NExTRAC WORKING GROUP ON DATA SCIENCE AND EMERGING TECHNOLOGY:

Progress on Phase 1 of Current Charge

Update to NExTRAC

07.14.22

Contents

NExTRAC Charge and Context

NIH seeks to understand how emerging technologies may enable the combination and use of human datasets, particularly from disparate sources (e.g., research and non-research settings), in an effort to anticipate potential benefits and risks for research participants, families, populations, and society. Ultimately, the goal is to assist NIH in developing research strategies and policy frameworks, informed by stakeholder input, to ensure that research progresses responsibly and meets expectations for the Nation's investment in research. Thus, the NExTRAC is charged to:

Phase 1:

- Define and characterize the types of research questions that require increasing granularity and aggregation of data about individuals that are likely to be addressed through emerging technologies. Please consider but do not limit the scope to:
    - Goals of such research studies and how they advance the NIH mission;
    - Emerging technologies that may generate potentially sensitive datasets;
    - Data types generated and their sources (e.g., digital health devices, EHR platforms) with an emphasis on exploring new data types or unique sources; and
    - Data science platforms and tools that facilitate data access, combination, and analysis (e.g., artificial intelligence, cloud computing).

Phase 2:

- For those questions and technologies defined above, consult with stakeholders to discuss and assess the value of and potential implications for individuals, groups (i.e., families, specific regions or populations), and society. Please consider, but do not limit the scope to:
    - Attitudes and perspectives about sharing participant data to advance biomedical research, specifically through the lens of balancing research risk (e.g., privacy, autonomy) with research deliverables; and
    - How these perspectives may evolve depending on the context of who is to benefit or assume risk, whether it be at the individual level, through the community, or broader society's expectations for public health advancement.

In addressing this charge, the working group shall convene consultations with stakeholders including, but not limited to, research participants, patient groups, ethicists and privacy experts, data scientists, technology developers across sectors, and public health officials.

Background and Overarching Topics

Over the course of Phase 1 of the effort to address the NExTRAC's charge (November 2021 – June 2022), the working group (WG) developed a draft list of types of research questions. While the initial set of issues considered was expansive, the group narrowed its focus to three major topics for informing health-related questions:

- **Data**: Use of novel data from outside of the traditional healthcare system
- **Algorithms**: Use of models and algorithms
- **Integration**: Data linkage and aggregation of disparate datasets from multiple sources

The WG acknowledges that there are many other areas of data science research that pose emerging issues, but this list of topics seems most salient to address the charge to the NExTRAC. Additionally, the WG notes that there is some overlap between these topics. The draft list of types of research questions that follows is organized by these topics, with general types of research questions as sub-bullets and exemplar research questions, including citations, to illustrate each type of research question.

Draft List of Types of Research Questions

The research questions below are those the WG has identified through its deliberations, literature reviews, and discussions with subject matter experts, as most responsive to meet Phase 1 of the charge. For these research questions, the potential implications for individuals, groups, and society of sharing such data through the lens of balancing research risk with research deliverables will be explored in Phase 2 discussions.

Types of research questions/major topics for informing health-related questions:

Topic 1: Use of novel data from outside of the traditional healthcare system

1. How are personal health data collected from outside of the traditional health system (wearables, fitness trackers, apps (e.g., period tracker apps), social media posts) being used to study health-related questions and predict health risks, at either an individual, family, group, or public health level? (Note that implications for privacy, consent, and other important principles will be explored in Phase 2.)
    - **Example research questions pertaining to social media data (often publicly available):**
        - Can patient-generated digital data from Facebook predict or detect relapse in psychiatric disorders? (Birnbaum et al., 2019)
        - Can real-time streams of secondary information related to suicide (e.g., tone of language used in posts, affiliation to particular user groups) be used to accurately estimate suicide fatalities in the US in real time? (Choi et al., 2020)
    - **Example research questions pertaining to wearables (often data are presumed private, but where the data are stored, who owns the data (raw and/or processed) may indicate otherwise):**
        - Can wearable microphones paired with accelerometers provide reliable long-term cardiopulmonary monitoring of patients? (Gupta et al., 2020)
        - Can chest-worn inertial sensors accurately detect a broad range of physiological signals (e.g., cardiac and respiratory parameters) and behaviors (posture, sleep, falling, swallowing) for real time tracking? (Rahmani, Berkcens, Weyn, 2021)
        - Can interstitial glucose levels be measured accurately and precisely through wearable technology for clinical decision making? (Cengiz and Tamborlane, 2009)

2. How can other consumer and lifestyle data from non-health-specific sources (e.g., credit card and consumer rewards data, sensors in the home) be used to study health-related questions and predict health risks? (Note that implications for privacy, consent, and other important principles will be explored in Phase 2.)
    - **Example research questions:**

- Can credit scores define cardiovascular disease risk? (Israel et al., 2014)
- Can personal AI assistants (e.g., Alexa) reliably detect health conditions through changes in speech patterns? (Anthes, 2020)
- Can in-home sensors be leveraged to create safer environments for people suffering from cognitive decline? (Vahia et al., 2020)

3. Can/how can integration of health data with data on the status of the social determinants of health (SDOH) enable better risk stratification of patient populations and better development of predictive algorithms (Cantor and Thorpe, 2018)? SDOH can include socio-economic status, housing status, education status, geographical environments in which people spend time (e.g., crime rates or environmental pollutants in a given neighborhood), and identity factors that advantage or disadvantage health status. (Note that implications for privacy, consent, and other important principles will be explored in Phase 2.)
    - **Example research questions:**
        - Can screening of the status of SDOH in electronic medical records enable tailored referrals to available community-based agencies? (Gottlieb et al., 2015)
        - Can integration of SDOH data into electronic health records facilitate clinical risk prediction and intervention? (Chen et al., 2020)

Topic 2: Use of models and algorithms

4. What is the role of computer-based technologies, such as artificial intelligence (AI), machine learning (ML), and automated image analysis, in advancing health decision-making?
    - **Example research questions:**
        - Can deep learning algorithms (i.e., algorithms that update themselves based on new information) be deployed for effective automated medical image analysis to replace clinical expertise for diagnosis of disease? (Pneumonia; Irvin et al., 2022), (Stomach, Intestinal Cancer; Braatz et al., 2022)
            - Does bias in algorithm development impact populations differentially by race, gender, and/or social/cultural identity? (Larrazabal, 2020)
        - When and how should human experts be inserted to monitor and analyze the output of health-related algorithms and models to ensure their fidelity? (Overview; Holzinger, 2016; Budd, Robinson, Kainz, 2021), (Clinician Expert; Boden et al., 2021), (Bioinformatics supervision; Goodwin and Demner-Fushman, 2020)
        - Can algorithms determine prognoses for patients on admission to the hospital? (Kamran et al., 2022)
        - How can the underlying computational strategies used by AI/ML algorithms be tracked and corrected for ensuring accuracy in clinical diagnoses? (Geirhos et al., 2020; While these algorithms can find solutions and provide

diagnoses based on the data that they have been given, it is often not clear how they came to those conclusions. Algorithms may be responding to spurious correlations or taking "shortcuts" to identify patterns in data.)

- What types of bias are involved in algorithms taking "shortcuts" to reach conclusions? (Wiens et al., 2020)

5. Can/how can natural language processing be deployed to analyze data held in health systems (e.g., Electronic Health Record (EHR), health insurance data, data from pharmacies) to provide insights about patient symptoms and disease classification?
    - **Example research questions:**
        - What symptom information can be found in unstructured narratives in EHRs? (Koleck et al., 2019)
            - How might large-scale language models exacerbate existing biases in medical research and clinical care? (Weidinger et al., 2021)
            - How do the identities and cultural backgrounds of doctors and patients impact the information derived using natural language processing of unstructured data?

Topic 3: Data linkage and aggregation of disparate datasets from multiple sources

6. Are there opportunities to standardize data formats – or deploy standardizing technologies – so that data from different countries and healthcare systems could be aggregated, linked, and shared across populations?
    - **Example research questions:**
        - Can large-scale observational research (enabled by standardizing data formats across countries and healthcare systems) outperform established guidelines and expert opinion? (OHDSI; Schuemie et al., 2020)
        - To what extent can the interoperability standards currently in use be applied for clinical decision support? (Taber et al., 2021)

7. Which disparate (and potentially conflicting) data sets (e.g., genomics, proteomics, clinical information, clinical imaging) can be linked and combined with harmonized (automated) data aggregators?
    - **Example research questions:**
        - Can multi-modal cancer data (e.g., EHR data, genomic data, health imaging data) be meaningfully pooled from numerous sources to improve its usefulness to the broader cancer research community? (NeuroLINCS)

8. Can/how can personal health libraries be used to combine individuals' health information across multiple different data streams to inform health outcomes?
    - **Example research questions:**

    - How can all of the behavioral data about a person's health and lifestyle be combined and managed by the individual to provide an accurate picture of their health (a "digital twin")? Can such digital twins be used to improve healthcare service for individuals? (Liu et al., 2019)

9. How can Privacy Preserving Record Linkage (PPRL; a method for integrating data while maintaining the security of privacy information) be used to combine data on individuals from multiple sources and with different identifiers for precision medicine and public health?
    - **Example research questions:**
        - Can PPRL identify new research opportunities for previous human research participants without sacrificing privacy?
        - Can PPRL safely combine public health data and clinical data without putting individuals at risk?

10. How should the research context (e.g., clinical, public health) and participants' consent status affect data linkage and aggregation?