Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1107X

Title: Virtual Clinical Trial of Contrast-Enhanced Tomosynthesis for Screening of Dense Breasts

Creator: Paul Kinahan - **ORCID:** 0000-0001-6461-3306

Affiliation: University of Washington (uw.edu) (washington.edu)

Principal Investigator: Cynthia Davis, Paul E Kinahan

Data Manager: Cynthia Davis, Paul E Kinahan

Project Administrator: Paul E Kinahan

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-21-166

Template: NIH-Default DMSP

Project abstract:

Contrast-enhanced (CE) imaging has demonstrated improvements for breast cancer detection in women with dense breasts. While CE-MRI is potentially a gold standard, it is not accessible as a widespread screening approach, so a priority for screening is the evaluation of contrast enhancement with 2D mammography (CE-M) and 3D digital breast tomosynthesis (CE-DBT). To compare the relative effectiveness of CE-M and CE-DBT in a timely and feasible manner, we are proposing an in silico 'virtual imaging clinical trial' that uses population-based and realistic models of breast anatomy, cancer biology, physics models of the image acquisition processes, and models of human observers.

Start date: 07-01-2024

End date: 06-30-2029

Last modified: 08-31-2023

Copyright information:

Virtual Clinical Trial of Contrast-Enhanced Tomosynthesis for Screening of Dense Breasts

Data Type

Types and amount of scientific data expected to be generated in the project: ***Summarize the types and estimated amount of scientific data expected to be generated in the project.***

Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

The primary type of scientific data generated (that will not be captured in publications) are synthetic medical images. The images will be stored the industry-standard Digital Imaging and Communications in Medicine (DICOM) standard. The synthetic images will be representations of contrast-enhanced mammography and contrast-enhanced digital breast tomosynthesis images.

Based on the VICTRE trial [Badano A, et al. JAMA Network Open. 2018], we expect to generate on the order of 30,000 contrast-enhanced mammography and contrast-enhanced digital breast tomosynthesis images (each) from approximately 3,000 virtual patients. E.g. 100 images of both types for each virtual patient.

The DICOM images will be approximately 50MB in size for each contrast-enhanced mammogram and 1 GB for a digital breast tomosynthesis image. The DICOM images will be stored on the Zenodo server at CERN.

Zenodo is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN. For each data set, a persistent digital object identifier (DOI) will be created, which makes the stored items easily citeable and searchable. Zenodo addresses the FAIR Principles definition as referenced from: Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Scientific data that will be preserved and shared, and the rationale for doing so: ***Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.***

The primary type of scientific data generated are synthetic mammography and digital breast

tomosynthesis images. The images will be stored the industry-standard Digital Imaging and Communications in Medicine (DICOM) standard. This unique data, with ground truth metadata, will be made available for other researchers for developing and/or testing image processing algorithms to be used with mammograms and digital breast tomosynthesis images.

Metadata, other relevant data, and associated documentation: Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

To facilitate the interpretation and reuse of the data, a README file and data dictionary will be generated and deposited into a repository along with all shared datasets. The README file will include synthetic patient and lesion information, , scanner descriptions, and acquisition protocol. The data dictionary will define and describe all variables in the dataset.

Related Tools, Software and/or Code

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

The image data generated via the simulations will be in the Digital Imaging and Communications in Medicine (DICOM) format, which is well-defined and widely adopted. Thus free and commercial tools to read and view the images are readily available.

Standards

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

Question not answered.

Data Preservation, Access, and Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived.

Imaging data and metadata will be deposited into Zenodo.

How scientific data will be findable and identifiable: Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

We will use Persistent Unique Identifiers (PIDs) to improve data findability across all dissemination outputs. PIDs used will DOIs for outputs (e.g., datasets, protocols), Research Resource IDentifiers (RRIDs) for resources, and Research Organization Registry (ROR) IDs and funder IDs for places, as much as possible to make data identifiable and findable. We will also use indexed metadata, such as MeSH terms with a unique URL to make scientific data easily findable. We will keep our records up to date with DOIs for our datasets and publications, ROR, and funder IDs to increase findability.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

All scientific data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 10 years after the funding period.

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing.

There are no factors that will impact access, distribution, and reuse for all other scientific data generated by this study.

Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

Access to scientific data will not be controlled.

Protections for privacy, rights, and confidentiality of human research participants: If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

Question not answered.

Oversight of Data Management and Sharing

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

The schema for metadata and storage of the images data will be based on the core Digital Imaging and Communications in Medicine (DICOM) standard, which will be the format the images are stored in. The schema will be proposed and reviewed by the leadership group. Once the schema is finalized, sample data will uploaded to Zendo to test adherence to the FAIR principles, and the schema will be modified if necessary.

As the research data sets are created, they will undergo a two step process between the University of Washington adn GE-Research teams. A member from one team will upload the data set, and a member of the other team will verify that that data are findable, downloadable, and correct.
