

Sage Bionetworks will build the data platform for a new federal program to accelerate rare disease diagnosis

The Rare Disease Data Commons will unify fragmented patient data, establish privacy-protected benchmark datasets, and run community AI challenges

SEATTLE, WA, UNITED STATES, August 31, 2026 /EINPresswire.com/ -- The average rare disease diagnosis takes six years. For some families it takes decades. More than 10,000 rare diseases affect roughly 30 million Americans, cost the U.S. health system an estimated \$1 trillion a year, and yet only about 5 percent have an approved treatment.

“

Rare disease data are locked into islands. A foundation typically funds one disease, collects one data type, and the result doesn't connect to what the next foundation collected for the next disease.”

Robert Allaway, PhD

Sage Bionetworks is partnering with the Advanced Research Projects Agency for Health (ARPA-H) to build the Rare Disease Data Commons (RDDC), the central data repository and benchmarking platform for ARPA-H's new Rare Disease AI/ML for Precision Integrated Diagnostics (RAPID) program. RAPID is awarding Sage Bionetworks up

to \$28 million over 4.5 years to advance the program's mission to transform rare disease diagnosis—helping patients get answers sooner while reducing costs for families and the health system.

ARPA-H's RAPID program is organized into four technical areas. The first focuses on assembling a large-scale longitudinal clinical dataset from electronic health records. The second technical area will collect patient-reported data and other modalities: genomics, imaging, video, voice, and wearable data. Sage's team leads the third technical area, building the platform that takes data in from both, harmonizes it, governs access, and benchmarks the diagnostic AI built on top of it. A fourth area, will validate tools in clinical settings. Patient experience partners work across all four areas throughout the program.

The primary obstacle the RDDC addresses is not a shortage of AI methods, rather, it is a shortage of usable data. Rare disease data sits in separate institutions, coded differently, and governed under terms that have to be renegotiated every time a new research group wants to work with it.

"Rare disease data are locked into islands. A foundation typically funds one disease, collects one data type, and the result doesn't connect to what the next foundation collected for the next disease." says Robert Allaway, PhD, Director of Rare Disease at Sage Bionetworks and Principal Investigator (PI) for the project. "We proposed a platform that connects these islands into an archipelago, solving data challenges shared across the rare disease community once instead of ten thousand times."

The RDDC will contain five connected modules that bring in data from the other RAPID teams and from outside contributors, make it consistent enough to compare across sources - building on tools created for the ARPA-H Biomedical Data Fabric Toolbox - and set clear rules for who can use it and how. Where data cannot be centralized, RDDC supports federated integration and interoperates with other trusted research environments rather than becoming another silo. "People living with rare diseases often drive their care odyssey, working closely with clinicians and researchers to gain insight into their diagnosis, connect to a community of others with similar conditions, and understand their prognosis and treatment options." says Milen Nikolov, PhD, Director of Product Innovation at Sage Bionetworks and co-PI for the project. "The RDDC modules advance breakthrough technology - translating disjoint, multi-modal datasets into reliable information usable by patients, researchers and clinicians - throughout this odyssey and across all rare diseases."

"For rare diseases, where no individual dataset may ever be large enough, our greatest opportunity is to learn across boundaries of individual rare diseases. In the Rare Disease Data Commons, we aim to meaningfully connect datasets from multiple rare diseases to allow advanced computational methods to uncover shared and latent patterns that would help improve diagnosis or identify novel therapeutic opportunities." says Jineta Banerjee, PhD, Associate Director of Advanced Data Analytics at Sage Bionetworks and co-PI for the project.

The most publicly visible deliverable is the Rare Disease Benchmark Dataset: a versioned, broadly accessible resource designed to give researchers and developers working on diagnostic AI a common, credible basis for evaluation. Today, groups developing rare disease AI methods often evaluate on proprietary data, making it difficult to compare results or know which methods are worth pursuing clinically. The benchmark would enable standardized evaluation across the field while maintaining strong protections for patient privacy through de-identification, governed access, and appropriate data-use controls. "It's incredibly challenging to make apples-to-apples comparisons between AI tools for rare diseases," says Allaway. "A group builds a model, reports on its own cohort, and there is no way to tell whether the next method is better or just due to differences in the test data. We are building a place where the evaluation is the same for everyone."

"About one in ten Americans lives with a rare disease, many of them are still searching for answers, and they have been waiting a long time for infrastructure built to their scale," says Luca Foschini, PhD, President & CEO of Sage Bionetworks. "Building in the open while protecting

privacy is a complex undertaking, but it's the only way we can create trust and share learnings for the benefit of all patients, rare disease and beyond."

Sage Bionetwork develops and operates Synapse, one of the NIH-listed generalist repositories, managing more than 4.5 petabytes of biomedical research data. The RDDC builds on 15 years of Sage's experience developing governed data commons for sensitive biomedical data. Sage is building the Rare Disease Data Commons with Cirro Bio (<https://cirro.bio/>), Netrias (<https://www.netrias.com/>), the Gyori lab at Northeastern University (<https://gyorilab.github.io/>), the Alsentzer lab at Stanford University (<https://alsentzerlab.org/>) and the Undiagnosed Diseases Network Foundation (<https://udnf.org/>). Organizations interested in contributing data or testing methods on the platform can contact the Sage Bionetworks team at rapid@sagebionetworks.org.

About Sage Bionetworks

Sage Bionetworks is a Seattle-based 501(c)(3) nonprofit research organization founded in 2009 to drive a new age of discovery through open science and radical collaboration. Sage develops and operates Synapse, an NIH-listed generalist data repository, and has spent 15 years building governed data commons for sensitive biomedical data on behalf of the research community. Learn more at <https://sagebionetworks.org>.

Funding acknowledgment

This research was, in part, funded by the Advanced Research Projects Agency for Health (ARPA-H). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government.

Robert Allaway
Sage Bionetworks
+1 206-928-8250
rapid@sagebionetworks.org

This press release can be viewed online at: <https://www.einpresswire.com/article/937893708>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.