hersh, Salsman, & Snyderman, 2015). Instructional videos, in comparison to textual information, have also been shown to be effective communication tools to increase memory

retention and patient satisfaction (Heinrich et al., 2019; Sharma et al., 2018). As important as these recommendations are for in-person studies, they will be even more critical for remote studies. Synchronous data collection would be preferred for lower health literacy participants, leveraging video conference and screen sharing technologies.

Technology access and skill level. It is recommended that a high proportion of persons with limitations or low language proficiency be recruited for formative medical device usability studies, as this will increase the number of use errors and increase accessibility of the final product (Wiklund, Kendler, & Strochlic, & 2016). In some cases, it may be easier to recruit these users and users with lower socio-economic status as they do not need to find transportation, child care, or use vacation time to attend a session. However, connectivity and internet access will remain a challenge for remote testing. While the use of digital technologies and internet access has become more widespread, a health disparity exists between young adults who predominantly use these tools and older adults who dominate the chronic disease population (Madrigal & Escofferey, 2019).

An additional barrier to online testing of patient-facing devices is the limited access to online resources and competence using technology. For example, chronic disease patients have low rates of online health information technology use despite the widespread availability (Ali et al., 2018), with substantial impact on the usability and acceptability of online testing platforms. Prior studies have shown that access to the internet and digital technologies affect a patient's willingness to use online services (Estacio, Whittle, & Protheroe, 2017). Given the existing barriers associated with technological competence in this older adult population, the move to remote testing, where technology is the sole platform for interaction, is expected to exacerbate these barriers. To mitigate issues with basic interactions (e.g., web navigation) or software installation, experimenters should provide resources for phone support prior to any online usability study.

Last, as with in person testing, there is an art to remote testing. Camera position, video and audio clarity, and good moderation are critical for detecting participant reactions in remote testing. Some of the reviewed solutions offer automated affect detection using facial analysis (e.g., EyeSee) but this would only be for tests with the participant looking at stimuli on the display rather than interacting with any physical object. Helping the participant to set up the test with the best lighting and angles possible for synchronous tests, and clear instructions for asynchronous tests will be needed to fully witness participant interactions.

Summary and Conclusions

We have summarized a variety of tools to conduct remote usability evaluations of medical devices and outlined important challenges. We provided tools appropriate for various research goals and ideas for extending other usability methods to remote use. We conclude that remote evaluation of medical devices is possible but challenging. Though some studies of cognitive and usability tasks suggest that results of remote tests are comparable to those in laboratory tests, such studies covered a limited range of tasks. Though a few researchers have attempted evaluations of 3-D devices, both virtually and physically, the literature is not strong enough for firm conclusions on comparison between remote and in-person testing. Fortunately, in many cases the type of data desired from usability tests (subjective assessments, qualitative impressions, learning) can be collected remotely, maintaining comparable data quality to lab-based testing. Last, aside from the technological hurdles, remote evaluations will require dedicated resources and attention toward recruiting representative users, who may be the most challenging to test online.

Acknowledging the challenges of moving studies to remote testing, we have created examples of remote testing using four medical device studies taken from the literature. Figure 3 provides four examples of how published medical device evaluation studies might be moved to an online format. The choices of platform and software are made based on the research goals of each study and the data needed to support those goals. Then, the remotely presented stimuli

are created to display on the chosen data collection platform. These studies were chosen to show the variety of research that may be moved online, from perceptual experiments, to mobile device interfaces, to 3D devices, and finally eye-tracking usability methods.

Remote usability testing is an emerging field that has the potential to increase efficiency of data collection. In addition it has the potential to allow access to user groups that are difficult to recruit if the correct precautions are taken. It is promising that initial work demonstrates equivalence between lab-based and remote testing, and that with the emergence of new approaches, remote testing can expand beyond subjective usability assessments.

Opportunities and tools for moving in-person usability studies of medical devices to remote testing. Laboratory studies and their research goals, shown left, are reimagined as online and remote studies that accomplish similar aims.

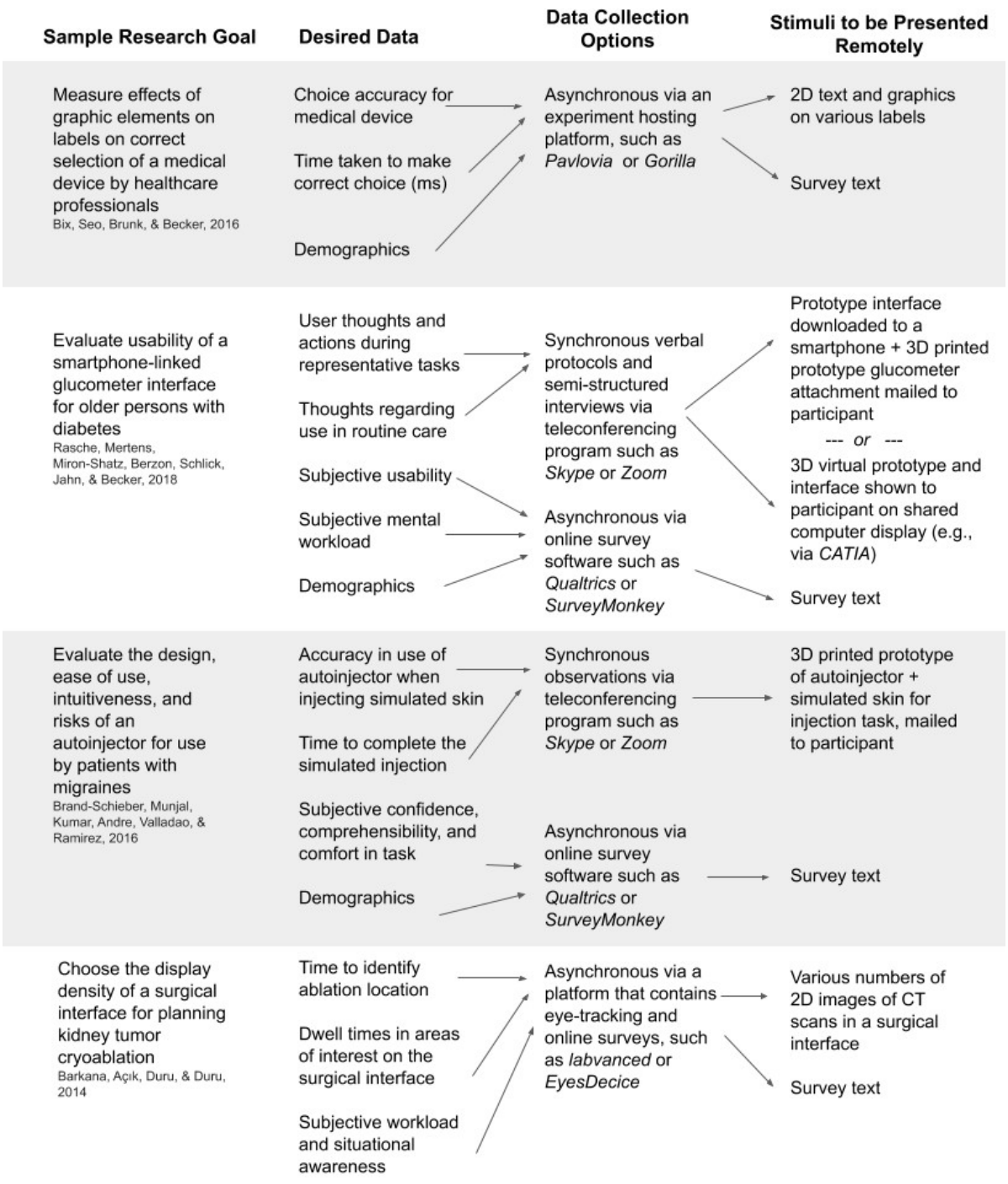

Figure 3. Example choices of remote tools or combinations of tools to meet research needs. Studies on the left were re-imagined as online, and tools that could provide the same or similar data are given. The types of stimuli that would be inputted into these tools is shown at right.

Key Points:

- Remote evaluation of medical devices will be necessary if the field is to progress during the restrictions of a pandemic.
- Many solutions are available, from those specialized to controlled experiments to those collecting qualitative data from a small number of participants.
- Novel attributes of some remote testing platforms include the ability to assess teams of participants, eye tracking, and enabling evaluation of 3D devices.
- Recruiting remote users with appropriate demographics to meet FDA obligations is expected to be more difficult than in person testing.
- We are cautiously optimistic that the tools for remote testing are at a point where medical devices can be design verified, with some able to be fully validated.

Acknowledgements

We are grateful to Richelle Huang for assistance with the literature review and to the reviewers for their helpful comments and suggestions.

References

Agency for Healthcare Research and Quality (2020, May 15), Health Literacy Measurement Tools (Revised), https://www.ahrq.gov/health-literacy/quality-resources/tools/literacy/index.html

Albrecht, T. L., Eggly, S. S., Gleason, M. E., Harper, F. W., Foster, T. S., Peterson, A. M., Orom, H., Penner, L. A., & Ruckdeschel, J. C. (2008). Influence of clinical communication on patients' decision making on participation in clinical trials. *Journal of Clinical Oncology*, *26*(16), 2666–2673.

Ali, S. B., Romero, J., Morrison, K., Hafeez, B., & Ancker, J. S. (2018). Focus Section health it usability: applying a task-technology fit model to adapt an electronic patient portal for patient work. *Applied Clinical Informatics*, *9*(1), 174–184.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 1405-1414). ACM.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388-407.

Arozullah, A., Yarnold, P., Bennett, C., Soltysik, R., Wolf, M., Ferreira, R., . . . Davis, T. (2007). Development and validation of a short-form, rapid estimate of adult literacy in medicine. *Medical Care, 45*(11), 1026-1033.

Asiedu, G. B., Finney Rutten, L. J., Agunwamba, A., Bielinski, S. J., St Sauver, J. L., Olson, J. E., & Rohrer Vitek, C. R. (2020). An assessment of patient perspectives on pharmacogenomics educational materials. *Pharmacogenomics*, *21*(5), 347–358.

Attridge, N., Noonan, D., Eccleston, C., & Keogh, E. (2015). The disruptive effects of pain on n-back task performance in a large general population sample. *Pain, 156*(10), 1885.

Barkana, D. E., Açık, A., Duru, D. G., & Duru, A. D. (2014). Improvement of design of a surgical interface using an eye tracking device. *Theoretical Biology and Medical Modeling, 11*(1). doi: 10.1186/1742-4682-11-S1-S4.

Belk, R. W., Caldwell, M., Devinney, T. M., Eckhardt, G. M., Henry, P., Kozinets, R., & Plakoyiannaki, E. (2018). Envisioning consumers: how videography can contribute to marketing knowledge. *Journal of Marketing Management, 34*(5-6), 432-458.

Benkert, R., Cuevas, A., Thompson, H. S., Dove-Meadows, E., & Knuckles, D. (2019). Ubiquitous yet unclear: A Systematic review of medical mistrust. *Behavioral Medicine*, *45*(2), 86–101.

Bix, L., Do Chan Seo, M. L., Brunk, E., & Becker, M. W. (2016). Evaluating varied label designs for use with medical devices: optimized labels outperform existing labels in the correct selection of devices and time to select. *PloS One, 11*(11).

Brand-Schieber, E., Munjal, S., Kumar, R., Andre, A. D., Valladao, W., & Ramirez, M. (2016). Human factors validation study of 3 mg sumatriptan autoinjector, for migraine patients. *Medical Devices, 9*, 131-137.

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1619-1628). ACM.

Burton, L., Albert, W., & Flynn, M. (2014). A comparison of the performance of webcam vs. infrared eye tracking technology. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), pp. 1437-1441). Sage CA: Los Angeles, CA: SAGE Publications.

Chynał, P., & Szymański, J. M. (2011). Remote usability testing using eye tracking. *IFIP Conference on Human-Computer Interaction* (pp. 356-361). Springer, Berlin, Heidelberg.

Clay, R. A. (2020). Conducting research during the COVID-19 pandemic. American Psychological Association. Retrieved on May 11, 2020 from https://www.apa.org/news/apa/2020/03/conducting-research-covid-19

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods, 40*, 428–434.

Devue, C., & Grimshaw, G. M. (2018). Face processing skills predict faithfulness of portraits drawn by novices. *Psychonomic Bulletin & Review, 25*(6), 2208-2214.

Diamantidis, C. J., Ginsberg, J. S., Yoffe, M., Lucas, L., Prakash, D., Aggarwal, S., ... & Fink, J. C. (2015). Remote usability testing and satisfaction with a mobile health medication inquiry system in CKD. *Clinical Journal of the American Society of Nephrology, 10*(8), 1364-1370.

El Guabassi, I., Bousalem, Z., Al Achhab, M., Mohajir, E. L., & Eddine, B. (2019). Identifying learning style through eye tracking technology in adaptive learning systems. *International Journal of Electrical & Computer Engineering* (2088-8708), 1-9.

Estacio, E. V., & Whittle, R., & Protheroe, J. (2017). The digital divide: Examining socio-demographic factors associated with health literacy, access and use of the internet to seek health information. *Journal of Health Psychology, 24*(12), 1668 - 1675.

Faust, F. G., Catecati, T., de Souza Sierra, I., Araujo, F. S., Ramírez, A. R. G., Nickel, E. M., & Ferreira, M. G. G. (2019). Mixed prototypes for the evaluation of usability and user experience: simulating an interactive electronic device. *Virtual Reality, 23*(2), 197-211.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: a unified JavaScript framework for online studies. *International Conference on Computational Social Science (Cologne).*

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847-857.

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is where the lab is: a comparison of online and lab data from a time-sensitive study of interruption. *Human Computation, 2*. 45-67.

Greenwood, D. A., Gee, P. M., Fatkin, K. J., & Peeples, M. (2017). A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. *Journal of Diabetes Science and Technology*, *11*(5), 1015-1027.

Güner, M. D., & Ekmekci, P. E. (2019). A survey study evaluating and comparing the health literacy knowledge and communication skills used by nurses and physicians. *Inquiry : a Journal of Medical Care Organization, Provision and Financing*, *56*, 1-10.

Heinrich, K., Sanchez, K., Hui, C., Talabi, K., Perry, M., Qin, H., Nguyen, H., & Tatachar, A. (2019). Impact of an electronic medium delivery of warfarin education in a low income, minority outpatient population: a pilot intervention study. *BMC Public Health*, *19*(1050), 1-7.

Hersh, L., Salzman, B., & Snyderman, D. (2015). Health literacy in primary care practice. *American Family Physician, 92*(2), 118 - 124.

Kamphuis, W., Essens, P. J., Houttuin, K., & Gaillard, A. W. (2010). PLATT: A flexible platform for experimental research on team performance in complex environments. *Behavior Research Methods, 42*(3), 739-753.

Kanai, S., Higuchi, T., & Kikuta, Y. (2009). 3D digital prototyping and usability enhancement of information appliances based on UsiXML. *International Journal on Interactive Design and Manufacturing (IJIDeM), 3*(3), 201-222.

Kantowitz, B. H. (1988). Laboratory simulation of maintenance activity. *Proceedings of the 1988 IEEE 4th Conference on Human Factors and Power Plants.* IEEE.

Koester, T., Brøsted, J. E., Jakobsen, J. J., Malmros, H. P., & Andreasen, N. K. (2017). The use of eye-tracking in usability testing of medical devices. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, 6*(1), 192–199. https://doi.org/10.1177/2327857917061042

Kruse, C., Karem, P., Shifflett, K., Vegi, L., Ravi, K., & Brooks, M. (2018). Evaluating barriers to adopting telemedicine worldwide: A systematic review. *Journal of Telemedicine and Telecare*, *24*(1), 4–12.

Kuutti, K., Battarbee, K., Sade, S., Mattelmaki, T., Keinonen, T., Teirikko, T., & Tornberg, A. M. (2001). Virtual prototypes in usability testing. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences* (pp. 1-7). IEEE.

Lange, K., Kühn, S., & Filevich, E. (2015). " Just Another Tool for Online Studies"(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One, 10*(6).

Lee, S. Y., Stucky, B. D., Lee, J. Y., Rozier, R. G., & Bender, D. E. (2010). Short assessment of health literacy-spanish and english: a comparable test of health literacy for spanish and english speakers. *Health Services Research*, *45*(4), 1105–1120.

Li, H., & Salah, H. (2017). Comparing four online symptom checking tools: preliminary results. *IIE Annual Conference Proceedings* (pp. 1979-1984). Institute of Industrial and Systems Engineers (IISE).

Lopez, V., Sanchez, K., Killian, M. O., Eghaneyan, B. H. (2018). Depression screening and education: an examination of mental health literacy and stigma in a sample of Hispanic women. *BMC Public Health*, 18(646), 1-8.

Madrigal, L., & Escoffery, .C. (2019). Electronic health behaviors among us adults with chronic disease: cross-sectional survey. *Journal of Medical Internet Research, 21*(3).

Maiti, A., Maxwell, A. D., & Kist, A. A. (2017). Using marker based augmented reality and natural user interface for interactive remote experiments. *4th Experiment@International Conference (exp.at'17) (*pp. 159-164). doi: 10.1109/EXPAT.2017.7984396.tracking

Marquard, J. L., & Zayas-Cabán, T. (2012). Commercial off-the-shelf consumer health informatics interventions: Recommendations for their design, evaluation and redesign. *Journal of the American Medical Informatics Association*, *19*(1), 137–142.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavioral Research Methods, 44,* 314–324.

Mejía-Gutiérrez, R., Carvajal-Arango, R. (2017). Design Verification through virtual prototyping techniques based on Systems Engineering. *Research, Engineering, and Design, 28,* 477–494. https://doi-org.prox.lib.ncsu.edu/10.1007/s00163-016-0247-

Mottelson, A., & Hornbæk, K. (2017). Virtual reality studies outside the laboratory. *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (pp. 1-10). ACM.

National Center for Education Statistics. (2006). *The Health literacy of america's adults: results from the 2003 national assessment of adult literacy.* U.S. Department of Education.

Office of the National Coordinator for Health Information Technology (ONC). (2015). *Certification Companion Guide: Safety-enhanced design.* Retrieved 1 July 2020 from https://www.healthit.gov/test-method/safety-enhanced-design

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam eye tracking using user interactions, *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI),* (pp. 3839--3845). AAAI.

Poureslami, I., Nimmon, L., Rootman, I., & Fitzgerald, M. J. (2017). Health literacy and chronic disease management: drawing from expert knowledge to set an agenda. *Health Promotion International*, *32*(4), 743–754.

Potthoff, J., Face, A. L., & Schienle, A. (2020). The color nutrition information paradox: Effects of suggested sugar content on food cue reactivity in healthy young women. *Nutrients, 12*(2), 312.

Rakkolainen, I., Raisamo, R., Turk, M. A., Höllerer, T. H., & Palovuori, K. T. (2016). Casual immersive viewing with smartphones*. Proceedings of the 20th International Academic Mindtrek Conference* (pp. 449–452). https://doi-org.prox.lib.ncsu.edu/10.1145/2994310.2994314

Rasche, P., Mertens, A., Miron-Shatz, T., Berzon, C., Schlick, C. M., Jahn, M., & Becker, S. (2018). Seamless recording of glucometer measurements among older experienced diabetic patients–A study of perception and usability. *PloS One, 13*(5).

Sauer, J., Sonderegger, A., Heyden, K., Biller, J., Klotz, J., & Uebelbacher, A. (2019). Extra-laboratorial usability tests: An empirical comparison of remote and classical field testing with lab testing. *Applied Ergonomics, 74*, 85-96.

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences, 10*(4), 251.

Servick, K., Cho, A., Guglielmi, G., Vogel, G., & Couzin-Frankel, J. (2020) Updated: labs go quiet as researchers brace for long-term coronavirus disruptions. *Science Magazine.* Retrieved 10 May, 2020 from https://www.sciencemag.org/news/2020/03/updated-labs-go-quiet-researchers-brace-long-term-coronavirus-disruptions#

Sharma, S., McCrary, H., Romero, E., et al. (2018). A prospective, randomized, single-blinded trial for improving health outcomes in rhinology by the use of personalized video recordings. *International Forum on Allergy and Rhinology, 8*(12), 1406–1411.

Thompson, K. E., Rozanski, E. P., & Haake, A. R. (2004). Here, there, anywhere: remote usability testing that works. *Proceedings of The 5th Conference on Information Technology Education* (pp. 132-137). ACM.

Von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods, 45*(1), 108-115.

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ, 3*, e1058. doi: 10.7717/peerj. 1058.

Biographies

All authors are members or partners of the hfMEDIC consortium (hfMEDIC.org), a partnership of academic and industry researchers dedicated to developing safer and more effective medical devices through human-centered design.

Anne Collins McLaughlin is currently a Professor in the Department of Psychology at North Carolina State University in Raleigh, NC. She earned her PhD in psychology in 2007 from the Georgia Institute of Technology. Her research interests include the study of individual differences in cognition, particularly those that tend to change with age, applied to various domains including medical device design.

Patricia R. DeLucia is currently a Professor in the Department of Psychological Sciences at Rice University in Houston, TX. She earned her PhD in psychology in 1989 from Columbia University. Her research interests include the human factors of health care (minimally-invasive surgery, telehealth, medication administration, patient safety, and medical device design).

Frank A. Drews is currently a Professor in the Department of Psychology at the University of Utah. He earned his PhD in psychology in 1999 from the Technical University of Berlin, Germany. His research interests include the design of medical devices and displays for healthcare providers and patients.

Monifa Vaughn-Cooke is currently an Assistant Professor in the Mechanical Engineering Department at the University of Maryland, College Park. She earned her PhD in industrial engineering in 2012 from The Pennsylvania State University. Her expertise is in the area of

human factors and healthcare, with a focus on improving human performance for medical device interaction.

Anil Kumar is currently an Associate Professor in the Industrial and Systems Engineering Department at San Jose State University. He earned his PhD in industrial engineering from Western Michigan University in 2007. His areas of specialty include product design and development (medical products and healthcare), ergonomics, human factors, work measurement and analysis, and safety.

Robert Nesbitt is currently the Director of Human-Centered Design and Human Factors at AbbVie, Chicago, Illinois. He has worked across a number of domains in industry, from Deere & Co. to Eli Lilly. In his role at AbbVie, he focuses on early-stage ethnographic or in-context understanding of patients' and users' needs, particularly for the design of combination medical products.

Kevin Cluff PhD PE, is Principal Consultant at BioWork Engineering, specializing in research and human factors for late stage combination products. Prior to BioWork, Kevin was a Principal Research Engineer for 16 years at a major biopharmaceutical company where he was responsible for HF studies and documentation of many successful FDA submissions.