Federal Register Notice: 89 FR 51554, Federal Register :: Networking and Information Technology Research and Development Request for Information on Digital Twins Research and Development, June 18, 2024.

Request for Information on the National Digital Twins R&D Strategic Plan

Jineta Banerjee, Ann Novakowski, Milen Nikolov
Institution: Sage Bionetworks

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.

RFI Response: Digital Twins R&D Plan

Responders: Jineta Banerjee, Ann Novakowski, Milen Nikolov
Institution: Sage Bionetworks

Disclosure statement: This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in the National Digital Twins R&D Strategic Plan and associated documents without attribution.

Response:

We appreciate NITRD's efforts towards collecting information about the status and needs for digital twins in various disciplines of science. While widely utilized in the aerospace and automobile industry to test innovations safely, digital twins' adoption is emerging yet limited in healthcare. In healthcare, drug development research remains a slow process due to limited understanding of disease biology and the need for extensive and labor-intensive clinical trials for drug validation. Computational approaches using digital twins to improve trial design or drug screening could revolutionize drug development research. With the increase in high throughput data generation in medicine and biomedical research and recent advances in multi-modal generative models, we are increasingly optimistic about the era of biomedical digital twins. Our response here will focus on addressing the following focus areas: *Data, Standards, Trustworthy.*

Lack of access to adequate and appropriate data is a considerable challenge for development as well as deployment of digital twins in medicine. Models that can generate digital twins of patients for various diseases will require training a combination of mechanistic, AI-based generative, and forecasting models. Developing such models will require a large amount of unbiased data from various scales of biology - cellular-level, tissue-level, and organism-level. It is unlikely that one medical or research organization will have all the required data and expertise for these models. So we foresee the need for coalitions of researchers from biology, medicine, AI, and physical sciences to generate such models based on data that are generated across institutions and disciplines as well as across a multitude of patient populations. To facilitate the cross-institutional research and development, large organized data management efforts will need to be in place for the success of such cross-institutional coalitions.

The nascency of digital twins research provides a unique opportunity to proactively develop purposeful tools and platforms to ensure transparent and patient-centered digital twin development and seamless data transfer between data-generating groups and data-using groups (AI/ML researchers or model generators). In our opinion, data management for digital twins would require considerations at two stages. One at the level of model generation, and second at the level of twin generation.

Model generation phase:

<u>Platform for data sharing and improvement:</u>

1) A cloud-based, scalable, institution-agnostic, and platform-agnostic data management

system will be needed that can ingest large amounts of data from multiple institutions and prepare them for egress as needed.
2) Digital twins and models generating the twins will need continuous updating and inclusion of new data modalities as necessary. So platforms that can provide an avenue for iterative feedback between data collectors and model generators will be needed to ensure collection of appropriate data and data modalities for model improvement.
3) Optionally, the ability to provide continuous benchmarking of models on gold standard datasets to examine the accuracy of models would be important in providing transparency for expectations in model output.
4) Platforms that can provide metrics on the level of disparity between data generated from different institutions and ability to harmonize data across institutions will enormously accelerate digital twin research.

Standards for data and metadata for model generation:
1) The data used in model generation would need to be quality checked and prepared to meet the FAIR standards with as much detailed metadata as possible to account for any and all nuances and biases of data capture.
2) Tools that provide easy ways to annotate data files with appropriate metadata leveraging automated capture from LIMS or electronic laboratory notebooks in addition to manual addition by researchers will become important to approach scalability while maintaining quality of data.
3) The data would likely come from various scales, i.e. cellular-level, tissue-level, and organism-level. Currently available data models would need to be enhanced to accommodate successful linking of such multiscale data.

Twin generation phase:
Once models to generate digital twins are developed, additional data management considerations would be required to facilitate generation and storage of digital twins of individual patients.

Such platforms should include the:
1) Ability to connect to clinical sites to enable data ingestion for individual patients.
2) Ability to continuously integrate newly acquired data with existing data from each patient.
3) Ability to preserve provenance of data from patients to the users generating digital twins to improve transparency of data use
4) Ability to store and update digital twin data using unique patient identifiers while being HIPAA compliant.

Data governance for model training and digital twins:
Since digital twins cannot be completely deidentified, special attention needs to be given to data governance. Robust governance frameworks will be essential to prevent privacy breaches, preserve data context, and minimize misuse or exploitation, e.g., using data to approve or reject health insurance claims or making generalized predictions that could harm specific groups. Adherence to data minimization principles would help mitigate these risks. Access to data and digital twins would need to be strictly controlled, with release limited to authorized parties in a secure environment. This access may be tailored to the stages of twin development, including data collection, twin exploration, and twin deployment to maintain compliance with research objectives and ethical guidelines.

Additionally, the digital twins themselves would need to be subject to control measures that protect the privacy and interests of the individuals they represent. A dynamic attribution and consent process will enable research participants to provide informed consent for twin deployment and monitor the status of their data in digital twin studies. The concept of “digital dignity” may be gaining traction as public awareness grows regarding the ubiquity of personal data collection and its uses in tracking, marketing, and other potentially invasive applications. Extending principles of digital dignity with unbroken data provenance to research participants would enable them to monitor the use of their data in current and future studies. Ethical, Legal, and Social Implications (ELSI) frameworks should also be considered in the return of results from digital twin studies. While this type of transparency could enhance model validity and clinical reliability in research outcomes when shared with clinical care teams, the insights gained would be carefully balanced against the potential benefits and risks to the participants.

Standards for models:

Since digital twins are as good as the generative models are, implementing specific standards for describing and deployment of such models will be key. We expect to see emergence of digital twin model repositories in conjunction with repositories for the data and twins. Such repositories would need to implement standards for describing models to make them findable and accessible. Special care should be taken to define model parameters including the accepted range of values and units of measurement. Model metadata should also include whether model parameters are cell-level, tissue-level, or organism level. Additionally, it would be important to surface metrics that measure congruence between specific parameters for real patients and those predicted for digital twins to provide transparency about the strengths and weaknesses of the models and the twins. We also expect that containerization of models and ability to be deployed by users other than the model generators will be encouraged to enable testing generalizability of the models.

Trustworthiness of digital twins:

Given that drug discovery and clinical decision support will be among the most important use cases of digital twins, special care needs to be taken to ensure the trustworthiness of the twins. For any data generated and used for model generation, special care should be taken to define and surface metrics regarding data quality and harmonization and should be updated continuously. Digital twin data predicted by these individualized models is generally accompanied by uncertainty of prediction. Such uncertainty metrics should be documented and surfaced adequately to prevent misinterpretation of the twin data.