



NIH REPLICATION PRIZE

Track 2: Replication Exemplars

Winning Submissions — Public Reference Document

Sponsored by the National Institutes of Health (NIH)

Administered by Floor23 Digital

Awards Ceremony · May 13, 2026 · Virtual

This document compiles the winning submissions from Track 2 of the NIH Replication Prize — the Replication Exemplars track. Where a team had two members submit independently, both are included under the single team entry. This document is intended as a public resource to advance the science of replicability.

ABOUT THE NIH REPLICATION PRIZE

The NIH Replication Prize was created to recognize those who are making science more transparent, rigorous, and reliable, and contribute to a scientific environment that values replication and encourages scientists to share their research.

Launched by the National Institutes of Health and administered by Floor23 Digital through the Innobear challenge platform, the competition invited researchers, clinicians, and innovators to propose new ideas on what to replicate and ways to improve replicability, as well as spotlight effective strategies that have been used to replicate research and improve replicability in their own scientific practice.

Replication is a key tool in the scientific process, as it is critical to assessing internal and external validity, transferability, and generalizability of research outcomes and methods. Replication research sets the stage for advancing scientific discovery and technologies that can be applied to improving the health and well-being of the population.

Track 2 of the Replication Prize asked scientists to share strategies that they have used to integrate replication into their research activities. These Replication Exemplars show the varied ways different scientists implement replication practices in their own work. This compilation ensures these strategies are shared with the broader research community.

THE NIH REPLICATION PRIZE TEAM

The NIH Replication Prize is led by a dedicated team within the NIH Office of Strategic Coordination – The Common Fund, whose commitment to advancing scientific rigor and reproducibility made this competition possible. Their vision and leadership have been instrumental in bringing this initiative to life.

From competition design to winner selection, this team has championed the integrity and impact of the Replication Prize at every stage — and it is their dedication that makes today's celebration possible.

Team Members

Kaitlyn Browning, <i>Strategic Planning, Evaluation, and Communications Analyst</i>	Sahana N. Kukke, <i>Program Leader</i>
Alicia Cavanaugh, <i>Scientific Program Analyst</i>	Katelynn Milora, <i>Strategic Planning, Evaluation, and Communications Analyst</i>
Rachel Diamond, <i>Policy, Reporting and Legislative Coordinator</i>	LeShawndra Price, <i>Re-Engineering, Research Enterprise Assistant Director</i>
Michelle Hamlet, <i>Program Leader</i>	Jason Shockey, <i>Operations Analyst</i>

A Note from the NIH Replication Prize Team

"Sincere thanks to the Replication Initiative Team, Floor23, and you, the participants, for making this important program possible!"

HOW TO USE THIS DOCUMENT

Each entry corresponds to one Track 2 winning team or individual, following the seven-section competition structure: (1) Replication Strategy, (2) Environment, (3) Impact, (4) Response to Barriers, (5) Timeline and Budget, (6) Collaboration, and (7) Sustainability/Transferability. Where a team had two members submit independently, both submissions are included under the single team entry, each clearly labeled by submitter. Submissions are presented word-for-word as submitted or edited for confidentiality, and lightly formatted for readability. The views and opinions expressed here are those of the authors and do not necessarily reflect those of the NIH, the Department of Health and Human Services (HHS), or the United States government.

Each submission stands on its own and can be read independently.

Understanding the Replication Category Icons

To help you navigate the submissions, each entry is marked with one or more icons. Each icon represents a Replication Category. These categories were developed by analyzing how each winner defined replication in their own words (See Appendix A), and identifying common themes that emerged from the work itself. A submission may carry more than one icon when its approach to replication spans multiple dimensions. For example, a study that both standardized its protocol rigorously and validated the underlying mechanism would carry both the Methodological Rigor and Mechanistic Validation icons.



Methodological Rigor: The submission focuses on the standards of documentation, procedural precision, and standardization that make a study repeatable, positing that replication depends on how thoroughly the original work was described and structured.



Multi-Site / Cross-Lab Consistency: The submission defines replication as the ability to produce consistent results across independent laboratories, research sites, or time periods, testing whether findings hold up when reproduced by different teams under comparable conditions.



Biological / Reagent Reproducibility: The submission centers on the reproducibility of biological materials themselves — such as antibodies, cell lines, or biological assays, recognizing that variation in these materials is possibly/likely a primary driver of irreproducibility in life science research.



Mechanistic Validation: The submission independently verifies the underlying biological or chemical mechanism.



Clinical / Translational: The submission extends replication into patient populations or real-world clinical settings, validating that findings from laboratory or preclinical research translate meaningfully to human health outcomes.



Computational / Data Replication: The submission uses computational methods, modeling pipelines, or shared datasets to independently re-execute a prior study, emphasizing transparency of code, data curation protocols, and reproducibility of analytical outputs.

Table of Contents

At a Glance: Track 2 Replication Exemplars..... 4

 ATCC Federal Solutions..... 6

 Antibody Team..... 11

  C20D3-Vitamin A Replicate..... 21

 Christina Kackos Brunelle..... 38

  ConductScience..... 44

  Jagadeesh Ramasamy..... 54

  John Worthley..... 59

  Krissandra Johnson-Nealey..... 65

 LATCH (LLM-Assisted Testing of Clinical Hypotheses)..... 75

  Magdalena Kasendra..... 94

  Michele Di Mascio..... 106

  Predictive/UNC Team..... 118

  SOAR (Spatial Omics Analysis Replication)..... 132

  Team PGRM (Phenotype-Genotype Reference Map)..... 141

  The RESI Team..... 147

Appendix A: “How did you define replication?” 156

Corresponding Authors..... 159

	Methodological Rigor		Biological / Reagent Reproducibility		Clinical / Translational
	Multi-Site / Cross-Lab Consistency		Mechanistic Validation		Computational / Data Replication

At a Glance: Track 2 Replication Exemplars

WINNER	SUMMARY
ATCC Federal Solutions	An organization committed to the advancement and credibility of sound science that leveraged over a century of authenticated biological resources and standards to help federal agencies generate reliable, reproducible public health research data. 
Antibody Team	An international consortium that replicated antibody reliability research to drive adoption of sequence-defined recombinant antibodies in place of inconsistent animal-derived reagents. 
C20D3-Vitamin A Replicate	Pharmaceutical co-founders who replicated their mechanistic hypothesis on vitamin A dimerization across labs, animal models, and human trials to advance a treatment for retinal disease.  
Christina Kackos Brunelle	A vaccine development and infectious disease researcher who applied rigorous replication methods to strengthen confidence in published findings. 
ConductScience	A research-instrument manufacturer that built tools to the exact specifications of published methods, so labs can replicate experiments with verified, citable equipment.  
Jagadeesh Ramasamy	A biomedical research director and translational scientist who replicated and extended his own landmark discovery on mitochondrial persistence in red blood cells, advancing understanding of inflammation and disease.  
John Worthley	A researcher who tested whether ACE inhibitors developed at sea level perform consistently at high altitude, using high-elevation "natural laboratories" to study hypoxia's effect on drug function.  
Krissandra Johnson-Nealey	A 20-year clinical trial site operations veteran who applied her frontline experience to improve data integrity and reproducibility in metabolic and cardiovascular research.  
LATCH	A multidisciplinary team of physician-scientists and graduate researchers who developed an LLM-assisted framework that converts written study descriptions into executable analyses, making replication faster and surfacing methodological gaps. 

Magdalena Kasendra	A biomedical research director in organoid biology who led organoid and organ-on-chip research at Cincinnati Children's, advancing standardized, reproducible platforms for drug discovery and toxicology testing. 
Michele Di Mascio	A former NIH scientist who applied two decades of HIV and viral dynamics research experience to validate and reproduce findings in infectious disease imaging. 
Predictive/UNC Team	A UNC-based lab and startup that used AI to extract real-world human toxicity data, building a reproducible alternative to animal testing for regulatory science. 
SOAR (Spatial Omics Analysis Replication)	A University of Washington and Fred Hutch collaboration that built standardized lab protocols and open-source software to make spatial proteomics research reproducible across institutions. 
Team PGRM (Phenotype-Genotype Reference Map)	A team of researchers based in Vanderbilt University Medical Center and the University of Michigan who built and validated the Phenotype-Genotype Reference Map (PGRM), a tool that lets biobanks assess and improve the quality of their phenotyping data. 
The RESI Team	Four biostatisticians who applied the robust effect size index to improve how effect sizes are reported and replicated across biomedical imaging and psychiatry research. 

TEAM WINNER

ATCC Federal Solutions



Leverages over a century of authenticated biological resources and standards to help federal agencies generate reliable, reproducible public health research data.

Team Members: Rebecca Bradford · Heather Couch · Sujatha Rashid · Ashley Castens · Manasi Tamhankar · Helen Navin · Michael Parker · Danielle Williams · Deb Hendrickson · Ann Wasko

For more than 100 years, ATCC has advanced its mission to enable reproducible science by providing authenticated, high-quality biological resources and standards. Building on this legacy, ATCC partners with U.S. government agencies to advance gold-standard science in support of public health initiatives. Anchored in robust quality systems, reproducible workflows, and responsive program management, the ATCC team enables federal researchers to generate reliable, credible data that accelerates discovery, informs sound decision-making, and enhances preparedness for emerging and ongoing health threats. ATCC is committed to the advancement and credibility of sound science by supporting initiatives for the U.S. government, academia, industry, and research foundations.

1. REPLICATION STRATEGY

1.1 Strategy Description

The ability to reproduce viral challenge material with consistent genomic, phenotypic, and functional characteristics across multiple production runs, ensuring that independent lots meet predefined quality standards and can be used interchangeably in research without additional manipulation.

1.2 Replication Strategy

Our team implemented a comprehensive replication strategy aimed at producing Well-Characterized Challenge Material (WCCM) for viruses to ensure reproducibility and consistency in preclinical research. This approach was built on a foundation of systematic optimization, rigorous validation, and innovative enhancements to minimize variability and guarantee lot-to-lot interchangeability. To further strengthen reproducibility, we integrated innovative methods such as bioinformatics pipelines for continuous genomic stability monitoring and variant detection. Collectively, this strategy established a highly controlled, reproducible process for generating WCCM that supports reliable, interchangeable use in virus research.

2. ENVIRONMENT

2.1 Location, Timing and Team Details

Location: ATCC

Timeframe: 2020–2025

Team Composition:

- **Laboratory Staff:** Manasi Tamhankar, Michael Parker, Helen Navin, Danielle Williams, Deb Hendrickson, Ann Wasko
- **Scientific Subject Matter Expert (SME):** Sujatha Rashid
- **Project Management:** Heather Couch, Ashley Castens
- **Leadership:** Rebecca Bradford

2.2 Activity Overview

The activities were conducted at ATCC between 2020 and 2025, encompassing work on pathogens. Each team member contributed specialized expertise: laboratory staff executed experimental workflows and quality control procedures; the SME provided scientific guidance on virology and genomic stability; project managers oversaw timelines, resource allocation, and compliance; and leadership ensured strategic alignment and program oversight.

2.3 Funding Support

The funding for this work came from the National Institute of Allergy and Infectious Diseases.

3. IMPACT

3.1 Replication Activity Impact

Following the implementation of our systematic replication strategy, we achieved transformative improvements in quality, scalability, and efficiency. We established lot-to-lot consistency, reducing genomic variation confirmed by next-generation sequencing and the ATCC variant analysis pipeline. Each production run generated high-yield, stable stocks, producing large quantities of WCCM with titers consistently meeting or exceeding target thresholds.

This scale and reliability enabled researchers to access interchangeable lots without additional manipulation, dramatically reducing variability in experimental outcomes. Furthermore, we streamlined quality control processes, reducing characterization turnaround time through standardized NGS workflows and integrated bioinformatics pipelines. These efficiencies accelerated material availability for urgent studies while maintaining rigorous safety and quality standards. Collectively, this impact represents a paradigm shift from fragmented, variable production to a highly controlled, reproducible system that supports robust, comparable data generation across institutions. By eliminating the need for end-user passages and ensuring genomic and phenotypic stability, our approach strengthens the foundation for reproducible science virus research and enhances preparedness for emerging infectious disease threats.

3.2 Long-Term Impact

Our replication strategy delivers enduring benefits that extend well beyond individual research projects, fundamentally strengthening global scientific infrastructure and pandemic preparedness. By depositing WCCMs into BEI Resources, we have enabled broad distribution to academic, government, and industry laboratories worldwide. This accessibility supports critical activities such as antiviral screening, vaccine development, and therapeutic evaluation, ensuring that researchers across the globe can work with standardized, high-quality materials that promote reproducibility and accelerate discovery. In addition to global research enablement, our approach directly advances pandemic preparedness. The scalable protocols we established allow for rapid production of challenge material for emerging pathogens, reducing response time during public health emergencies. This capability ensures that when novel threats arise, validated and interchangeable viral stocks can be deployed quickly to support preclinical testing and countermeasure development. Collectively, these long-term impacts create a sustainable framework for reproducible science, global collaboration, and rapid innovation in the face of emerging infectious disease threats.

3.3 Research Rigor Improvements

Our replication strategy introduced critical enhancements that directly address long-standing challenges in reproducibility and experimental integrity. First, we eliminated uncontrolled passages by end-users, a common source of variability that historically led to genomic drift and inconsistent titers. By providing WCCM that is ready for immediate use “out of the vial,” we removed this uncontrolled step, ensuring that all researchers start with standardized material and reducing variability in downstream assays. Second, we integrated genomic variability screening into standard operating procedures, leveraging advanced bioinformatics tools. These pipelines enable continuous monitoring of genomic stability and rapid detection of low-frequency variants, ensuring that each lot

meets stringent integrity criteria. This proactive approach builds confidence in the fidelity of the challenge material and supports transparent, data-driven quality assurance.

Finally, we standardized documentation practices by issuing Certificates of Analysis and detailed product sheets for every production lot. These documents provide comprehensive information on titer, purity, genomic characterization, and stability metrics, offering researchers full visibility into material quality. This level of transparency not only strengthens reproducibility but also aligns with best practices for FAIR (Findable, Accessible, Interoperable, Reusable) data principles, enabling global collaboration and trust in shared resources. Collectively, these improvements represent a significant leap forward in research rigor, transforming viral challenge material production from a fragmented, variable process into a highly controlled, quality-assured system that underpins reproducible science.

3.4 Dissemination and Adoption

Our replication strategy delivers enduring benefits that extend well beyond individual research projects, fundamentally strengthening global scientific infrastructure and pandemic preparedness. By depositing WCCMs into BEI Resources, we have enabled broad distribution to academic, government, and industry laboratories worldwide. This accessibility supports critical activities such as antiviral screening, vaccine development, and therapeutic evaluation, ensuring that researchers across the globe can work with standardized, high-quality materials that promote reproducibility and accelerate discovery. In addition to global research enablement, our approach directly advances pandemic preparedness. The scalable protocols we established allow for rapid production of challenge material for emerging pathogens, reducing response time during public health emergencies. This capability ensures that when novel threats arise, validated and interchangeable viral stocks can be deployed quickly to support preclinical testing and countermeasure development. These improvements not only support current needs but also position the facility as a critical resource for future outbreak response. Collectively, these long-term impacts create a sustainable framework for reproducible science, global collaboration, and rapid innovation in the face of emerging infectious disease threats.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

During the development and implementation of our replication strategy, we encountered several significant barriers that required innovative solutions to ensure success.

Barrier 1: Variability in Virus Propagation Conditions. Early efforts revealed substantial variability in viral titers and genomic integrity due to differences in cell lines, infection parameters, and harvest timing. This variability threatened reproducibility and lot-to-lot consistency. **Solution:** We conducted systematic optimization studies to define standardized conditions for cell line selection, seeding density, multiplicity of infection (MOI), and days post-infection (DPI). These controlled experiments established a reproducible baseline that minimized sequence drift and maximized yield.

Barrier 2: Genomic Drift from End-User Passages. Historically, challenge materials required additional passages by end-users, introducing uncontrolled variability and genomic changes. **Solution:** We developed an “out-of-the-vial” usability model, producing WCCM that is ready for immediate use without further manipulation. This eliminated uncontrolled passages and significantly reduced experimental variability across institutions.

Barrier 3: Limited Genomic Surveillance and QC Integration. Initial workflows lacked robust genomic monitoring, making it difficult to detect low-frequency variants and ensure material integrity. **Solution:** We integrated NGS and bioinformatics pipelines into standard operating procedures, enabling real-time genomic variability screening and variant detection. This innovation strengthened trust in material quality and compliance with reproducibility standards.

Barrier 4: Fragmented Documentation and Transparency. Prior processes lacked standardized documentation, creating challenges for traceability and reproducibility. **Solution:** We implemented a rigorous documentation framework, including CoAs and detailed product sheets for every lot. These resources provide comprehensive data on titer, purity, genomic characterization, and stability, ensuring transparency and supporting FAIR principles.

4.2 Organizational Levels Addressed

At the point of distribution, users historically required additional passages to achieve experimental titers, introducing genomic drift outside ATCC's controlled environment.

Solution: ATCC implemented an “out of the vial” model, producing WCCM at ready to use titers and eliminating the need for downstream passaging.

5. TIMELINE & BUDGET

5.1 Timeline

- Eight months.

5.2 Associated Costs

- Labor: Virology, QC, bioinformatics teams.
- Materials: Cell culture reagents, sequencing kits, sterility testing.

5.3 Required Resources

- Personnel: Virologists, molecular biologists, bioinformaticians, project/program managers.
- Technology: NGS platforms, variant analysis pipeline, data management systems.

6. COLLABORATION

6.1 Key Collaborations

Successful implementation of replication activities required strategic partnerships to ensure scientific rigor and operational feasibility. In addition to external collaborations, internal partnerships between virologists and bioinformaticians played a critical role, enabling integrated workflows for virus production, genomic characterization, and advanced data analysis. These cross-functional efforts ensured that challenge materials were both phenotypically consistent and genomically stable. Collectively, these collaborations facilitated resource sharing, streamlined processes, and strengthened the infrastructure necessary for producing well-characterized challenge materials for emerging infectious disease research.

7. SUSTAINABILITY / TRANSFERABILITY

The replication activities are sustainable because they are embedded in SOPs for viral production and quality control at ATCC. These SOPs are adaptable across different virus families and scalable for emerging pathogens.

7.1 Current Sustainability Plans

- **Resource Sharing:** WCCM lots deposited into BEI Resources guarantee ongoing access for the research community.

7.2 Scaling and Application

- **Rapid Response Capability:** The strategy can be scaled for pandemic preparedness by applying the same optimization framework to novel viruses.

- **Broader Application:** Methods for genomic QC and variant screening can be applied to vaccine seed stock production and therapeutic development pipelines.

7.4 Future Integration Plans

- **Training Programs:** Develop workshops and virtual modules for replication best practices in virology.
- **Policy Alignment:** Advocate for inclusion of replication QC standards in institutional and federal guidelines.
- **Networking and Dissemination:** Expand collaborations through NIAID consortia and scientific conferences; publish optimization data in peer-reviewed journals.
- **Digital Tools:** Create a centralized database for WCCM characterization data to enhance transparency and reproducibility.

TEAM WINNER

Antibody Team



An international consortium addressing research reliability by driving adoption of sequence-defined recombinant antibodies in place of undefined animal-derived reagents.

Team Members: Valérie Allamand · Jeffery Bajramovic · Hannah Cable · Pierre Cosson · Doris Lou Demy · Stefan Dübel · Mark Fowler · Katherine Groff · Michael Hust · Ravindran Kumaran · Jens Kurreck · Raphael Munoz-Ruiz · Samera Rafiq · Ouarda Saib · Esther Wenzel · Kilian Zilkens

This multidisciplinary team brings together experts from academia, industry, biotechnology, and non-profit organizations to advance research reproducibility through the adoption of sequence-defined recombinant antibodies. Team members represent internationally recognized institutions and companies, including Abcalis, Abcam, ABCD Antibodies, the French Center for the 3Rs (FC3R), the National Institute of Health and Medical Research (INSERM), PETA, Technische Universität Berlin, Technische Universität Braunschweig, Unilever, the University of Geneva, and Utrecht University.

Collectively, the team has decades of experience in antibody discovery and engineering, assay development, characterization, regulatory science, education, and open science infrastructure. Their work spans development of fully human recombinant antibodies, large scale transitions away from animal-derived reagents, curation of repositories of antibody sequences and validation data, and integration of recombinant antibodies into university curricula.

United by a shared commitment to scientific rigor, transparency, and ethical innovation, the team collaborates across sectors and disciplines to deliver practical solutions that improve reproducibility in biomedical research while replacing animal use.

1. REPLICATION STRATEGY

1.1 Strategy Description

Advancing Replicability in Biomedical Research Through the Integration of Recombinant Antibodies (rAbs)

The replication strategy involves the systematic transition to sequence-defined rAbs across multiple domains to integrate reliability into standard research practice. This addresses direct replication by controlling artifacts via refining experimental procedures to decrease unintended influences that can distort the results. Transitioning to rAbs improves reagent quality and consistency, which directly affects the dependent variable measurement (e.g., signal strength, specificity, background noise).

1.2 Outcomes Evaluation Methods

Outcomes related to replicability in the context of transitioning to rAbs are evaluated based on the intrinsic properties of the new reagents, rigorous empirical characterization, and successful implementation and dissemination strategies. Initiatives outlined emphasize comprehensive characterization of antibodies, and the rAbs are evaluated for specificity and affinity. Multiple rAbs are tested using various techniques, including Western blot, slot blot analysis, and immunostaining on cultured cells. Testing includes comparison with negative and positive controls. Further, rAb sequence and characterization data are published.

Categories: Basic and pre-clinical

2. ENVIRONMENT

The initiatives outlined below reflect multi-stakeholder efforts to advance research replicability by systematically addressing the reliability challenges inherent in non-recombinant antibodies. These efforts span three key

dimensions: (1) the development of rAbs to ensure consistency and specificity; (2) the characterization and adoption of these reliable reagents across research and industry settings; and (3) educational curriculum integration to train future scientists in reproducible practices. By focusing on sequence-defined rAbs, these initiatives aim to strengthen the foundational tools of biomedical research, enabling both individual studies and broader scientific communities to achieve more replicable outcomes.

2.1 Initiatives to Increase Replicability via Development of rAbs

2.1.1 *Development of a monoclonal rAb against the diphtheria toxin*

Fully human sequence-defined recombinant monoclonal antibodies capable of neutralizing diphtheria toxin (DT) were developed using phage display technology. Equine-derived diphtheria antitoxin suffers from short supply and quality issues related to batch variability, potential contamination, and inherent risk of immunological reactions in recipients. To create a reproducible alternative, researchers at the Institute for Biochemistry, Biotechnology, and Bioinformatics at Technische Universität (TU) Braunschweig immunized volunteers with the diphtheria vaccine, collected blood, and constructed a human immune phage display library. Through two panning strategies—microtiter plate panning and solution-based panning combined with functional screening—the team generated 400 recombinant antibodies. Of these, 61 unique antibodies were characterized as antibody fragments and 35 were converted to full IgG1 format. Epitope mapping confirmed that the antibodies targeted all three domains of the diphtheria toxin: enzymatic, translocation, and receptor-binding. Combinations of 2-3 antibodies obtained strong neutralization capacity. It has been published in *Scientific Reports* (Wenzel et al., 2020), and its broad news coverage underscores public interest in the topic.

Research took place at TU Braunschweig in Germany from 2018-2019 in consultation with the National Institute for Biological Standards and Control (NIBSC), Division of Bacteriology, UK, PETA Science Consortium International (PSCI), and the WHO Collaborating Centre for Diphtheria and Streptococcal Infections of the UK Health Security Agency, UK. Leadership was provided by:

- Dr. Esther Wenzel was the lead researcher who designed and performed the experiments and analyzed data as a PhD student. She is Co-founder and COO of the biotech company Abcalis GmbH, which develops rAbs.
- Prof. Dr. Michael Hust served as project advisor, participated in the design of the study, and directed the experimental work. He is the Director of the Department of Medical Biotechnology. His research focuses on antibody engineering.
- Prof. Dr. Stefan Dübel, a Full Professor of Biotechnology and Director of the corresponding department, advised the project. Dr. Dübel initiated the "Antibody factory" of the German National Genome Research Network and is editor of the "Handbook of Therapeutic Antibodies" and other antibody engineering books. He is co-founder of several biotech companies and co-inventor of animal-free antibody generation by phage display.
- Dr. Paul Stickings leads the Toxins Group at NIBSC and served as project advisor.
- Jeffrey Brown served as project coordinator and is the PSCI Senior Scientist for Pharmaceuticals and Medical Devices.

Technische Universität Braunschweig provided ongoing institutional support, including laboratory space and equipment and researcher support. PSCI provided €134,000 for the development of the sequenced human antibodies neutralizing diphtheria toxin in vitro.

Based on the success of the project, a similar effort is being replicated at TU Braunschweig, with the Ensenada Center for Scientific Research and Higher Education in Mexico contributing its expertise in the isolation and characterization of toxins. This project is developing fully human recombinant monoclonal antibodies capable of neutralizing black widow spider venom. Antibody candidates selected via phage display from a naïve antibody gene library and capable of in vitro neutralization of alpha-latrotoxin of the European black widow have been isolated (Ruschig et al., 2024).

2.1.2 *IL-8 ELISA kit development*

Researchers commonly use paired capture and detection antibodies in sandwich enzyme-linked immunosorbent assays (ELISA). However, no ELISA kits currently available on the market utilize fully sequence-defined antibodies. To address this gap, our initiative focuses on developing fully sequence-defined capture and detection antibodies for interleukin-8 (IL-8), one of the most studied and clinically relevant cytokines in immunology (R&D Systems, 2025).

The 77-amino-acid isoform of IL-8 was selected as the target for antibody generation. Abcalis panned an in-house human-derived naïve antibody phage library using this recombinant 77-aa isoform of IL-8 protein from R&D Systems. Identified unique clones were expressed via transient transfection in HEK cells to produce rAbs, which were then screened for functionality. Sandwich ELISA pairs were assessed by evaluating specificity and affinity across different clones at varying antigen concentrations. The most effective capture antibody was selected and paired with various detection antibodies. These pairs were further tested under different conditions, including coating concentrations and blocking reagents.

Following initial screening, the antibodies of the best-performing pair were used to generate mutation libraries for optimization. The best-performing antibody pair was defined by:

- Lowest autoreactivity, assessed via ELISA against a panel of immune-related proteins.
- Highest sensitivity, demonstrated by the ability to detect low antigen concentrations with a clear dose-response curve.

Subsequently, Unilever tested the replicability of the performance of the antibodies identified at Abcalis via ELISA, and they optimized ELISA conditions for the top-performing pair within its laboratories (e.g., for antigen suppliers, blocking buffer type and duration, etc.).

During Unilever's evaluation, performance testing showed that the best Abcalis antibody pair achieved a detection limit of approximately 2.54 pg/mL (≈ 0.316 pM) for the 77-aa IL-8 isoform, with a calibration curve comparable to leading kits. Isoform specificity was confirmed in real biological samples: the Abcalis ELISA detected IL-8 in HepG2 supernatants (expected to contain 77-aa IL-8), and ongoing work is focused on optimising detection in PBMC supernatants (expected to contain 72-aa IL-8) once suitable pairs are available. The detection range for the Abcalis pair was comparable to the DuoSet kit (31–2,000 pg/mL).

The capture antibody also retained binding to the 72-aa isoform. To enhance detection of the 72-aa isoform, candidates from the matured antibody libraries will be screened to identify clones with cross-reactivity or specificity for this target. This strategy aims to ensure comprehensive coverage of both IL-8 isoforms in future assay iterations. Once completed, the cocktail of antibody pairs will reliably identify both the 72-aa and 77-aa IL-8 isoforms, and each antibody will be fully sequence-defined to enable consistent production in perpetuity.

Antibody discovery and optimization was conducted at Abcalis in Germany. Approximately \$84,000 in funding was provided by Unilever and PSCI for this work to be conducted from 2023 - July 2025, and \$12,700 will be provided for Abcalis to identify antibody isoforms targeting the 72-amino acid IL- protein. Internal funding and laboratory resources were provided by Unilever for assay optimization work within Unilever's laboratory conducted in August - September 2025.

Leadership was provided by:

- Dr. Kilian Zilkens conducted all laboratory work and is the lead scientist for antibody discovery at Abcalis GmbH. With a PhD in biotechnology, he has 10 years of experience in antibody development, including at Merck KGaA and Yumab GmbH.
- Ouarda Saib is a Senior Researcher at Unilever, based in Sharnbrook, Bedford, and Samera Rafiq is a Safety Scientist in Chemistry at SERS (Safety, Environmental & Regulatory Science) Unilever. They advised on the development and optimization of the antibody assays.
- Dr. Mark Fowler has more than 20 years of experience as a cell biologist at SERS Unilever. He optimized assay conditions of the developed antibodies within Unilever's facilities.

2.1.3 Abcam's initiatives to increase research reproducibility through rAb development

Starting in 2013, Abcam implemented a strategic large-scale transition from polyclonal and hybridoma-derived monoclonal antibodies to rAbs to address variability and improve research reliability. Alongside the transition to sequence-defined antibody formats, there were a number of parallel initiatives: >10,000+ ascites-derived products removed from sale; >50,000 polyclonals removed or replaced since 2015; and ~3,000 products switched to in vitro production or recombinant format. Today, all new antibody products developed in-house are in recombinant format, and Abcam has both the largest portfolio of recombinant primary antibodies (>32,000) as well as ELISA kits that use rAb pairs (>1,500).

Transitioning to more robust and renewable antibody formats is challenging. Development costs are higher to identify the right antibody-secreting B cell, cloning and sequencing methodologies are more costly than polyclonal approaches, and timelines can also be longer. However, consistent performance across batches and a reduction in lot-to-lot variability supports replication of study results and reduced off-target effects from monoclonal specificity supports accuracy of research outcomes. Sequence-defined antibodies are amenable to engineering –

including mutations to fine-tune binding and performance but also species-switching of Fc regions to enable variable secondary detection and use of multiple fluorochromes in one assay. Additionally, there is a significant impact on manufacturers' commitment to the 3Rs – in vitro production of rAbs reduces animal-derived antibody production and avoids genetic drift which can impact hybridoma-derived antibodies.

The organization CiteAb, a reagent search and data service for life science researchers, releases a report of the top cited commercial products each year (Williams, 2025). Abcam has been working to ensure high-quality recombinant formats of their top cited antibody products, with a range of options now available for key targets that work equivalently, or better, than the top cited polyclonal, ascites, or hybridoma alternatives (e.g., ki67, beta actin, alpha tubulin, GFAP).

To address manufacturing methodologies, Abcam introduced biophysical quality control methods of each lot of rAbs in 2023, further supporting batch-to-batch consistency as the 'fingerprint' of the antibody protein was validated by mass spectrometry, dynamic light scattering, and HPLC. These techniques give robust data on antibody aggregation, purity, sequence identity, and integrity.

This work was carried out across Abcam's international development and production facilities, including in Waltham, MA, Hangzhou (China), and Cambridge (UK). Leadership was provided:

- Dr. Alejandra Solache Diaz leads global R&D at Abcam Ltd. Key strategic initiatives under her leadership include the implementation of rAb development and a significant move away from non-recombinant formers, and the establishment of multiple validation programs, several of which have received industry awards.
- Dr. Ravindran Kumaran leads the development of Abcam's neuroscience products and collaboration programs to increase research reproducibility through improved antibody development and validation. He has extensive experience in neurodegenerative disease research developed through his studies at Kings College London, University College London and 10 years postdoctoral research at the NIH.
- Hannah Cable has >15 years of experience in antibody development and leads new product development and collaborations at Abcam. Her work on this project included development of advanced validation standards aligned to Uhlen and Bradbury's antibody validation pillars.
- Dr. Junli Yu leads antibody discovery in Abcam's China laboratories. Her role on this project included development of biophysical quality control assays for the rAb portfolio and implementation of advanced antibody discovery methods.

2.2 Initiatives to Increase Replicability via Characterization and Use of Developed rAbs

2.2.1 Characterization of antibodies used for life science purposes

This initiative aimed to formally characterize rAbs and their ability to reliably and specifically identify targets across different platforms and models as compared to traditional antibodies. Spearheaded by the GIS FC3R (French Centre for 3Rs), it was conducted in phases involving national collaboration, rigorous testing, and comprehensive dissemination of results. In partnership with the French Association of Histotechnology and antibody suppliers, the project was initiated in 2024 by conducting a national survey to identify which animal-derived antibodies were most commonly used on histology platforms in France. Twelve national platforms (laboratories) participated in the survey, which successfully identified 36 high priority antibodies to be replaced, based in part on frequency of use and a lack of confidence in the existing, non-sequence defined antibodies. From 2024 – present, FC3R has been coordinating a pilot study, which focuses on the characterization and validation of 12 target proteins selected for testing. For each selected target protein, partner suppliers provided three rAbs. Histology platforms at sites across France performed immunofluorescence characterization and validation of these antibodies using relevant positive and negative controls as described below.

- CD3: The APEX Platform (at INRAE [the National Research Institute for Agriculture, Food, and the Environment, a French public scientific and technical research establishment]) tested four recombinant anti-CD3 antibodies on tissues. Two, Ab16669 (Abcam) and CST 85061 (Cell Signaling Technology), were validated in mice, rats, and macaques. The other two, AbAb YTH 12.5 (Absolute Antibody) and "ABCD_AD036" (ABCD Antibodies), did not produce a signal under the tested conditions. The validated rAbs exhibited immunostaining properties identical to or superior to those of the reference polyclonal antibody Dako A0452.
- CD31 – or PECAM-1 (Platelet endothelial cell adhesion molecule): The RHEM Platform (CNRS, 2025) tested four recombinant anti-CD31 antibodies on tissues. Two (Ab182981 and Ab281583) were validated

in pigs, mice, and rats. Another (AbAb JC70) was validated in humans only. “ABCD_AD217” did not produce specific staining under the tested conditions.

- Chromogranin A (CHGA): The INRAE Internal Histology Workshop at St Gilles, a laboratory facility, tested two recombinant anti-CHGA antibodies on tissues. One (Ab283265) was validated in pig, mouse, and rat species, and the other (CST 36468) did not produce a signal under the tested conditions.
- Glial fibrillary acidic protein (GFAP): The APEX platform tested three recombinant anti-GFAP antibodies on tissues. All three, Ab278054, “ABCD_AK148”, and CST 80788, were validated in mice, rats, and macaques. All exhibited immunostaining properties identical to or superior to those of the reference polyclonal antibody Dako Z0334.
- Ki67 (Kiel, clone 67): The RHEM platform tested three recombinant anti-Ki67 antibodies on mouse tissues (intestine and spleen). All three (Ab16667, Ab279653, and MA5-14520) produced a strong, biologically relevant signal with little or no background noise. The staining patterns obtained with the three rAbs were similar to each other, but not comparable to that obtained with the platform's reference monoclonal antibody (which was less intense and more widespread), and which has since been discontinued. These results suggest that more specific staining was achieved with the rAbs than with the antibody they were intended to replace.
- KRT14 (Keratin 14): The INRAE Internal Histology Workshop at St Gilles tested four recombinant anti-KRT14 antibodies on cells and tissues. One (CST 74956S) was validated in cows, goats, and sows; one (Abcam Ab119695) detected KRT14 in cows (2D cell culture and tissue sections), and in sows, but not in goats; one (AbAb RCK107) detected bovine KRT14 only in 2D cell culture, and the labeling was partial; and one (AbAb 2GA) did not produce a signal under the tested conditions.
- Occludin: The INRAE Internal Histology Workshop at St Gilles tested four recombinant anti-Occludin antibodies on cells and tissues. One (Abcam Ab216327) was validated in pigs (cells and intestinal sections), but results were mixed in cows. Two (AbAb 1-3 and AbAb 37-5) showed encouraging results on pig and bovine tissue sections, but the testing conditions need optimization. One (CST 91131) did not produce a signal under the tested conditions.
- CD68: The H2P2 Platform (Université de Rennes, 2025) is testing five rAbs against human CD68, including AbAb FA-11, Abcam Ab303565 (multiclonal), “ABCD_AD140”, “ABCD_AD141”, and CST E307V.

The final phase to make detailed protocols and data widely available began in September 2025. Results and data are being shared through publications on the FC3R website (Short Notes) and in the journal *Antibody Reports*.

Resources to support this initiative were provided by numerous stakeholders. The histology platforms stated above provided laboratory space and equipment, personnel, and cell and tissue samples to perform the histology work at cost. The antibody companies listed above provided samples of antibodies for testing. FC3R provided €40,000 to the platforms for the characterization of the rAbs and spearheaded the project. Leadership was provided by:

- Dr. Doris Lou Demy served as project advisor, directed the design of the study, and coordinated logistical aspects of the national survey, antibody procurement, laboratory characterization, and data dissemination.
- As director of FC3R and a researcher at the Cochin Institute at INSERM, Dr. Athanassia Sotiropoulos served as project advisor.

In a congruent approach, from 2024 onwards, the 3Rs Centre Utrecht (3RCU) with ABCD antibodies started an initiative to cross-validate eight rAbs directed against common cell biological targets (GFP, Giantin, KDEL, Cisd1) and tags (FLAG [DYKDDDDK], Myc, HA, 6xHis). Across the Utrecht Science Park campus multiple research groups are testing, validating, and incorporating these antibodies in their workflows. Resources to support this initiative were provided by ABCD and 3RCU as in-kind contributions. Total costs equal approximately €100,000, and leadership was provided by:

- Dr. Jeffrey Bajramovic, director of the 3RCU, served as project advisor, directed the design of the study, and coordinated logistical aspects.
- Dr. Raphael Munoz-Ruiz (see also below) is a post-doctoral fellow at the University of Geneva and the Director of Operations for ABCD Antibodies SA.

2.2.2 University characterization, database, and dedicated journal for rAbs

The Geneva Antibody Facility at the University of Geneva aims to advance the use of rAbs to improve experimental replicability by employing a multi-platform approach:

- The ABCD Database – a curated, open-access repository of sequenced antibodies with known targets.

- *Antibody Reports* – an open-access technical journal publishing detailed validation data for rAbs.
- Antibody discovery service – discovery of new rAbs for researchers at economical prices.
- Antibody production service– producing existing rAbs for researchers at economical prices.

Through its discovery service, the facility has produced thousands of rAbs for the scientific community since 2014, aiming to make high-quality reagents widely accessible for fundamental research. Many of these antibodies are cataloged in the manually curated ABCD (AntiBodies Chemically Defined) Database (<https://web.expasy.org/abcd/>), which contains primary sequence information essential for reagent identity and reproducibility. As of December 2025, the database includes > 32,000 sequenced antibodies. To further disseminate rAbs, the University of Geneva launched ABCD Antibodies SA, a spin-off company offering low-cost production of these open-access antibodies.

Launched in 2018, *Antibody Reports* is an international, academic open access online journal edited by the University of Geneva and indexed in DOAJ. Designed to focus on technical, experimental validation of rAbs, it is a unique resource currently with nearly 200 publications with functional data on rAbs in specific applications, enabling researchers to optimize antibody use and replicate experimental results.

Initial setup required: basic molecular biology and chemistry equipment, a small laboratory, roughly \$100,000 for materials, and one dedicated staff member. Leveraging existing infrastructure, the total launch cost was approximately \$500,000. Ongoing support comes from institutional funds and revenue from custom antibody development and catalog sales through the facility and ABCD Antibodies SA. Leadership was provided by:

- As a Full professor in the Department of Cell Physiology and Metabolism, Faculty of Medicine at the University of Geneva, Prof. Pierre Cosson founded the facility, and he continues to oversee their continuous growth. He is also a co-founder of ABCD Antibodies SA.
- Dr. Raphael Munoz-Ruiz is a post-doctoral fellow at the University of Geneva and the Director of Operations for ABCD Antibodies SA. In this position, he has overseen the founding and growth of ABCD Antibodies.
- Philippe Hammel serves as the lead technician at the Geneva Antibody Facility, overseeing rAb development.

2.2.3 Validation of novel rAbs specific to extracellular matrix components

Prompted by batch-to-batch variability between commercially available animal-based antibodies commonly used in Dr. Allamand's group, this project aimed to validate rAbs specifically targeted against components of the extracellular matrix (ECM) produced by ABCD Antibodies SA. The experimental validation of the selected rAbs started in April 2025 and is on-going. Several methodologies were employed to validate the novel rAbs on biological material derived from human cell cultures (primary dermal fibroblasts and immortalized myogenic cells): immunocytochemistry on PFA-fixed cell and ECM layer, slot-blotting on conditioned medium (CM) and/or protein extracts, Western blotting on CM and cell layer. Immunostaining on tissue cryosections remains to be performed.

The tested rAbs were directed against the following targets:

- Fibronectin-1: Among the four rAbs tested, two ("ABCD_AR861", "ABCD_AA629") gave strong positive signals by immunostaining and slot-blot, and they remain to be validated by Western blot. These two antibodies performed significantly better than a commercial mouse monoclonal Ab regularly used in the lab (HFN from DSHB). A third rAb ("ABCD_AU776") only gave a positive signal by slot-blot on CM. Another rAb was tested although its initial concentration was very low and it did not yield any positive result.
- Tenascin C: Four rAbs were tested and gave positive signals by immunostaining on fixed dermal fibroblasts and by slot-blot, but two ("ABCD_AS071", "ABCD_AS103") performed better compared to the other two ("ABCD_AE902" and "ABCD_AF606").
- Laminin alpha-2 chain (LAMA2): One rAb ("ABCD_AW449") was tested by immunofluorescence on fixed dermal fibroblasts and myogenic cells. The signal was shown to be specific as it disappears on cells from patients with mutations in the LAMA2 gene. This is also confirmed by using commercially available antibodies. However, there is a strong background present when "ABCD_AW449" is tested via slot-blot.
- Matrix metalloproteinase 9 (MMP9): Two rAbs ("ABCD_AF171", "ABCD_AU857") are promising as they yielded positive signals by slot-blot on conditioned medium from dermal fibroblasts and protein extracts from different cell lines (fibroblasts, HEK293, COS7, lymphoblasts, myogenic cells). Two other rAbs ("ABCD_AW361", "ABCD_AF497") gave a weaker signal on the same protein extracts. Due to the lack of negative control for these antibodies, siRNA-mediated knock-down of MMP9 was initiated but remains to be finalized.

Overall, several rAbs performed very well and could be used quite diluted (1µg/ml). Slot-blot proved to be an efficient method to screen the various antibodies on conditioned medium from cultured cells and total protein extracts from cell layers. Antibody use is dependent on application, making data such as this extremely useful to the wider scientific community. Antibody validation will continue in 2026 in order to disseminate the results through publication in the journal *Antibody Reports*, and dissemination during seminars and conferences.

Leadership was provided by:

- As a research director at the Center for Research in Myology (Sorbonne University-Inserm), Dr. Valérie Allamand, PhD, served as project supervisor, directed the design of the project, and performed some experiments.
- Dr. Raphael Munoz-Ruiz, as the Director of Operations for ABCD Antibodies SA, oversaw the choice of rAbs to be tested with Dr. Allamand, as well as the funding and growth of ABCD Antibodies, and coordinated logistics to supply the rAbs.
- Dr. Louise Benarroch, as a postdoctoral fellow in the Center for Research in Myology (Sorbonne University-Inserm), performed some experiments.
- Lauren P Strubble, as a Master Student from University Paris Cité (France), performed most experiments between April and August 2025, under the co-supervision of Dr. Allamand and Dr. Benarroch.

2.3 Initiatives to Increase Replicability by Incorporating rAbs into University Courses

2.3.1 Integration of rAbs in practical courses

This initiative developed and implemented a graduate-level laboratory course within Technische Universität Berlin (TUB) that incorporated the use of phage display-derived sequenced antibodies to enhance scientific reproducibility and ethical practices. The course utilizes both animal-free (phage display-derived) and rAbs primarily within the Western blot analysis module for detecting the target protein GP73 and the loading control actin. An HRP-conjugated phage display-derived anti-actin antibody was employed as a fully animal-free alternative to the conventional approach. The conventional method previously used in the teaching setting relied on a monoclonal mouse anti-actin primary antibody followed by an HRP-conjugated goat anti-mouse polyclonal secondary antibody, both of which were undefined. The HRP-conjugated phage display-derived antibody is fully reproducible and streamlines the protocol by removing the secondary antibody incubation step. To validate use of the new antibody, they were tested in Western blot, with both the conventional and the phage display-derived detection systems producing a single, specific band, demonstrating comparable performance. Thus, moving forward the phage display antibody can be used as a fully replicable option to detect actin, which serves as a widely used internal standard and loading control to confirm equal protein loading during Western blotting.

For detecting the target protein Golgi Protein 73 (GP73), a mixed detection strategy was implemented. Because no phage display-derived antibodies were available for the target protein GP73 and resources precluded the development of a custom antibody, the course used a monoclonal primary antibody against GP73 (mouse anti-GP73) in combination with a recombinant HRP-conjugated secondary antibody (m-IgG2a BP-HRP: sc-542731). HeLa cells were transfected with increasing concentrations of siRNA targeting GP73 (siGP73) (10, 50, or 80 nM) or a non-targeting control siRNA (siCon, 80 nM). GP73 proteins were detected via Western blot using the anti-GP73 primary antibody and recombinant secondary antibody. Results indicated a concentration-dependent reduction in GP73 protein levels when transfected with siGP73.

Dr. Jens Kurreck spearheaded the development and implementation of this course, which was carried out in March 2025. He is a Full Professor for Applied Biochemistry at the Technische Universität Berlin (TUB) with a primary focus in bioprinting. Financial support was provided by PSCI (\$5,600 in 2024), the Berlin Animal Protection Commissioner (P776 and NAM Teaching Award 2022, €5,000), and the Einstein Foundation Berlin (Einstein Center 3R, EZ-2020-597-2, €83,000).

This course can be disseminated to other universities using materials that Dr. Kurreck has provided in Nessar et al. (2025).

Team Contributions

- **Prof. Dr. Jens Kurreck** conceived and led the project, designed the overall educational concept, supervised its implementation, and coordinated the integration of recombinant antibodies into the laboratory course at Technische Universität Berlin.

- **Alexandra Nessar** (M.Sc. student) performed the majority of the experimental work and contributed substantially to the validation of the teaching protocol.
- **Viola Röhrs** (technical assistant) supervised and supported the laboratory work, performed experiments, and helped establish the practical workflow.
- **Mathias Ziersch** (PhD student) coordinated the integration of the project into the university laboratory course and established the link to the practical teaching curriculum.
- **Ahmed S. M. Ali** (PhD student) contributed to the adaptation of the cell culture system to fetal bovine serum (FBS)-free conditions, thereby supporting the implementation of animal-component-free methodologies used alongside the project.
- **Julia Moradi** (M.Sc. student) contributed to the project during its early stages under the supervision of **Dr. Johanna Berg**.
- **Dr. Anke Kurreck** contributed a key figure to the associated publication and provided scientific support throughout the project.

3. IMPACT

3.1 Replication Activity Impact

For research to be replicable, the reagents, including antibodies, must be reliable. Antibodies profoundly affect research replicability, as they are essential assay-specific tools that form the backbone of numerous applications across scientific disciplines. Antibodies are crucial for detecting and quantifying molecules, biological compounds, viruses, and cells in research and diagnostic testing applications in medicine.

The reproducibility crisis in life sciences has highlighted the limitations of traditional antibody formats. As widely used research tools, antibodies have historically contributed to irreproducible results due to batch variability and undefined composition, as well as poor validation. Reviews of large-scale validation efforts have found that only about 49% of commercial, non-recombinant antibodies tested could be validated to recognize only their targets (Bradbury and Plückthun, 2015). Early large-scale projects such as the Human Protein Atlas highlighted the importance of rigorous antibody validation and the variability in antibody performance (Berglund et al., 2008). Using a standardized characterization approach, a study by Ayoubi et al (2023) quantified antibody performance across 614 commercial antibodies for 65 neuroscience-related protein targets. Their findings demonstrated that more than 50% of overall antibodies failed in one or more of the three tested applications, with rAbs performing significantly better than monoclonal or polyclonal antibodies. This is a critical finding relevant to research reproducibility. Extrapolation of results indicate 20-30% of protein studies use ineffective antibodies, with the potential to not only slow down research progress but also hinder progress towards therapeutic outcomes (Virk et al., 2019).

It has been estimated that unreliable antibodies may account for half of the worldwide spending on protein-binding reagents, resulting in the waste of tremendous amounts of money, time, and experimental samples (Bradbury and Plückthun, 2015).

Modern initiatives such as YCharOS (Antibody Characterization through Open Science) have confirmed the scale of this problem and provided actionable solutions. YCharOS has characterized more than 800 commercial antibodies using knockout cell line controls and standardized protocols, revealing that more than half fail specificity tests. Crucially, their findings show that sequence-defined rAbs consistently outperform traditional monoclonals and polyclonals in validation studies, offering superior specificity and reproducibility (Monteiro et al., 2025; Biddle and Virk, 2023). These results reinforce the need for systematic transition to rAbs and transparent validation frameworks to ensure reliability across biomedical research.

Recombinant antibodies are selected from libraries of genes encoding slightly different antibody proteins for their affinity to bind to target antigens. The sequences of rAbs can be reproduced by DNA synthesis, resulting in nearly no lot-to-lot variability. Further, only rAbs inherently offer the advantage of being structurally defined by their polypeptide sequence, which provides a unique identifier directly linked to specificity.

Impacts specific to each initiative are outlined within the respective section in section 2.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

- **Financial:** Securing funding to rigorously evaluate reagents and develop more replicable, sequence-defined antibodies is a substantial challenge. To address this barrier, these initiatives have been undertaken in a collaborative nature.
- **Capacity:** Millions of antibodies are available for purchase from hundreds of suppliers internationally. Thus, transitioning to rAbs is a substantial endeavor that requires prioritization. Stakeholders involved in these initiatives had to prioritize which antibodies to develop and characterize and which to use within studies and practical courses.
- **Availability:** While the availability of off-the-shelf rAbs is steadily increasing, it remains limited compared to traditional antibodies. This gap presents a significant barrier to their broader implementation in both research studies and academic coursework. Overcoming this challenge will require sustained multi-partner collaborations, increased funding to support development and distribution, and greater awareness of the critical role that defined reagents play in ensuring research replicability.
- **Technical:** Today, there are various robust and reliable methods available for the generation and production of rAbs (Dübel, 2024; Wenzel et al., 2026). As with all antibodies, rAbs must be fit-for-purpose. Not all recombinant antibodies developed bind sufficiently to their target molecules or exhibit sufficient specificity toward them from the outset. Due to their recombinant nature, however, they can be easily optimized for their intended purposes, which may require several rounds of in vitro antibody maturation.

5. TIMELINE & BUDGET

Timeline and budget specific to initiatives are outlined in section 2, Environment.

6. COLLABORATION

6.1 Key Collaborations

As described in section 2, success in this area is founded on a coordinated, multi-stakeholder effort involving diverse organizations, including academic and research institutions, NGOs, biotechnology companies, and industry partners.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Current and Future Sustainability Plans

Stakeholders such as those in section 2 are actively engaged across sectors in laying the foundation to increase replicability of studies via broader adoption of rAbs and have begun disseminating initiatives as described above. Building on this momentum, this team plans to expand these efforts through continued interdisciplinary collaboration, targeted initiatives to characterize and apply rAbs, and the replication of successful educational models. We will continue collaborative efforts to expand the availability of rAbs within catalogs for commercial and academic use, publish characterization data to enhance transparency and data availability, and replicate scalable curriculum models.

8. REFERENCES

1. Ayoubi, R., Ryan, J., Biddle, M.S., Alshafie, W., Fotouhi, M., Gonzalez Bolivar, S., et al. (2023). Scaling of an antibody validation procedure enables quantification of antibody performance in major research applications. *eLife*, 12, RP91645. <https://doi.org/10.7554/eLife.91645.2>
2. Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L., Asplund, A., Al-Khalili Sziguarto, C., et al. (2008). A gene-centric human protein atlas for expression profiles based on antibodies. *Molecular and Cellular Proteomics*, 7(10), 2019–2027. <https://doi.org/10.1074/mcp.R800013-MCP200>
3. Biddle, M.S., Virk, H.S. (2023). YCharOS open antibody characterisation data: lessons learned and progress made. *F1000Research*, 12, 1344. <https://doi.org/10.12688/f1000research.141719.1>
4. Biddle, M., Stylianou, P., Rekas, M., Wright, A., Sousa, J., Ruddy, D., et al. (2024). Improving the integrity and reproducibility of research that uses antibodies: a technical, data sharing, behavioural and policy challenge. *mAbs*, 16(1), 2323706. <https://doi.org/10.1080/19420862.2024.2323706>
5. Bradbury, A., Plückthun, A. (2015). Reproducibility: standardize antibodies used in research. *Nature*, 518(7537), 27–29. <https://doi.org/10.1038/518027a>
6. Bradbury, A.R.M., Trinklein, N.D., Thie, H., Wilkinson, I.C., Tandon, A.K., Anderson, S., et al. (2018). When monoclonal antibodies are not monospecific: hybridomas frequently express additional functional variable regions. *mAbs*, 10(4), 539–546. <https://doi.org/10.1080/19420862.2018.1445456>
7. Centre National de la Recherche Scientifique. (n.d.). Réseau Histologie Expérimentale et Modèles (RHeM). Retrieved December 11, 2025, from <https://www.rhem.cnrs.fr/>
8. Dübel, S. (2024) Can antibodies be "vegan"? A guide through the maze of today's antibody generation methods. *Mabs*, 16(1). <https://doi.org/10.1080/19420862.2024.2343499>
9. Monteiro, F.L., Voskuil, J.L.A., Williams, C. (2025). YCharOS protocol for antibody validation. *Nature Protocols*, 20(6), 1389–1390. <https://doi.org/10.1038/s41596-024-01108-6>
10. Nessar, A., Röhrs, V., Ziersch, M., Ali, A.S.M., Moradi, J., Kurreck, A., et al. (2025). Promoting ethical and reproducible cell culture: implementing animal-free alternatives to teaching in molecular and cell biology. *Frontiers in Toxicology*, 7, 1670513. <https://doi.org/10.3389/ftox.2025.1670513>
11. McGill University, The Neuro (Montreal Neurological Institute–Hospital). (2025, December 17). Open solution for improving research reproducibility. Retrieved December 18, 2025, from <https://www.mcgill.ca/neuro/channels/news/open-solution-improving-research-reproducibility-362210>
12. R&D Systems. (n.d.). Human IL-8/CXCL8 Quantikine ELISA Kit (D8000C). Retrieved December 16, 2025, from https://www.rndsystems.com/products/human-il-8-cxcl8-quantikine-elisa-kit_d8000c (452 citations)
13. R&D Systems. (n.d.). Human IL-8/CXCL8 DuoSet ELISA Kit (DY208-05). Retrieved December 16, 2025, from https://www.rndsystems.com/products/human-il-8-cxcl8-duoset-elisa_dy208 (502 citations)
14. Ruschig, M., Nerlich, J., Becker, M., Meier, D., Polten, S., Cervantes-Luevano, K., et al. (2024). Human antibodies neutralizing the alpha-latrotoxin of the European black widow. *Frontiers in Immunology*, 15, 1407398. <https://doi.org/10.3389/fimmu.2024.1407398>
15. Uhlen, M., Bandrowski, A., Carr, S., Edwards, A., Ellenberg, J., Lundberg, E., et al. (2016). A proposal for validation of antibodies. *Nature Methods*, 13, 823–827. <https://doi.org/10.1038/nmeth.3995>
16. Université de Rennes. (n.d.). Histopathologie. Retrieved December 16, 2025, from <https://histopathologie.univ-rennes.fr/>
17. Virk, H.S., Rekas, M.Z., Biddle, M.S., Wright, A.K.A., Sousa, J., Weston, C.A., et al. (2019). Validation of antibodies for the specific detection of human TRPA1. *Scientific Reports*, 9, 18500. <https://doi.org/10.1038/s41598-019-55133-7>
18. Wenzel, E.V., Bosnak, M., Tierney, R., Schubert, M., Brown, J., Dübel, S., et al. (2020). Human antibodies neutralizing diphtheria toxin in vitro and in vivo. *Scientific Reports*, 10, 571. <https://doi.org/10.1038/s41598-019-57103-5>
19. Wenzel, E.V., Al-Halabi-Frenzel, L., Demiral, S., Frenzel, A., Gnann, D.C., Keitel, M.C.E., et al. (2026). The landscape of antibody production systems: Recombinant expression for research, diagnostic and therapy. *Frontiers in Bioengineering and Biotechnology*, 14, 1765764. <https://doi.org/10.3389/fbioe.2026.1765764>
20. Williams, R. (2025, October 13). Top antibodies in research: 100 most cited products, 2025 update. CiteAb Blog. <https://blog.citeab.com/top-100-antibodies-in-research-2025-update/>

TEAM WINNER

C20D3-Vitamin A Replicate



Co-founders of Alkeus Pharmaceuticals who replicated their mechanistic hypothesis on vitamin A dimerization across labs, animal models, and human trials to advance a treatment for retinal disease.

Team Members: Ilyas Washington · Leonide Saad

Dr. Ilyas Washington and Dr. Leonide Saad are co-founders of Alkeus Pharmaceuticals and have led a translational research program targeting retinal degenerative diseases.

Dr. Washington (PhD, UCLA; BA, Bard College) trained under Ken Houk in computational chemistry and completed an NIH-sponsored postdoctoral fellowship at Columbia University with Koji Nakanishi and Nicholas Turro. His work elucidated the role of vitamin A dimerization in lipofuscin accumulation and retinal degeneration and led to the invention of deuterated vitamin A (C20-D3 vitamin A; ALK-001/gildeuretinol) to modulate this process without disrupting the visual cycle.

Dr. Saad (PhD, MIT) has expertise in tissue engineering, drug development, and clinical translation. As CEO of Alkeus, he has led the advancement of ALK-001 from preclinical validation through multi-center clinical trials and regulatory milestones, including FDA Breakthrough Therapy designation for Stargardt disease, and a positive Phase 2 clinical trial in Stargardt disease.

Together, the team C20D3-Vitamin A Replicate implemented a multi-system replication strategy spanning independent laboratories, animal models, and human trials, enabling translation of a mechanistic hypothesis into a clinical-stage therapeutic program supported by the NIH, FDA, retinal foundations, and private investment.

SUBMISSION BY: ILYAS WASHINGTON

1. REPLICATION STRATEGY

1.1 Description of the Replication Activities

The replication activities comprised a deliberate, stepwise confirmation of a single mechanistic hypothesis: slowing the rate-limiting step of vitamin A dimerization through selective deuteration at carbon 20 (C20D3-vitamin A) is sufficient to prevent the biochemical, functional, and structural pathology underlying Stargardt disease and related retinal degenerations. Replication began with in vitro chemical and biochemical systems, where the kinetic isotope effect was isolated from biological confounders. Kaufman et al. (2010) demonstrated that deuteration at carbon 20 slowed A2E formation approximately 7-fold and ATR-dimer formation approximately 12-fold, establishing a reproducible, quantitative relationship between selective isotope substitution and dimerization rate. These results were replicated across varying deuteration ratios, showing a graded response in which $\geq 80\%$ replacement of native vitamin A with deuterated vitamin A resulted in $\sim 80\%$ inhibition of vitamin A dimer formation. Replication then progressed to in vivo disease models, where the same mechanism was tested against biological endpoints relevant to human pathology. In albino Abca4 knockout mice (Ma et al. 2010), dietary replacement of vitamin A with C20D3-vitamin A reduced retinal A2E and ATR-dimer concentrations to levels comparable with wild-type mice, representing reductions of approximately 60–80% relative to controls. Fundus autofluorescence—a surrogate for lipofuscin accumulation—was reduced by approximately 70% after three months of treatment, and dark-adapted ERG responses were preserved at 12 months, demonstrating functional replication using an electrophysiological endpoint. To replicate these findings under conditions more closely resembling the human retina, pigmented Abca4 knockout mice were then studied by an independent laboratory at the University of Oxford (Nuffield Laboratory of Ophthalmology) in the laboratories of Dr. Robert MacLaren and at University of Bonn (Issa et al. 2015).

After nine months of treatment, fundus autofluorescence measured at both 488 nm and 790 nm excitation wavelengths was reduced to near wild-type levels, representing an approximate 50–60% reduction compared with untreated Abca4 knockout controls. Electron microscopy confirmed a $\sim 50\%$ reduction in RPE lipofuscin granule burden, and no adverse effects on scotopic or photopic ERG responses were observed, demonstrating replication

of efficacy without functional toxicity. The most stringent replication occurred in *Abca4/Rdh8* double knockout mice, which develop overt RPE atrophy analogous to human disease (Zhang et al. 2021A, 2021B). In this model, long-term administration of C20D3-vitamin A for up to 18 months resulted in an average 75% reduction in fundus autofluorescence signals and an ~85% reduction in quantifiable vitamin A cycle byproducts (including A2E and oxo-A2E) throughout the study. Critically, 100% of control animals developed RPE atrophic pathology by 13–18 months of age, whereas 0% of treated animals exhibited any RPE atrophy at any time point examined. Functional replication was observed as preserved dark-adaptation kinetics, with treated animals recovering ~71% of baseline ERG b-wave amplitude at 30 minutes post-bleach compared with ~53% in controls. Long-term toxicology and tissue-distribution studies confirmed retinal incorporation of deuterated vitamin A without systemic disruption of vitamin A–dependent physiology. The final stage of replication occurred in translational and clinical settings (Upcoming publication).

Successful replication across chemical, biological, and preclinical domains enabled GMP manufacturing, regulatory engagement, and progression into multi-phase clinical trials. In 2021, the FDA granted Breakthrough Therapy Designation for the treatment of Stargardt disease. Clinical trials involving 200+ patients over two years have been completed, and a Phase 3 trial is ongoing. At this stage, replication was defined not by repeated experiments but by the consistent preservation of mechanism-driven outcomes under real-world clinical constraints.

1.2 Working Definition of Replication Used by the Team

The team adopted a definition of replication that emphasized durability of inference rather than procedural repetition. Replication was understood as the demonstration that a scientific claim remains valid when tested across different models, measurement strategies, time horizons, and decision-relevant endpoints. Under this definition, a result was not considered replicated simply because it could be repeated under identical conditions. Instead, replication required that the same causal mechanism produced consistent outcomes when evaluated using orthogonal assays, distinct biological systems, and increasingly translational criteria. This approach intentionally exceeded minimum reproducibility standards in order to reduce the probability of false confidence—a central concern motivating the Replication Prize. Importantly, the team viewed replication as inseparable from use. A finding that could not withstand replication under conditions relevant to regulatory, clinical, or manufacturing decisions was treated as insufficient, regardless of statistical significance in early studies.

1.3 Scientific Category

This work spans multiple categories but is best characterized as translational and pre-clinical research with direct clinical extension. The replication strategy intentionally bridged basic mechanistic chemistry, disease-model biology, pharmacology, toxicology, and human clinical investigation.

1.4 Replication Gap Addressed

The replication gap addressed by this work sits at the intersection of three persistent failures in biomedical research: (1) the inability of promising mechanistic findings to translate into durable disease modification, (2) the fragility of preclinical results when tested longitudinally or across biological contexts, and (3) the absence of replication standards that meaningfully inform regulatory and clinical decision-making. In retinal degenerative diseases such as Stargardt disease and age-related macular degeneration, numerous therapeutic strategies have demonstrated early biochemical or functional effects—often reductions in surrogate markers such as lipofuscin or short-term functional improvements—yet have failed to prevent progressive retinal atrophy when tested over time. These failures reflect a replication gap not of experimental repetition, but of durability: early findings do not reliably persist when evaluated against structural endpoints, advanced disease models, or clinically relevant time horizons. A second replication gap concerns mechanistic specificity. Vitamin A metabolism is essential for vision and systemic health, and many prior approaches attempted to suppress disease by broadly inhibiting the visual cycle or reducing vitamin A availability. While such approaches could reduce toxic byproducts, they frequently introduced unacceptable side effects, limiting translational viability. The field therefore lacked replicated evidence that pathological vitamin A dimerization could be selectively inhibited without disrupting essential vitamin A biology. This distinction could not be resolved by single-model or short-duration studies.

A third gap involves the historical treatment of kinetic isotope effects in biology. Although well established in chemistry, kinetic isotope effects had rarely been replicated across whole-organism disease models, particularly in ways that demonstrated structural disease prevention rather than biochemical modulation alone. Prior to this work, there was no replicated evidence that a subtle kinetic modification could propagate through complex biological systems to alter disease trajectory. This replication strategy explicitly addressed these gaps by requiring that the same mechanistic intervention:

- Produce consistent quantitative effects across multiple genetically distinct disease models
- Maintain efficacy over long durations (up to 18 months in vivo)
- Prevent overt retinal pigment epithelium atrophy, not merely delay biochemical accumulation
- Preserve retinal function without introducing systemic toxicity
- Generate data of sufficient rigor to support regulatory submissions and large-scale clinical testing

By closing these gaps, the replication effort transformed an isolated mechanistic insight into a durable, decision-grade scientific claim. This directly advances the contest's broader goal of elevating replication from narrow confirmation toward a process that determines whether scientific findings are truly fit for use.

1.5 Rationale for the Strategy and Why Replication Was Needed

Replication was essential because the intervention targeted a foundational biological process—vitamin A metabolism—that is both ubiquitous and indispensable. Any therapeutic strategy modifying this pathway required extraordinary confidence that beneficial effects in the retina would not compromise systemic physiology. Replication therefore served two purposes. Scientifically, it tested whether selective inhibition of vitamin A dimerization could decouple pathological byproduct formation from essential vitamin A function. Institutionally, it generated the level of evidence required for regulatory acceptance, long-term investment, and patient exposure. This strategy aligns directly with the contest's broader objective of elevating replication from a confirmatory afterthought to a central driver of scientific credibility and societal trust.

1.6 Innovative Elements of the Replication Strategy

The replication strategy incorporated several innovations:

- Conceptual replication across multiple disease models rather than reliance on a single system
- Longitudinal replication over disease-relevant timescales (up to 18 months)
- Independent replication by external laboratory (University of Oxford/University of Bonn)
- Integration of replication with regulatory and manufacturing milestones
- Treatment of replication as a translational pipeline rather than isolated experiments.

2. ENVIRONMENT

2.1 Organizational and Research Context

All replication activities were led by a single laboratory under the direction of the applicant, providing continuity of scientific leadership while enforcing rigorous internal replication standards. The laboratory operated across academic, translational, and industry-linked settings, enabling replication to proceed without artificial boundaries between discovery and application. This environment allowed replication decisions to be driven by scientific uncertainty rather than publication incentives, consistent with the goals of the Replication Prize.

2.2 Location, Dates, and Duration of Activities

Replication activities occurred across three institutional settings:

- 2007–2018: Columbia University Medical Center, Department of Ophthalmology, New York, NY
- 2010–Present: Alkeus Pharmaceuticals, Boston, MA (co-founded by applicant)
- 2017–Present: biOOrg3.14, New York, NY (founded by applicant) Independent replication was conducted at the University of Oxford (Nuffield Laboratory of Ophthalmology) and University of Bonn (Department of Ophthalmology) between 2013–2015.

Individual studies ranged from short-term kinetic experiments to multi-year animal and clinical investigations. Total duration of systematic replication: 18 years.

2.3 Team Composition, Expertise, and Roles

The team included undergraduates, graduate students, postdoctoral fellows, physicians, and PhD scientists spanning chemistry, ophthalmology, pharmacology, toxicology, and regulatory science. The applicant served as principal investigator on the preclinical and clinical work and as senior author on the primary publications. Independent replication was conducted by Drs. Peter Charbel Issa, Alun R. Barnard, Philipp Herrmann, and Robert E. MacLaren at Oxford/Bonn, with the applicant contributing reagents, analytical methods, and co-authorship.

2.4 Funding Sources and Resource Support

Funding was provided by the National Institutes of Health, the National Eye Institute, the U.S. Food and Drug Administration, and private eye disease foundations, including the International Retinal Research Foundation. These sources enabled sustained replication across stages rather than fragmented, short-term validation.

3. IMPACT

3.1 Pre- vs Post-Replication Metrics

Before systematic replication, the work existed as a mechanistic hypothesis supported by discrete experimental observations: slowed vitamin A dimerization in vitro and early reductions in toxic byproducts in limited animal models. At this stage, uncertainty remained high regarding durability, generalizability, safety, and relevance to human disease. The findings were insufficient to justify long-term studies, regulatory engagement, or clinical exposure. Post-replication, the evidence base changed qualitatively and quantitatively. Across multiple independent disease models and endpoints, replication produced convergent metrics demonstrating robust disease modification:

- Biochemical replication: Vitamin A cycle byproducts (A2E, ATR-dimers, oxo-A2E) were reduced by approximately 60–85% across models, with suppression maintained for up to 18 months in vivo.
- Functional replication: Dark-adapted ERG responses were preserved in treated animals, including a ~34% relative improvement in dark-adaptation recovery at 12 months in *Abca4/Rdh8* double knockout mice.
- Structural replication: In the most stringent disease model, 0% of treated animals developed RPE atrophy compared with 100% of controls, establishing complete prevention of a defining pathological endpoint.
- Translational replication: The same intervention supported GMP manufacturing, regulatory submissions, FDA Breakthrough Therapy Designation (2021), Phase 2 clinical trials (200 patients, 2 years), Phase 3 clinical trials (300+ patients ongoing), and creation of over 60 technical jobs. These post-replication metrics shifted the work from hypothesis-driven exploration to outcome-validated intervention.

3.2 Effects on Research Rigor and Confidence

Replication fundamentally altered the confidence with which the findings could be interpreted and used. Early results suggested plausibility; replicated results established reliability. Each additional layer of replication—chemical, biological, pathological, pharmacological—narrowed uncertainty and reduced reliance on extrapolation. Replication demonstrated that the therapeutic effect was mechanism-constrained rather than artifact-driven. The graded relationship between deuteration level and dimer suppression, the consistency of outcomes across genetically distinct models, and the absence of off-target pharmacology collectively increased confidence that observed benefits were causally linked to the intended mechanism. This increase in rigor had practical consequences. It enabled:

- Long-term dosing without evidence of cumulative toxicity
 - Acceptance of the data package by regulatory bodies
 - Investment in large-scale manufacturing and clinical infrastructure
- Replication thus functioned as a risk-reduction strategy, converting scientific uncertainty into quantified confidence.

3.3 Effects on Acceptance, Interpretation, or Use of Findings

The replicated findings help to reshape how Stargardt disease and related retinal degenerations are interpreted. Rather than being viewed solely as inexorable degenerative processes, these conditions could be reframed as disorders driven by a modifiable biochemical pathway whose intervention yields durable structural and functional benefit. This shift influenced multiple downstream uses of the findings. Clinically, it justified trials designed to test disease modification rather than symptomatic delay. Regulators were able to evaluate the intervention against long-term safety and efficacy expectations rather than short-term biomarker changes. More broadly, the work demonstrated that replication-intensive strategies can accelerate, rather than impede, translation by preventing late-stage failure. By showing that replication strengthens interpretation and guides appropriate use, this work exemplifies the contest's goal of embedding replication at the center of scientific progress.

3.4 Dissemination and Adoption Beyond the Original Team

Results were disseminated through six peer-reviewed publications in high-impact journals (PNAS, Journal of Biological Chemistry, Translational Vision Science & Technology), U.S. Patent No. 9,486,432, other pending US patents, and invited lectures at major scientific conferences including the Association for Research in Vision and Ophthalmology (ARVO; 12,000+ members from 75+ countries) and business incubators such as MIT's Innovation Lab. The work has been covered by Forbes, The Wall Street Journal, Nature, The Boston Globe, The Boston Herald, and nationally televised on FOX News. The research has received multiple awards recognizing its translational impact.

4. RESPONSE TO BARRIERS

4.1 Scientific or Technical Barriers Encountered

The primary scientific barrier was the intrinsic difficulty of replicating a subtle mechanistic effect—a kinetic isotope effect—across complex biological systems and over clinically relevant time horizons. Unlike interventions that fully inhibit or replace biological pathways, the strategy here relied on a modest, rate-limiting kinetic modification that had to propagate through retinal biochemistry, cellular physiology, tissue structure, and ultimately disease outcome. This introduced several technical challenges. First, the effect size at the molecular level, while robust in vitro, risked attenuation in vivo due to biological redundancy and compensatory mechanisms. Demonstrating that slowed dimerization could meaningfully reduce toxic byproduct accumulation without disrupting essential vitamin A-dependent processes required long-duration studies and sensitive, orthogonal endpoints. Second, many standard retinal disease models fail to develop overt structural pathology, limiting their usefulness for testing true disease prevention. Overcoming this required replication across multiple genetically distinct models, including the *Abca4/Rdh8* double knockout mouse, which develops reproducible RPE atrophy. Conducting longitudinal studies in these models—often exceeding one year—posed logistical, financial, and experimental challenges uncommon in preclinical research. Third, selective replication demanded rigorous control over vitamin A sourcing, isotopic enrichment, dietary formulation, and tissue incorporation. Ensuring consistent replacement of native vitamin A with C20D3-vitamin A across cohorts and over time was essential to attributing observed effects to mechanism rather than variability.

4.2 Institutional or Cultural Barriers

Institutionally, the work faced the common barrier that replication-intensive research is often undervalued relative to novelty-driven discovery. Longitudinal replication across multiple models does not yield rapid publications, yet it consumes substantial resources and time. Securing sustained funding for replication rather than exploratory expansion required alignment with funders willing to prioritize rigor and translation. Culturally, the prevailing incentive structure in biomedical science often encourages early claims of efficacy based on limited data. The decision to delay clinical advancement until replication had been achieved across stringent endpoints ran counter to pressures to move quickly, particularly in competitive translational fields. This approach required a deliberate commitment to replication as a prerequisite for credibility. Additionally, integrating replication into regulatory and manufacturing pathways required navigating organizational boundaries between academia, industry, and

government. Replication data had to meet standards acceptable across these domains, further increasing the burden of proof.

4.3 Solutions Implemented and Lessons Learned

These barriers were addressed by embedding replication as a core design principle rather than a post hoc validation step. From the outset, experiments were structured to test durability, generalizability, and mechanism specificity. Models were selected not for convenience but for their ability to falsify the central claim. Long-term studies were prioritized despite their cost because they provided decisive information unavailable from short-term assays. Quantitative thresholds—such as prevention of RPE atrophy rather than reduction of biomarkers—were used as go/no-go criteria for advancement. A key lesson was that replication accelerates translation when aligned with decision-making. By resolving uncertainty early and definitively, the strategy prevented costly late-stage failures and enabled confident progression into manufacturing and clinical trials. More broadly, this work demonstrates that replication is most effective when treated as infrastructure: an integrated process that supports discovery, translation, and societal impact simultaneously.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

The replication strategy unfolded over approximately eighteen years, with each phase contingent on successful replication at the prior stage. This gated progression ensured that advancement reflected reduced uncertainty rather than momentum alone.

- 2007–2010: Foundational chemical and biochemical replication establishing the kinetic isotope effect on vitamin A dimerization, including quantitative characterization of A2E and ATR-dimer suppression (Kaufman et al.).
- 2010–2012: Initial in vivo replication in albino Abca4 knockout mice, establishing biochemical and functional relevance and confirming preservation of retinal physiology (Ma et al.).
- 2013–2015: Independent replication in pigmented Abca4 knockout mice at Oxford/Bonn to test robustness in a model more reflective of human retinal biology (Issa et al.).
- 2015–2021: Longitudinal replication in Abca4/Rdh8 double knockout mice, including structural pathology endpoints and extended follow-up to 18 months (Zhang et al.).
- 2016–Present: Translational replication encompassing pharmacology, toxicology, GMP manufacturing, regulatory submissions, and multi-phase clinical trials.

Each transition required predefined replication criteria to be met before proceeding, preventing premature escalation.

5.2 Costs and Resources Required

Costs scaled with biological and translational complexity. Early chemical and animal studies required modest laboratory resources, while long-term animal studies, toxicology, and GMP manufacturing represented significant investments. Funding from federal agencies and private foundations enabled continuity, avoiding fragmented replication efforts. Importantly, resources were concentrated on replication rather than parallel exploratory programs, reflecting a strategic allocation toward uncertainty reduction.

5.3 Efficiency and Cost-Effectiveness of the Strategy

Although replication-intensive, the strategy proved cost-effective by eliminating weak or non-durable hypotheses early. By resolving key uncertainties before clinical deployment, the program avoided late-stage failures that typically consume disproportionate resources. Replication thus functioned as an efficiency mechanism rather than a cost burden.

6. COLLABORATION

6.1 Existing Partnerships Leveraged

The replication effort leveraged long-standing collaborations with academic institutions, federal agencies, contract research organizations, and industry partners. These partnerships provided complementary expertise in ophthalmology, chemistry, pharmacology, toxicology, regulatory science, and manufacturing. Support from the National Institutes of Health, the National Eye Institute, and the U.S. Food and Drug Administration enabled replication to proceed at a scale and rigor uncommon in academic laboratories.

6.2 New Partnerships Established

The Oxford/Bonn collaboration arose specifically to provide independent validation and was formalized through co-authorship and ongoing scientific exchange. As replication advanced into translational and clinical stages, new partnerships were established with GMP manufacturers, clinical research organizations, and clinical trial sites. These collaborations were formed only after replication milestones were achieved, ensuring alignment around validated evidence.

6.3 Roles of External Collaborators

External collaborators contributed specialized capabilities—such as independent confirmation of key findings using fundus autofluorescence imaging, toxicology studies, manufacturing, and clinical execution—while scientific leadership and replication design remained centralized. This structure preserved conceptual coherence while enabling scale.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Sustainability of Replication Activities

The replication activities described here were not episodic or grant-bound; they were embedded as a permanent operating principle of the laboratory and its translational pipeline. Replication became the mechanism by which scientific credibility was maintained, partnerships were formed, and decisions were justified. Because the work progressed only after replication criteria were met, replication naturally sustained itself as new questions arose and new stakeholders became involved. As the program advanced, replication was repeatedly revisited—not because prior results were doubted, but because each new context (new model, new endpoint, new regulatory threshold) imposed additional credibility requirements. This recursive use of replication ensured that confidence did not erode as complexity increased.

7.2 Scalability to Other Teams or Contexts

The replication strategy proved scalable because it relied on principles rather than proprietary infrastructure. Reproducing the work across three distinct animal models—albino *Abca4* knockout mice, pigmented *Abca4* knockout mice, and *Abca4/Rdh8* double knockout mice—demonstrated that the core findings were not model-specific artifacts. Independent confirmation through collaboration further reinforced this conclusion. As additional collaborators and potential partners became involved, transparency became central to scalability. In multiple instances, partners requested replication studies or analyses that the core team believed were scientifically redundant. These studies were nevertheless conducted because they were necessary to establish confidence external to the originating laboratory. This openness to replication beyond internal necessity made the work feel accessible, inspectable, and trustworthy to outsiders.

7.3 Plans for Continued or Expanded Replication

Replication continues as the compound advances through late-stage clinical trials. Ongoing and planned efforts include replication of clinical outcomes across sites, confirmation of long-term safety profiles, and post-trial analyses designed to test durability of benefit. Replication thus evolves with the lifecycle of the intervention, shifting from preclinical validation to clinical and real-world confirmation. As of 2025, interim data from the TEASE-3 study demonstrate halted disease progression in early-stage Stargardt patients treated with ALK-001 (C20D3-vitamin A), reinforcing mechanistic durability in humans. Additional FDA designations, including Rare Pediatric Disease and Fast Track status, further support expanded replication through accelerated trial pathways and post-approval monitoring. Importantly, replication was also essential for capital formation. Independent

reproduction of results across models and contexts provided the confidence required to raise capital, scale manufacturing, and initiate clinical trials. In this sense, replication functioned not only as a scientific safeguard but as a prerequisite for translation.

7.4 Generalizability of the Strategy Across Fields

The replication framework demonstrated here is broadly generalizable. Its core elements—multiple independent models, long-term endpoints, transparency with collaborators, and willingness to replicate beyond internal sufficiency—apply to any field where complex systems and high stakes demand durable evidence. Researchers in other therapeutic areas—particularly those involving chronic diseases, subtle mechanistic interventions, or long development timelines—can adopt this same gated replication framework to reduce translational failure and build stakeholder confidence before committing to costly late-stage development. By treating replication as a means of building shared confidence rather than merely confirming correctness, this strategy aligns scientific rigor with practical adoption. The result is work that others can trust, invest in, and build upon. This approach exemplifies how replication can serve as connective tissue between discovery, collaboration, and societal impact.

8. REFERENCES

1. Kaufman Y, Ma L, Washington I. Deuterium enrichment of vitamin A at the C20 position slows the formation of detrimental vitamin A dimers in wild-type rodents. *J Biol Chem*. 2011;286(10):7958–65. Epub 20101112. doi: 10.1074/jbc.M110.178640. PubMed PMID: 21075840; PubMed Central PMCID: PMC3048682.
2. Ma L, Kaufman Y, Zhang J, Washington I. C20-D3-vitamin A slows lipofuscin accumulation and electrophysiological retinal degeneration in a mouse model of Stargardt disease. *J Biol Chem*. 2011;286(10):7966–74. Epub 20101214. doi: 10.1074/jbc.M110.178657. PubMed PMID: 21156790; PubMed Central PMCID: PMC3048683.
3. Charbel Issa P, Barnard AR, Herrmann P, Washington I, MacLaren RE. Rescue of the Stargardt phenotype in Abca4 knockout mice through inhibition of vitamin A dimerization. *Proc Natl Acad Sci U S A*. 2015;112(27):8415–20. Epub 20150623. doi: 10.1073/pnas.1506960112. PubMed PMID: 26106163; PubMed Central PMCID: PMC4500285.
4. Zhang D, Robinson K, Washington I. C20D3-Vitamin A Prevents Retinal Pigment Epithelium Atrophic Changes in a Mouse Model. *Transl Vis Sci Technol*. 2021;10(14):8. doi: 10.1167/tvst.10.14.8. PubMed PMID: 34878528; PubMed Central PMCID: PMC8662574.
5. Zhang D, Robinson K, Saad L, Washington I. Vitamin A cycle byproducts impede dark adaptation. *J Biol Chem*. 2021;297(3):101074. Epub 20210812. doi: 10.1016/j.jbc.2021.101074. PubMed PMID: 34391781; PubMed Central PMCID: PMC8427233.
6. Washington I. Compositions and methods for treating macular degeneration. U.S. Patent No. 9,486,432. November 8, 2016.

Generative AI tools were used to assist with organization, editing, and clarity of presentation. All scientific content, data interpretation, and conclusions are those of the applicant.

SUBMISSION BY: LEONIDE SAAD

1. REPLICATION STRATEGY

1.1 Replication Overview

In 2007, Dr. Washington was assistant professor at Columbia University. He invented a new molecule, gildeuretinol acetate, a deuterated form of vitamin A (C20-D3-vitamin A or code name ALK-001), with the potential to treat forms of blindness such as Stargardt disease and age-related macular degeneration. After the invention, he needed to ensure that his initial results were replicable. Doing this as a team proved to be the most efficient

route. Stargardt disease is a leading form of blindness in children and young adults, caused by accelerated formation of vitamin A dimers in the retina, associated primarily with ABCA4 genetic defects (recessive).

The replication strategy involved a two-stage process: animal and other preclinical data replication, followed by clinical validation. Information presented here is the result of a 15+ year valiant effort to bring a therapy from the bench to patients.

Preclinical replication successfully confirmed the mechanism of action of gildeuretinol acetate both in test tubes and in a variety of relevant mice and other models, resulting in several publications.(1-10) His foundational work demonstrating that C20-D3-vitamin A slows lipofuscin accumulation in an Abca4 knock-out mice models was further independently replicated (11) with a partnership with Oxford University and the then start-up, Alkeus Pharmaceuticals.

Clinical replication then proceeded under the sponsorship of Alkeus Pharmaceuticals, resulting in positive findings from the phase 2 clinical trial “TEASE-1”, which evaluated the therapy in patients with advanced Stargardt disease, as well as in the TEASE-3 study, which focused on pre-symptomatic Stargardt patients.

Methods for Evaluating Outcomes: - Pre-clinical: Efficacy was evaluated using Quantitative Autofluorescence (qAF) imaging to measure lipofuscin (a biomarker associated with vitamin A dimers) accumulation in the retinal pigment epithelium (RPE), demonstrating a significant reduction in the treated group. Electroretinography (ERG) was also used to show that treatment with C20-D3-vitamin A could preserve electrophysiology of the retina. - Clinical: Outcomes were evaluated by measuring the rate of disease progression, specifically the rate of atrophic lesion growth in the retina using Short Wavelength Fundus Autofluorescence (FAF) imaging.

Innovative methods:

- Use of deuterium substitution (C20 D3 vitamin A) to exploit the kinetic isotope effect and slow vitamin A dimerization, addressing the primary chemical cause of disease. At the time, there were no therapies approved that contained a deuterium atom.
- TEASE 3 design that enrolls early stage, mostly pre-symptomatic patients and employs congenic, sibling controls, providing a sensitive test of therapeutic benefit at the earliest stages of the disease.

1.2 Working Definition

Replication was defined as independent, multimodal verification of the central hypothesis “slowing vitamin A dimerization prevents retinal pigment epithelium damage in mice and halts progression of retinal atrophy in humans.” This encompassed: (1) independent validation of molecular mechanism and efficacy in pigmented and albino animal models using quantitative biomarkers; and (2) confirmation of clinical benefit across patient populations with Stargardt disease and age related macular degeneration

Category Identification: Pre-clinical, Clinical, and Translational

2. ENVIRONMENT

2.1 Activity Context

The replication work took place within a high priority drug development program aimed at converting a novel small molecule chemistry concept into a therapy for Stargardt disease - a rare, inherited retinal disorder caused by ABCA4 mutations that leads to toxic vitamin A dimer (lipofuscin) accumulation in the retinal pigment epithelium. No FDA approved treatment exists for this condition. The central hypothesis was that replacing native vitamin A with the deuterated analogue C20-D3-vitamin A could slow dimerization, reduce lipofuscin formation, and thereby preserve RPE and retinal health. Alkeus Pharmaceuticals, Inc. sponsored the translation from academic proof of concept to multicenter clinical trials intended for regulatory filing.

2.2 Location, Timing, and Team Details

Pre-clinical/Translational Replication:

- **Location:** Primarily at the Columbia University Medical Center (New York, NY) for initial development and proof-of-concept, and the Nuffield Laboratory of Ophthalmology, University of Oxford (Oxford, UK) for the critical independent pre-clinical replication study.
- **Timing:** The foundational work was published in 2011, with the independent replication study in pigmented Abca4 mice published in 2015.
- **Lead Team Details:** Ilyas Washington, PhD: (Academic Inventor, Columbia University) Expertise/Role: Synthetic chemistry and mechanism of action; primary responsibility for the deuterated vitamin A concept and testing. Peter Charbel Issa, MD, and Robert E. MacLaren, MD, PhD: (Principal Investigators, Oxford University) Expertise/Role: Independent confirmation of pre-clinical phenotype rescue; responsible for the blinded, objective replication study using specialized quantitative retinal imaging (qAF) and ERG. Leonide Saad, PhD: Alkeus Pharmaceuticals CEO and founder, in charge with organization of the research, and preparation of regulatory dossier and submissions, as well as manufacturing of drug in preparation of clinical trials

Clinical Replication (TEASE-1 and TEASE-3 Studies):

- **Location:** Over seven clinical sites primarily across the United States.
- **Timing and Duration:** TEASE-1 (Advanced disease/Symptomatic Patients): Randomized, controlled trial with an initial 2-year treatment period. Data read-out occurred in 2021. TEASE-3 (Early-Stage Patients): Open-label study initiated around 2015, with patients continuing to show stable vision for up to seven years.
- **Lead Team Details:** Leonide Saad: Executive oversight, strategic development, securing funding, and study director and principal investigator of the NIH and FDA grants Ilyas Washington: Scientific coordination, chemistry, and co-principal investigator on the same grants Christine Kay, M.D., Michael Gorin, MD, PhD, Stephen Tsang, MD, PhD, Hendrik Scholl, M.D., and others: Clinical investigators in charge of clinical oversight, monitoring, patient enrollment, and data collection.

2.3 Funding Support

The replication and development effort was supported by a combination of federal grants, organizational/foundation support, and private investment.

- NIH/National Eye Institute: - R44 EY021988 (2016 2017) – \$2.1 M for Phase 2a ALK 001 trial in geographic atrophy. - R01 EY021207 – \$2.0 M to investigate vitamin A dimer pathology in retinal degeneration.
- FDA Office of Orphan Products Development: - R01 FD004098 (2013, 2017) – \$0.4 M for Phase 1/2 Stargardt disease studies. - R01 FD006316 (2018 2023) – \$1.75 M supporting TEASE Phase 2 trials.
- Organizational/Foundation: - BrightFocus Foundation grant (2010) – early academic funding for deuterated vitamin A research. Private/Corporate: - Alkeus Pharmaceuticals, Inc. – \$150 M venture capital invested to execute late stage clinical trials and manufacturing.

These combined resources enabled independent pre clinical validation, translational scaling, and multi site clinical replication of the ALK 001 therapeutic approach.

3. IMPACT

3.1 Replication Activity Impact

The replication activities had a profound and quantifiable impact to enable the development of C20-D3-vitamin A:

Pre-clinical Validation and Mechanism Confirmation:

- Pre-replication state (Initial 2011 study): Showed that C20-D3-vitamin A treatment in an albino Abca4 knockout, albino mouse model (prone to lipofuscin accumulation) resulted in an 80% decrease in the toxic dimer A2E and significantly slowed lipofuscin accumulation and degeneration. This established the proof-of-concept.

- Post-replication state (Independent 2015 study): The replication in a more clinically relevant, pigmented Abca4 mouse model by an independent lab demonstrated that C20-D3-vitamin A treatment reduced lipofuscin-related fundus autofluorescence (qAF) by 47% and electrophysiologically preserved retinal function to near wild-type levels. This independently verified that the drug's mechanism of slowing vitamin A dimerization was responsible for correcting the Stargardt phenotype, moving the therapy from a promising academic concept to a robust pre-clinical candidate.

Clinical Efficacy and Patient Cohort Expansion:

- Pre-replication state (TEASE-1, Advanced Patients): Showed a positive but limited effect, with a ~22% reduction in the growth rate of retinal atrophic lesions against untreated patients in the advanced stage of Stargardt disease. This confirmed efficacy in symptomatic patients.
- Post-replication state (TEASE-3, Early-Stage Patients): This crucial replication study in a more challenging and informative cohort demonstrated a dramatically superior outcome, with participants showing no disease progression over multiple years (up to seven years of treatment). This changed the therapeutic expectation from merely slowing progression in advanced disease to potentially halting the disease when administered early, highlighting the optimal therapeutic window.

3.2 Long-Term Impact

The long-term impact is the acceleration of a potential disease-modifying therapy for Stargardt disease, a blinding condition with no current approved treatment.

- Regulatory milestones: The replicated data secured FDA Breakthrough Therapy, Orphan Drug, Rare Pediatric Disease, and Fast Track designations for ALK-001.
- Paradigm shift: Approval would represent the first disease modifying therapy for Stargardt disease and would validate a treatment strategy based on genetic screening and early biochemical intervention.
- Extension to other indications: The mechanistic proof has enabled Phase 2a trials of ALK 001 in dry age related macular degeneration, broadening the potential impact beyond Stargardt disease.

3.3 Research Rigor Improvements

The replication activities significantly improved research rigor by: - Model relevance: Transition from albino to pigmented mouse models eliminated species specific artifacts and better approximated human retinal physiology.

- Objective endpoints: Use of quantitative imaging biomarkers (qAF pre clinically; short wavelength FAF clinically) and electrophysiology provided unbiased, reproducible measures of efficacy.
- High fidelity controls: TEASE 3 employed untreated affected siblings as congenic controls, maximizing statistical power and confidence in the observed effect size. These practices collectively raised the standard of rigor for pre clinical and clinical retinal research.

3.4 Dissemination and Adoption

Describe how these activities have been disseminated to and adopted by others, if applicable - Scientific communication: The independent replication was published in PNAS (2015) and cited extensively; clinical results have been presented at AAO, ARVO, and other major meetings and appear in peer reviewed journals, influencing current discussions of standard care for Stargardt disease.

- Clinical uptake: Retina specialists now incorporate the data when counseling patients with early stage disease about emerging oral disease modifying options, and trial designers for other inherited retinal dystrophies are modeling early intervention protocols on TEASE 3's sibling control design.
- Industry influence: The success of a deuterated vitamin approach has spurred other companies to explore kinetic isotope effect strategies for stabilizing bioactive molecules, extending the concept to multiple therapeutic areas.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

Barrier 1: Scientific Validity and Model Translation

- **Description:** The initial, successful proof-of-concept study was conducted in an albino mouse model. This raised a critical question: would the mechanism (slowing vitamin A dimerization) be equally effective in human eyes, which contain the pigment melanin in the retinal pigment epithelium (RPE)? Melanin could potentially interact with the toxic lipofuscin, creating a translational barrier.
- **Resolution:** The research team overcame this by facilitating an independent replication in a pigmented Abca4 mouse model by an external laboratory (University of Oxford). This rigorous, independent study confirmed that C20-D3-vitamin A treatment reduced the disease-driving lipofuscin accumulation by 47% and preserved retinal function, even in the presence of melanin. This independent, more stringent verification eliminated the translational concern and validated the core mechanism.

Barrier 2: Demonstrating Optimal Clinical Efficacy

- **Description:** The first clinical trial (TEASE-1) focused on patients with advanced, symptomatic Stargardt disease. While the drug slowed the progression of retinal damage (22% reduction in lesion growth), the effect size was modest because significant, irreversible damage had already occurred. This suggested that a "slow-down" therapy might not provide sufficient clinical benefit to the most affected patients.
- **Resolution:** The clinical replication strategy was pivotally shifted to the TEASE-3 study, which focused on the pre-symptomatic/early-stage patient population (children and adolescents). This was a high-risk group to recruit but provided the critical data. The results, demonstrating no disease progression over multiple years, definitively proved that the therapy is a disease-halting agent when applied early, effectively overcoming the barrier of limited clinical benefit observed in the later-stage cohort.

Barrier 3: Financial Resource Constraints for a Rare Disease

- **Description:** Developing a drug for a rare disease like Stargardt (known as an Orphan Drug) requires substantial, sustained capital. As a small biotech company with a handful of employees, Alkeus Pharmaceuticals faced the challenge of funding multi-year, multi-center clinical trials.
- **Resolution:** The team employed a strategic blended funding approach. Early academic work was launched with foundational grants from the BrightFocus Foundation. The company then successfully secured over \$2.0 Million in competitive federal grants from the NIH's National Eye Institute (NEI) and over \$2.1 Million from the U.S. FDA's Office of Orphan Products Development (OOPD) (grants like R01 FD004098 and R01 FD006316). This federal support de-risked the early clinical program, which later proved the concept in TEASE-1, enabling the company to then secure over \$150 Million in private investment to complete the final stages of the clinical replication (TEASE-3) and regulatory path. Alkeus Pharmaceuticals has now grown to 50+ employees.

Barrier 4: Patient Recruitment in a Rare Population

- **Description:** Recruiting genetically-confirmed, yet asymptomatic, children for the TEASE-3 replication study presented a logistical barrier due to the rarity of the condition and the challenge of treating children without symptoms.
- **Resolution:** The team built a strong collaboration with patient advocacy groups to reach this specific, rare patient demographic. The innovative trial design used affected, but untreated, siblings as controls, which maximized the clinical rigor of the small cohort, provided a highly relevant comparison, and streamlined the recruitment effort by focusing on families already engaged with the disease.

4.2 Organizational Levels

These barriers spanned the translational pipeline:

- Scientific/Methodological – model selection and validation (team & academic collaborators).
- Clinical/Strategic – trial design and patient stage selection (clinical sites).
- Financial/Organizational – blended grant and private funding (company leadership).
- Community/Logistical – recruitment via advocacy groups (patient community).

Collectively, the identified challenges and their systematic resolutions enabled successful replication from bench to bedside.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

The replication activities were executed over a sustained period of over a decade, moving strategically from foundational science to pivotal clinical trials. The timeline began with the Foundational Discovery Phase (2007–2014), during which the initial C20-D3-vitamin A concept was developed and published, showing a significant reduction in the toxic dimer A2E in an albino mouse model. This was supported by an NIH/NEI R01 grant award to Dr. Washington.

This was immediately followed by the Pre-clinical Validation Phase (2013–2015). A critical step in 2015 was the independent, external replication and publication in the pigmented *Abca4* mouse model, which verified the core mechanism and translated the findings into a more clinically relevant setting. Concurrently, the first clinical applications were being launched, with the initial FDA grant (FD004098) secured in 2013 to start the Phase 1 Study of ALK-001 for Stargardt Disease.

The program then progressed into the Early Clinical Replication Phase (2016–2017). This period included the start of the Phase 2a clinical trial in geographic atrophy (EY021988 funding in 2016 and 2017) and continued FDA funding for the Stargardt Phase 1 trial.

The most resource-intensive phase was the Core Clinical Replication (TEASE Program), which began in 2018 with the first grant funding (FD006316) for the ALK-001 Phase 2 Treatment of Stargardt Disease (TEASE). This funding continued from 2018 through 2024, supporting the execution of the TEASE studies, which targeted the crucial, early-stage patient cohort of TEASE-3. The extended clinical replication program culminated in 2023 with definitive data demonstrating a potential disease halting effect in pre symptomatic patients.

5.2 Associated Costs

The replication outcomes were achieved through a combination of federal grants and significant private investment, totaling over \$4.25 million in federal non-dilutive funds specifically identified for these trials.

Federal funding for the Stargardt replication efforts totaled approximately \$2.16 million across two key FDA grants:

- The initial Phase 1 Study of ALK-001 for Stargardt Disease (FD004098) received grants in 2013 and 2017 totaling \$415,984.
- The main Phase 2 Treatment of Stargardt Disease (FD006316), which funded the TEASE clinical replication, received grants in 2018, 2019, 2021, and 2023 totaling \$1,750,000.

In addition, a federal grant for a related indication, A Phase 2a clinical trial of ALK-001 in geographic atrophy (EY021988), provided over \$2.08 million (grants in 2016 and 2017) which supported the drug's overall safety and development profile. Beyond federal support, the later-stage, long-term clinical replication activities were financially supported by substantial private equity, notably including a Series B financing round announced in 2023 that provided over \$150 Million to complete the clinical program and prepare for regulatory filing.

5.3 Required Resources

Achieving the robust replication outcomes required a highly specialized convergence of resources:

- Financial Capital: The multi-year clinical trials necessitated sustained funding, provided by the over \$4.25 million in Federal grants and the later infusion of private capital.
- Specialized Pharmaceutical Compound: The drug itself, C20-D3-vitamin A (ALK-001), a chemically unique deuterated retinol, required specialized synthetic chemistry and Good Manufacturing Practice (GMP) production and quality control to ensure a pharmaceutical-grade product for human trials.
- Expert Personnel: A diverse team was essential, including: Academic Chemists/Pharmacologists (e.g., Dr. Washington) for developing and validating the drug's mechanism. Independent Pre-clinical Scientists

(e.g., Dr. MacLaren and team) for the rigorous, objective, and unbiased replication in a clinically relevant animal model. Clinical Operations and Regulatory Specialists (e.g., Dr. Saad and team) to design, manage, and execute multi-center, FDA-regulated clinical trials (TEASE-1 and TEASE-3).

- **Infrastructure and Equipment:** Access to specialized, quantitative measurement tools was critical for both pre-clinical and clinical replication, including quantitative fundus autofluorescence (qAF) and electroretinography (ERG) in animal models, and long-term, high-resolution fundus autofluorescence (FAF) and optical coherence tomography (OCT) imaging systems in clinical sites to track lesion growth and progression over years.
- **Patient Access:** Strong relationships with patient advocacy groups were required to successfully identify and enroll the rare population of pre-symptomatic Stargardt patients required for the pivotal TEASE-3 study.

6. COLLABORATION

6.1 Necessary Partnerships

The successful implementation and replication of the ALK-001 program required a multi-tiered network of partnerships across the academic, industry, and clinical sectors.

- **Academic Research Invention (Columbia University):** Essential for the foundational discovery, initial proof-of-concept, and the subsequent transfer of intellectual property (IP) to the developing company.
- **Independent Pre-clinical Validation (Oxford University):** Absolutely necessary to provide rigorous, unbiased replication of the core mechanism of action in the most clinically relevant animal models (e.g., pigmented Abca4 mice), which increased the scientific validity and regulatory confidence in the therapy.
- **Biotech/Pharmaceutical Partner (Alkeus Pharmaceuticals, Inc.):** Critical for manufacturing the pharmaceutical-grade drug, managing the complex regulatory process (IND application, FDA designations, Clinical trial sponsor), and securing the funding necessary to run multi-year clinical trials.
- **Federal and Foundation Funding Agencies (NIH and FDA):** Essential for securing non-dilutive grants that de-risked the early-stage clinical work, allowing the program to progress to the point where it could attract larger private investments.
- **Multi-center Clinical Trial Network:** A network of specialized retina centers and experienced Principal Investigators (PIs) was required to enroll the rare Stargardt patient population, manage the treatment, and collect the high-quality, long-term clinical outcome data (FAF, BCVA) needed for clinical replication.
- **Patient Advocacy Groups:** Necessary for connecting with and educating the rare disease community to facilitate the challenging recruitment of patients.

6.2 Leverage Existing Partnerships

Existing relationships were key to rapidly translating the discovery into the clinic:

- **Academic-Industry Transfer:** The foundational partnership between Dr. Ilyas Washington (Columbia University, academic discoverer) and Dr. Leonide Saad (Alkeus Pharmaceuticals, CEO) was leveraged to efficiently transfer the deuterated vitamin A IP. This allowed the smooth transition from lab bench to pharmaceutical development.
- **Federal Grant Expertise:** The duo leveraged their experience with NIH and FDA grant mechanisms (e.g., R01, R44) to secure over \$4.2 million in federal non-dilutive funding, which was critical for financing the clinical replication activities (TEASE-1, TEASE-3, and the related Geographic Atrophy trial called “SAGA”). These partnerships were built on a track record of successful research execution.
- **Scientific Peer Network:** Dr. Washington’s pre-existing scientific relationships and Dr. Saad’s business experience were leveraged to engage the team at the Nuffield Laboratory of Ophthalmology, University of Oxford (Dr. MacLaren and Dr. Charbel Issa) for the vital, independent pre-clinical replication. This external, objective validation, published in PNAS, was a powerful leverage point for subsequent regulatory and investor discussions.

6.3 Building New Partnerships

The most crucial steps in the replication process required the intentional formation of new partnerships:

- **Clinical Trial Network Establishment:** New partnerships were forged with retina specialists across the U.S. (e.g., Dr. Christine Kay at University of Florida, Dr. Michael Gorin at UCLA, Dr. Hendrik Scholl at Johns Hopkins) to establish the multi-center TEASE clinical trial network. These PIs and their institutions were essential for navigating the logistics of a rare disease trial, enrolling the diverse patient population, and ensuring the long-term follow-up required for the replication endpoints.
- **Patient Community Engagement:** New, formalized partnerships were established with major patient advocacy organizations, such as the Foundation Fighting Blindness (FFB). This collaboration was instrumental in identifying and recruiting the extremely difficult-to-find cohort of early-stage and pre-symptomatic patients needed for the high-impact TEASE-3 replication study.
- **Private Investment Partners:** Following the successful, grant-funded initial clinical replication phase (TEASE-1), new partnerships were built with venture capital firms including Bain Capital Life Sciences. These investors were convinced by the initial replicated data and committed over \$150 Million to finance the completion of clinical trials and prepare the company for future trials and commercialization.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Sustainability, Transferability, and Generalizability

Sustainability: The therapeutic effect of ALK-001 (gildeuretinol) has been shown in 24-month long clinical trials in TEASE-1. ALK-001 is administered as a once-daily oral capsule, avoiding the burdens of injectable biologics and supporting long-term patient adherence. Because the compound is a deuterated analogue of the natural nutrient vitamin A, it's been well-tolerated to date. Importantly, ALK-001 acts on the chemical root cause of Stargardt disease by slowing vitamin A dimerization and thereby reducing lipofuscin formation without disrupting the visual cycle. The TEASE-3 data demonstrate that treated patients also remained without further vision loss for up to seven years, confirming durable biological efficacy.

Transferability: The replication activities are highly transferable.

- **Genotype/phenotype:** Stargardt disease results from pathogenic variants in ABCA4. ALK-001 bypasses the defective protein and targets a universal biochemical reaction (vitamin-A dimerization), so the therapy is expected to be effective across the spectrum of ABCA4 mutations, ensuring broad applicability within the Stargardt population.
- **Geographic region:** As an oral medication that does not require specialized administration equipment or surgical facilities, ALK-001 could be distributed through standard pharmacy channels worldwide, facilitating adoption in any clinical setting.

Generalizability: The principle of slowing toxic vitamin A dimers extends beyond Stargardt disease to other retinal disorders where vitamin A dimer accumulation has been implicated in the disease pathology. This includes:

- **Geographic atrophy/Dry AMD:** In the SAGA trial, ALK-001 produced a statistically significant preservation of low-luminance visual acuity and a trend toward slower lesion growth, indicating that the mechanism translates to age-related macular degeneration, a far larger patient base.
- **Other dystrophies:** Cone rod dystrophy, retinitis pigmentosa, Best disease and certain pattern dystrophies also feature RPE vitamin A dimer accumulation; therefore, ALK-001 warrants investigation in these conditions as a potentially disease-modifying agent.

7.2 Current Sustainability Plans

The current sustainability plans are focused on regulatory approval and commercial transition to ensure patient access:

1. **Regulatory Approval and Priority Review:** Alkeus Pharmaceuticals is leveraging the FDA's multiple special designations (Breakthrough Therapy, Rare Pediatric Disease, Fast Track, and Orphan Drug) to accelerate the development and commercial submission of ALK-001 in the treatment of Stargardt disease.
2. **Manufacturing and Supply Chain:** Work is taking place to secure a validated supply chain for the large-scale, GMP manufacturing of ALK-001, ensuring long-term supply for the global patient population.

3. Long-term Follow-up (TEASE-4): The ongoing TEASE-4 open-label extension study is a core sustainability plan, offering continued access to ALK-001 and collecting multi-year, real-world data (now reaching up to 7+ years) to continue to generate data on the drug's long-term safety and efficacy profile.

7.3 Scaling and Application

The replication activities provide key insights for scaling and broader application:

- **Scaling Up the Stargardt Patient Population:** The success in the early-stage TEASE-3 cohort provides the blueprint for possible screening and early treatment. Application can be scaled up through ophthalmic screening or genetic testing of ABCA4 carriers, allowing a possible treatment initiation before symptoms appear, which would be an ultimate success metric.
- **Application to Geographic Atrophy (GA):** The data from the SAGA trial shows the potential to apply the mechanism to a much larger context: GA, which affects millions of patients. The mechanism of action is transferred by applying the same oral therapy to an older, non-genetic, but closely related pathologically patient population. The oral nature of the drug would make it easily scalable for the high-volume treatment required by the AMD population. A clinical trial in earlier stage AMD could be performed as well.
- **Application to Personalized Medicine:** The core principle of correcting an aberrant reaction rate (Vitamin A dimerization) could be applied to other metabolic disorders where minor differences in isotope weight could be used to slow detrimental enzymatic or chemical reactions without disrupting essential biological pathways.

7.4 Future Integration Plans

Future plans are focused on ensuring maximum therapeutic impact and global reach:

1. Continued Data Replication and Dissemination: committed to presenting and publishing the final, long-term data from the TEASE-3 replication study in peer-reviewed journals and international ophthalmology conferences (AAO, ARVO, ASRS). This continuous dissemination is crucial for educating the global medical community.
2. Targeted Physician Training Programs: committed to develop and implement training programs for comprehensive ophthalmologists and optometrists to recognize early signs of Stargardt disease and the need for genetic screening, shifting toward preventative care rather than intervention after irreversible vision loss.
3. Combination Therapies Replication: In the future, new replication studies could be designed to evaluate ALK-001 in combination with other emerging therapies (e.g., gene therapy, cell therapy, or other therapeutic modalities) for Stargardt disease, aiming to provide a synergistic benefit, especially for patients with advanced disease.
4. Information Gathering for Health Economics: Committed to conduct post-marketing studies to gather real-world evidence on the long-term impact on patient quality of life (QoL), to facilitate integration into institutional policy and national healthcare systems.
5. International Regulatory Integration: Initiate parallel regulatory filings (e.g., Europe, Japan) to ensure global patient access.

8. REFERENCES

1. Bagchi AD, Dale S, Verbitsky-Savitz N, Andrecheck S, Zavotsky K, Eisenstein R. Examining effectiveness of medical interpreters in emergency departments for Spanish-speaking patients with limited English proficiency: results of a randomized controlled trial. *Ann Emerg Med*. 2011;57(3):248-56 e1-4. Epub 2010/08/04. doi: 10.1016/j.annemergmed.2010.05.032. PubMed PMID: 20678825.
2. Heath M, Hvass AMF, Wejse CM. Interpreter services and effect on healthcare - a systematic review of the impact of different types of interpreters on patient outcome. *J Migr Health*. 2023;7:100162. Epub 2023/02/24. doi: 10.1016/j.jmh.2023.100162. PubMed PMID: 36816444; PubMed Central PMCID: PMCPCMC9932446.
3. Lopez Vera A, Thomas K, Trinh C, Nausheen F. A Case Study of the Impact of Language Concordance on Patient Care, Satisfaction, and Comfort with Sharing Sensitive Information During Medical Care. *J*

- Immigr Minor Health. 2023;25(6):1261-9. Epub 2023/02/26. doi: 10.1007/s10903-023-01463-8. PubMed PMID: 36840903; PubMed Central PMCID: PMCPMC9959935.
4. Reaume M, Labossiere MN, Batista R, Van Haute S, Tangri N, Rigatto C, et al. Patient-Physician Language Concordance and Cardiovascular Outcomes Among Patients With Hypertension. *JAMA Netw Open*. 2025;8(2):e2460551. Epub 2025/02/19. doi: 10.1001/jamanetworkopen.2024.60551. PubMed PMID: 39969882; PubMed Central PMCID: PMCPMC11840650.
5. Chu JN, Wong J, Bardach NS, Allen IE, Barr-Walker J, Sierra M, et al. Association between language discordance and unplanned hospital readmissions or emergency department revisits: a systematic review and meta-analysis. *BMJ Qual Saf*. 2024;33(7):456-69. Epub 2024/01/02. doi: 10.1136/bmjqs-2023-016295. PubMed PMID: 38160059; PubMed Central PMCID: PMCPMC11186734.
6. Bartlett G, Blais R, Tamblyn R, Clermont RJ, MacGibbon B. Impact of patient communication problems on the risk of preventable adverse events in acute care settings. *CMAJ*. 2008;178(12):1555-62. Epub 2008/06/04. doi: 10.1503/cmaj.070690. PubMed PMID: 18519903; PubMed Central PMCID: PMCPMC2396356.
7. Martin SS, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation*. 2024;149(8):e347-e913. Epub 2024/01/24. doi: 10.1161/CIR.0000000000001209. PubMed PMID: 38264914; PubMed Central PMCID: PMCPMC12146881.
8. Shannon EM, Blegen MB, Orav EJ, Li R, Norris KC, Maggard-Gibbons M, et al. Patient-surgeon racial and ethnic concordance and outcomes of older adults operated on by California licensed surgeons: an observational study. *BMJ Open*. 2025;15(3):e089900. Epub 2025/03/04. doi: 10.1136/bmjopen-2024-089900. PubMed PMID: 40032373; PubMed Central PMCID: PMCPMC11877244.
9. John-Baptiste A, Naglie G, Tomlinson G, Alibhai SM, Etchells E, Cheung A, et al. The effect of English language proficiency on length of stay and in-hospital mortality. *J Gen Intern Med*. 2004;19(3):221-8. Epub 2004/03/11. doi: 10.1111/j.1525-1497.2004.21205.x. PubMed PMID: 15009776; PubMed Central PMCID: PMCPMC1492154.
10. Kazi DS, Elkind MSV, Deutsch A, Dowd WN, Heidenreich P, Khavjou O, et al. Forecasting the Economic Burden of Cardiovascular Disease and Stroke in the United States Through 2050: A Presidential Advisory From the American Heart Association. *Circulation*. 2024;150(4):e89-e101. Epub 2024/06/04. doi: 10.1161/CIR.0000000000001258. PubMed PMID: 38832515.
11. Hanna M, Shaefer HL, Finegood E, Agarwal S, Zamani-Hank Y, LaChance J. Hardship and Hope: The Relationship Between Unconditional Prenatal and Infant Cash Transfers, Economic Stability, and Maternal Mental Health and Well-Being. *Am J Public Health*. 2025;115(12):2020-9. Epub 2025/09/11. doi: 10.2105/AJPH.2025.308244. PubMed PMID: 40934444; PubMed Central PMCID: PMCPMC12614019.
12. Reese TJ, Padi-Adjirackor NA, Griffith KN, Steitz B, Patrick SW, Leech AA, et al. Comparative effectiveness of buprenorphine adherence with telemedicine vs. in-person for rural and urban patients. *J Am Pharm Assoc (2003)*. 2025;65(2):102318. Epub 2024/12/30. doi: 10.1016/j.japh.2024.102318. PubMed PMID: 39733804; PubMed Central PMCID: PMCPMC11903133.

INDIVIDUAL WINNER

Christina Kackos Brunelle

A vaccine development and infectious disease researcher applying rigorous replication methods to strengthen confidence in published findings.

Christina Kackos Brunelle is a Senior Consultant at Deloitte with a background in infectious disease research. Dr. Kackos received her Ph.D. in Biomedical Sciences in 2022 at St. Jude Children's Research Hospital Graduate School of Biomedical Sciences. Dr. Kackos specializes in vaccine development, data management, and scientific communications.

1. REPLICATION STRATEGY**1.1 Description of Replication Activities**

Influenza viruses pose a unique public health danger due to their recurrent seasonal epidemics and pandemic potential. Influenza virus risk assessments, unknown virus identification, and characterization of immunological correlates of protection are performed using a common set of assays and animal models. Therefore, it is critical that these assays and models perform consistently across laboratories to ensure accurate results that reliably inform public health and therapeutic interventions.

In recognition of the importance of congruent results across laboratories, the influenza research community has undertaken efforts to understand the sources of experimental variability and to identify how these inconsistencies can be mitigated across 1) common influenza virus characterization assays and 2) the ferret model of influenza risk assessments. These activities fall under the basic category of research. For the purpose of these studies, the community defined replication as consistency in experimental results across participating laboratories. The specific metrics and evaluations used to define and assess consistency for each study are described below.

The replication studies we describe here are examples of how experimental variability is characterized by the flu community in two critical assays and an animal model. They leveraged pre-existing influenza research programs to improve the quality of their research and promote trust in their findings. Not only do they highlight the dedication of the field to high-quality research but also underscore the value of cultivating research networks with a shared mission and resources.

1.1.1 External Quality Assessment (EQA) of Hemagglutination Inhibition and Microneutralization Assays

Hemagglutination inhibition assays (HAI) and microneutralization assays (MN) are used to characterize influenza viruses as well as measure host antibody responses. An EQA study conducted by Waldock et al. was designed to assess intra- and inter-laboratory performance of these assays and engaged laboratories across industry, academia, public health, and contract research organizations [2]. Each participating laboratory was provided three pre- and twenty post-vaccination samples, as well as four duplicate samples, two serum standards composed of pooled human sera, and a negative control. Participating laboratories were instructed to perform HAI and MN assays with the provided samples using four common seasonal influenza viruses that were not provided with the samples. Additionally, laboratories were asked to assay the samples using either in-house or standardized HAI and MN protocols. Laboratories and results were anonymized to remove bias from statistical analysis.

HAI and MN results fall within a specific range of values based on dilutions used in the assays. Intra-assay variability was evaluated using maximum-minimum ratios of the four duplicate samples within the serum panel. Intra-laboratory variability was assessed using maximum-minimum ratios of three assay replicates. In both cases, a ratio of >3.5 was defined as an unacceptable level of intra-assay/lab variability and these samples were excluded from inter-laboratory comparisons. Additionally, any samples with a Geometric Mean Titer (GMT) below the limit of detection were also excluded from statistical analyses.

For inter-laboratory comparisons, a reference titer was calculated from the GMT of the sample across all laboratories. Individual laboratory titers were then divided by the reference titer to obtain a Geometric Mean Ratio

(GMR). The same approach was used to evaluate differences in assay protocols, however in this case the reference titer was defined as the GMT of a given sample when tested using each protocol.

The serum standard that had the most consistent results across laboratories was used to determine a calibration factor by calculating the ratio of the standard's titer in a given run to the GMT of the standard. This calibration factor was then applied to all other sample results to normalize the titers for more direct comparisons between laboratories.

Following data analysis, a report containing the results was provided to each participating laboratory and included evaluation of intra-assay, intra-laboratory, and inter-laboratory variability.

1.1.2 Ferret Models for Influenza Risk Assessment

Ferrets are a commonly used animal model for influenza virus research, particularly in the context of virus transmission, due to their natural susceptibility to influenza and development of clinical disease similar to humans [3]. However, their large size and complex husbandry needs necessitate small sample sizes, which when coupled with variability in experimental protocols between laboratories, as well as heterogeneity in the ferret population, can result in discrepant results. A 2022 study by Belser et al. assessed variations in ferret transmission results between laboratories when given a standardized set of protocols and viruses to evaluate the robustness of the model [4].

In this study, laboratories were given two pre-characterized viruses and standardized protocols for virus preparation, donor ferret infection (i.e., virus dose, route of infection, volume), and contact ferret exposure. Four donor-contact pairs were evaluated per virus per laboratory, and respiratory samples and clinical data (temperature and weight) were collected throughout the study period (fourteen days) alongside environment data, including room temperature and humidity. Sera was collected at the end of the experiment. Respiratory samples were assayed using three methods to obtain viral titers and these titers were normalized to 50% tissue culture infectious dose (TCID₅₀) per milliliter. Endpoint sera were assayed via HAI using in-house protocols. All data were anonymized and statistical analyses were performed by an independent statistician.

Transmissibility was categorized as moderate to high ($\geq 50\%$ transmission across donor-contact pairs) or low ($\leq 25\%$ transmission across donor-contact pairs). Concordant results were defined as a virus exhibiting the same transmissibility across laboratory ferret groups and discordant results were those in which a virus displayed different transmissibility across the study.

2. ENVIRONMENT

2.1 External Quality Assessment (EQA) of Hemagglutination Inhibition and Microneutralization Assays

This study engaged sixteen laboratories across the United Kingdom, United States, Australia, Japan, South Africa, France, Norway, Italy, Spain, and the Netherlands through the FLUCOP project [5]. The study was designed by Joanna Waldo, Catherine Caillet, and Othmar Engelhardt. Laboratory work was performed by researchers across collaborating institutions and all authors reviewed and approved the manuscript prior to submission. A complete list of authors and contributions can be found in the publication [4].

The study was initiated during the SARS-CoV-2 pandemic and results were published on February 28, 2023. The study was funded by the Innovative Medicines Initial Joint Undertaking (IMI JU) with additional support from the European Union Seventh Framework Programme, European Federation of Pharmaceutical Industries and Associations companies' in-kind contributions, and the Melbourne WHO Collaborating Centre for Reference and Research on Influenza, funded by the Australian Government Department of Health.

2.2 Ferret Models for Influenza Risk Assessment

Eleven laboratories within the United States, Hong Kong, Australia, China, and The Netherlands participated in the study. The study was organized under the working group on the standardization of the ferret model for influenza risk assessment and was led by Jessica Belser, Eric Lau, and Hui-Ling Yen. A complete list of authors and institutions is provided in the publication.

Funding was provided by the National Institute for Allergies and Infectious Disease Centers for Excellence for influenza Research and Surveillance (CEIRS) [6]. The study was published in August of 2022.

3. IMPACT

3.1 External Quality Assessment (EQA) of Hemagglutination Inhibition and Microneutralization Assays

For the HAI assay, 98% of minimum-maximum ratios fell within the acceptable range when analyzing intra-assay results, demonstrating that intra-assay performance was consistently good across laboratories. Similarly, 90% of intra-laboratory results had a ratio of <3.5 when assayed against influenza A viruses. Variability increased when influenza B viruses were used with 12.6% and 15.1% of B Victoria and B Yamagata results excluded from further analyses, respectively. These results were consistent with inter-laboratory variability that also found results were more consistent for influenza A viruses than B. The authors suggest that this may be due to differences in antigens used by each laboratory. This is further supported by the effect of standard inclusion on results. When normalized to the serum standard, variation was reduced across all viruses, with the most significant reduction of 82% variability observed in B Yamagata results. Finally, the authors found that use of a standardized protocol improved variability in intra-assay and intra-laboratory results but were only marginally effective at promoting inter-laboratory consistency.

Similar intra-assay trends were found when MN data were analyzed, though variability was observed at higher rates than HAI. The MN assay is inherently more prone to variability due to the more complex nature of the assay. Notably, intra-laboratory variability was consistent across laboratories without subtype-specific differences. Inter-laboratory variability of MN results was elevated compared to HAI results, and again, most prominent when influenza B viruses were assayed. However, similar reductions in variability were observed once results were normalized to the serum standard as in HAI. Lack of sufficient data points prevented comparison of experimental protocols.

3.2 Ferret Models for Influenza Risk Assessment

Transmissibility of both viruses used in the study – a representative of the 2009 H1N1 pandemic influenza virus (A(H1N1)pdm09) and an avian influenza virus (ruddy turnstone/09) – was first evaluated by detection of infectious virus and presence of antibodies in contact animals. Infectious virus and seroconversion were found in 32/44 (72.7%) and 35/44 (79.5%) of contacts exposed to A(H1N1)pdm09, respectively, indicating high transmissibility. In contrast, ruddy turnstone/09 was found to have low transmissibility with only 6/43 (14%) contact ferrets positive for infectious virus or seroconversion to the virus. Quantification of viral titers found that ruddy turnstone/09 viral particles were shed at lower levels than A(H1N1)pdm09.

The authors then evaluated environmental conditions of each laboratory to determine if these parameters affected transmissibility and found that while temperature was consistent across groups humidity varied greatly. However, no statistically significant associations between virus titers over the 14-day sampling period and humidity were found. Additionally, experimental cage setups varied across laboratories, including cage dimensions, air flow, and distance between cages. Laboratories with directional airflow reported higher transmissibility than those without airflow directionality, but no statistically significant difference between these groups were found. Other cage parameters, such as size and spacing, were also not found to have a significant correlation with transmissibility.

Using these results, the authors next evaluated how confidence in virus transmissibility is impacted by aggregating data from multiple groups using two methods. The first method assumes that a virus will have moderate to high or low transmissibility. If results are concordant across laboratories and then pooled, as few as three donor-contact pairs are needed to produce a probability of $>80\%$ that the virus will be moderately to highly transmissible when transmission is observed in $>75\%$ of animals. When ≤ 1 transmission event is observed there is an 85% probability that the virus has low transmissibility. When this method was applied to the results of the study, confidence in the transmissibility of A(H1N1)pdm09 and ruddy turnstone/09 were $>99.9\%$ and 94.8% , respectively.

The second method utilized a voting system in which a conclusion (vote) is drawn based on transmissibility in each laboratory and an overall conclusion is reached based on these votes. Assuming four donor-contact pairs per laboratory, the authors found three laboratories were needed to conclude moderate to high transmissibility with $>90\%$ confidence when votes are concordant. Five concordant votes are needed to predict low transmissibility with $>90\%$ confidence, indicating more laboratories are needed to identify low transmissibility

viruses than moderate to high. When the results of the study were applied to this method, the confidence in the conclusion that A(H1N1)pdm09 is moderately to highly transmissible was 99.8% and that ruddy turnstone/09 was lowly transmissible was 90.1%

Based on these findings, the authors concluded that results were concordant for both A(H1N1)pdm09 and ruddy turnstone/09 with high and low transmissibility, respectively. Therefore, laboratories with different environments and facility designs can reach consistent results when using standardized experimental protocols and viruses.

3.3 Conclusions

Inconsistencies in biological research are expected and unavoidable. These variations stem from multiple sources including differences in research facilities, model systems, experimental protocols and techniques, reagents, and more. Some of these variations are inherent and can be key to the research questions being asked, such as differences in human samples, animal subjects, and even virus replication dynamics, as already discussed. It is critical to disentangle these acceptable complexities in study design from inconsistencies in methods and reagents.

These two studies are exemplars of the work being done by the influenza community to better understand and improve replication within the field. The findings of these studies emphasize that while some inherent variability is unavoidable, it can be mitigated through the use of standardized protocols, reagents, and normalization of results to standard samples. The studied methodologies are regularly used to support global risk assessments of influenza virus strains, which in turn helps public health leaders make crucial decisions for pandemic preparedness. It is therefore critical that we understand the reliability and inherent variability in these methods. The influenza research community is uniquely positioned to continue characterizing and implementing these measures due to their well-established collaborations through externally funded research networks.

Publication of these studies is the first step in ensuring their impact. By making the study designs and conclusions publicly available, the authors have provided a framework for others to implement similar efforts into their own research. Additionally, these studies have provided evidence to support the validity of results generated from HAI and MN assays and the ferret model, while highlighting potential methodological pitfalls that should be reviewed critically when interpreting these data.

4. RESPONSE TO BARRIERS

4.1 External Quality Assessment (EQA) of Hemagglutination Inhibition and Microneutralization Assays

The authors initially reached out to forty-two laboratories to participate in the study. Sixteen laboratories performed the assays and provided results; the authors attribute this low response rate to the SARS-CoV-2 pandemic. Additionally, the use of human sera in the study inserts additional uncertainty in the results as humans have complex cofactors that influence immunity and assay outcomes, such as comorbidities and variable immune histories. To mitigate these effects the authors implemented the use of GMTs compared to individual results from each laboratory in their statistical analyses. The authors acknowledge that the limited number of participating laboratories does reduce the effectiveness of this method but were still able to draw conclusions about intra-assay, intra-laboratory, and inter-laboratory replication.

4.2 Ferret Models for Influenza Risk Assessment

Standardization of all experiment parameters was impossible due to institutional differences in animal facilities and research staff. Therefore, the authors identified factors that could be easily controlled and documented all variable factors. Additionally, the ferret model is challenging to work with and expensive, limiting the number of laboratories that could participate in the study. While the results obtained from the eleven participating institutions are convincing, the authors note that additional studies expanding the number of laboratories and influenza viruses used for infection should be expanded.

5. TIMELINE & BUDGET

5.1 External Quality Assessment (EQA) of Hemagglutination Inhibition and Microneutralization Assays

The study was initiated during the SARS-CoV-2 pandemic and results were published on February 28, 2023. The study was funded by the Innovative Medicines Initial Joint Undertaking (IMI JU) with additional support from the European Union Seventh Framework Programme, European Federation of Pharmaceutical Industries and Associations companies' in-kind contributions, and the Melbourne WHO Collaborating Centre for Reference and Research on Influenza, funded by the Australian Government Department of Health [4].

Access to sufficient quantities of sera to send to each participating laboratory were required to complete the study. The partnerships and research network that made the study possible will be detailed in the next section.

5.2 Ferret Models for Influenza Risk Assessment

Funding was provided by the National Institute for Allergies and Infectious Disease Centers for Excellence for influenza Research and Surveillance (CEIRS) [6]. The study was published in August of 2022.

Ability to properly house and execute ferret experiments was necessary for participating laboratories to complete the study.

6. COLLABORATION

6.1 Key Collaborations

Neither study would have been possible without FLUCOP and CEIRS (now CEIRR). Centers of Excellence for Influenza Research and Response (CEIRR) is an international network composed of Centers whose purpose is to study influenza transmission and pathogenesis as well as immune responses to infection and vaccination. Research conducted by CEIRR Centers is used to support influenza outbreak preparedness and response [7]. The FLUCOP project was specifically designed to improve and standardize existing assays used in influenza research with the goal of using these results to develop new and better assays [5].

Though collaboration is a cornerstone of the scientific enterprise, establishment of research networks promotes partnerships between laboratories while minimizing friction in establishing these relationships. Additionally, these funding mechanisms can promote reproducibility and replication in science by requiring awarded institutions to engage in efforts such as those presented here.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Current and Future Sustainability

As protocols for assessing assays and models have been established through these studies, and robust research partnerships already exist within the FLUCOP and CEIRR programs, similar efforts should be undertaken to standardize additional, commonly used methods and models across research communities, especially communities that inform public health preparedness or national security. However, studies alone are not enough. A concerted effort from each research community will be needed to implement the optimized protocols and reagents, including consensus on technical approaches, training for researchers, and validation of reagents and biological materials that can be easily obtained through existing repositories, such as BEI [8]. Data management and sharing practices will also need to be incorporated to allow researchers to replicate experiments, perform secondary analysis of data, and evaluate standardized protocol/reagent performance. Many programs, including CEIRR and the Collaborative Influenza Vaccine Innovation Centers (CIVICs) include data sharing frameworks and workflows, further highlighting the value of large-scale research networks in improving research quality and transparency.

These studies, and the FLUCOP and CEIRR programs, can serve as a model for other fields to understand the sources of variability in their key assays. This will require investment from funders and buy-in from the research communities, but the merit of these programs on expanding the breadth and quality of scientific research is undeniable.

8. REFERENCES

1. Spackman, E., Cardona, C., Muñoz-Aguayo, J., & Fleming, S. (2016). Successes and Short Comings in Four Years of an International External Quality Assurance Program for Animal Influenza Surveillance. *PloS one*, 11(10), e0164261. <https://doi.org/10.1371/journal.pone.0164261>
2. Waldoock, J., Weiss, C. D., Wang, W., Levine, M. Z., Jefferson, S. N., Ho, S., Hoschler, K., Londt, B. Z., Masat, E., Carolan, L., Sánchez-Ovando, S., Fox, A., Watanabe, S., Akimoto, M., Sato, A., Kishida, N., Buys, A., Maake, L., Fourie, C., Caillet, C., ... Engelhardt, O. G. (2023). An external quality assessment feasibility study; cross laboratory comparison of haemagglutination inhibition assay and microneutralization assay performance for seasonal influenza serology testing: A FLUCOP study. *Frontiers in immunology*, 14, 1129765. <https://doi.org/10.3389/fimmu.2023.1129765>
3. Buhnerkempe, M. G., Gostic, K., Park, M., Ahsan, P., Belser, J. A., & Lloyd-Smith, J. O. (2015). Mapping influenza transmission in the ferret model to transmission in humans. *eLife*, 4, e07969. <https://doi.org/10.7554/eLife.07969>
4. Belser, J. A., Lau, E. H. Y., Barclay, W., Barr, I. G., Chen, H., Fouchier, R. A. M., Hatta, M., Herfst, S., Kawaoka, Y., Lakdawala, S. S., Lee, L. Y. Y., Neumann, G., Peiris, M., Perez, D. R., Russell, C., Subbarao, K., Sutton, T. C., Webby, R. J., Yang, H., Yen, H. L., ... Working group on the standardization of the ferret model for influenza risk assessment (2022). Robustness of the Ferret Model for Influenza Risk Assessment Studies: a Cross-Laboratory Exercise. *mBio*, 13(4), e0117422. <https://doi.org/10.1128/mbio.01174-22>
5. FLUCOP – standardisation and development of assays for assessment of influenza vaccine correlates of protection. (n.d.). <https://flucop.eu/>
6. Centers of Excellence for Influenza Research and Response (CEIRR). (2024, June 14). Nih.Gov. <https://www.niaid.nih.gov/research/centers-excellence-influenza-research-response>
7. CEIRR Network. (n.d.). www.ceirr-network.org. Retrieved July 9, 2026, from <https://www.ceirr-network.org/>
8. BEI Resources Repository. (2024, September 27). Nih.Gov. <https://www.niaid.nih.gov/research/bei-resources-repository>

INDIVIDUAL WINNER

ConductScience



A research-instrument manufacturer that builds tools to the exact specifications of published methods, so labs can replicate experiments with verified, citable equipment.

ConductScience is a science-manufacturing company that builds replication-grade research instruments across the life sciences. The catalog spans behavioral apparatus, microfluidic chips, histology systems, microscopy, computer vision, automated procedure software, and a methods publishing platform. Founded in 2013 with MazeEngineers, ConductScience manufactures every instrument to the methods-section spec published in the originating paper, with every change tracked, dated, and citable. Instruments are deployed across more than 1,200 research institutions in over 50 countries and cited in more than 280 peer-reviewed publications.

Shuhan He, MD (Founder) leads replication strategy. Louise Corscadden, PhD is Director of Science and Development and Lead of the Technology Transfer Program. Boyu Peng, MS leads ConductVision, the science-first computer-vision platform. Santosh Adhikari, MS leads computer-vision algorithms. Yaning (Abby) Zheng, MS leads the scientometric platform that translates the world's scientific literature into a procurement layer for replicable science. ConductScience Fellows Annelise Silva, Andres Diaz, and Lauren Beaudin lead Scholarly Education, Research, and Open Data Standards at ConductScience.org.

1. REPLICATION STRATEGY

1.1 Strategy Description

Maze-based assays are foundational to preclinical behavioral neuroscience, providing primary measures of learning, memory, anxiety-like behavior, and motivation in animal models [1,2]. In routine research practice, however, nominally identical maze tasks frequently yield non-comparable results across laboratories, even when studies report using the same “standard” assay [3]. Small and often undocumented differences in apparatus construction, materials, lighting and sensors, procedural timing and handling, and analytic definitions (for example, zone entry or trial completion) can meaningfully alter measured behavior, undermining cross-study comparability and interpretation [4-6]. This strategy emerged from repeated cross-laboratory attempts to reproduce nominally identical rodent maze assays, revealing how undocumented differences in apparatus, execution, and analytic definitions undermine cross-site comparability. We address this gap with an integrated replication framework that preserves complete methods across sites by standardizing five elements: (1) apparatus, (2) procedures, (3) analytic definitions, (4) translation of investigator-developed methods into deployable standards, and (5) a published record that preserves these components for reuse. This submission documents an operational exemplar maintained for more than a decade across independent laboratories and evaluated using field-level benchmarks, within-laboratory transitions, and external validation.

Replication activity 1: Hardware Fidelity (via MazeEngineers)

MazeEngineers is a custom physical apparatus construction and standardization service for maze-based behavioral assays. It addresses hardware-level variation introduced when “standard” equipment is locally rebuilt or inconsistently sourced [6]. Apparatus designs are anchored directly to specific, highly cited reference papers and translated into precise, build-ready specifications tied to those citations rather than offered as generic devices. For assays with substantial historical variation, structured libraries of citation-specific variants are released to standardize materials, surface finish, and sensor placement while preserving protocol-level differences.

Replication activity 2: Procedural Fidelity (via ConductMaze)

ConductMaze is a software system for automated protocol execution and task control. It standardizes procedural execution by replacing manual trial timing and operator-driven task progression that can vary across runs and sites. Task steps, decision rules, and timing logic are defined in advance and executed by the system in response to the animal's behavior, enabling procedures to be reproduced, shared, and executed consistently across laboratories without reliance on operator judgment.

Replication activity 3: Analytic Fidelity (via ConductVision)

ConductVision is a video-based behavioral analysis software system that generates standardized tracking and outcome measures. It computes behavioral variables using predefined, documented analytic rules and exports results directly to the Neurodata Without Borders (.NWB) community standard [7]. Each dataset preserves both measured values and the metadata required to interpret and recompute outcomes, enabling behavioral data to be inspected, compared, and reused across laboratories using common tools.

Replication activity 4: Research Tool Technology Transfer

This activity is a technology transfer pathway for research tools. Investigator-developed prototypes are translated into deployable method packages that combine apparatus specifications, fixed procedure logic, and defined analytic outputs, reducing reliance on narrative descriptions or restrictive licensing [8]. This pathway has been validated through the formal translation of citation-locked academic inventions into standardized research tools.

Replication activity 5: Method Dissemination and Reuse (via ConductScience.org)

ConductScience.org is a web-based dissemination and publishing infrastructure for reusable research methods. Methods are published as complete packages with manuscripts submitted alongside method-defining assets (protocols, datasets, and design files), issued persistent identifiers (DOIs), and rendered for inspection of method-critical materials. This preserves complete specifications and reduces site-to-site variation introduced by inference or reconstruction (Table 4) [32].

Replication-enabling capability	ConductScience.org	Traditional publishing	Contribution to replication
Semantic enrichment of method assets	Method components are linked to normalized terms and persistent identifiers	Method descriptions are typically provided as free text without enforced identifier binding	Stabilizes the meaning of variables and method components across studies, software versions, and institutions
Alignment to Common Data Elements	Variables and units are mapped to NIH-aligned Common Data Elements	CDE alignment is not routinely supported	Improves cross-study comparability and supports aggregation beyond the originating study
Repository connectivity	Methods and associated assets maintain bidirectional links with external repositories	Data are commonly attached as supplementary files or static links	Preserves the relationship between methods and underlying data over time
Inspection of method-critical files	Complex assets such as 3D designs, structured protocols, and datasets are viewable directly within the publication	Assets are limited to static figures or downloadable files	Allows independent inspection of method-defining materials without author mediation
Structured sharing and reuse workflows	Machine-readable metadata and standardized export pathways support consistent reuse	Sharing practices vary by author and journal	Improves portability, auditability, and downstream reuse
Publishing of complete method packages	Manuscript, protocol files, datasets, and metadata are published together as a unified method package	Manuscripts are published independently of many method-defining assets	Reduces omissions and lowers barriers to faithful re-execution

Table 4. ConductScience.org publishing capabilities that support replication, compared with traditional publishing

1.2 Methods Used To Evaluate Outcomes

To assess whether replication activities 1–5 produced measurable improvements, we used complementary quantitative and qualitative outcome measures aligned with the Impact section, spanning pre/post comparisons, research rigor and transparency, dissemination and adoption, and independent external validation. Field-level quantitative benchmark comparison We conducted a retrospective analysis of peer-reviewed publications citing

MazeEngineers or ConductScience apparatus. Because pre-standardization datasets from deploying laboratories are not available retrospectively, we used Crabbe, Wahlsten, and Dudek (1999) as the pre-replication reference because it is a widely cited, purpose-built multi-laboratory study that directly quantified site-to-site variability under intentionally “nominally identical” behavioral assay conditions across independent laboratories [3]. This makes it an appropriate historical baseline for the magnitude of uncontrolled inter-laboratory variability that hardware standardization aims to reduce. We screened 174 open-access papers, restricted extraction to assay classes evaluated in the benchmark (Elevated Plus Maze, Elevated Zero Maze, Open Field Test, Y-Maze), and identified 58 independent control cohorts from 41 publications. For each cohort, coefficient of variation (CV) was calculated from reported means and dispersion (converting SEM to SD where required) and compared to the benchmark using one-sample t-tests [3].

Site-level qualitative longitudinal evidence

We summarized pre/post transitions using user-reported observations from ConductVision adopters describing changes in timing precision, robustness under real facility conditions, and perceived data consistency relative to prior manual or semi-manual scoring workflows. Analytic reproducibility and data integrity We tested conversion of ConductVision and ConductMaze outputs to the .NWB community open standard, evaluating preservation of event intervals, metadata required for recomputation, and semantic mapping via a translation dictionary and ontology identifiers. Dissemination, adoption, and long-term impact. We audited order records to quantify deployment scale and assay breadth, conducted a bibliometric analysis of peer-reviewed publications citing standardized tools, and tracked university inventions from licensing through independent laboratory deployment as standardized method packages.

Independent external validation

We monitored independent scientific press coverage and federal or academic programmatic recognition as external indicators of field-level credibility and validation of the replication strategy. Innovative elements The core innovation was treating replication as standing research infrastructure rather than a one-off experimental effort. This included creating a standardized way to build apparatus directly from published protocols, redesigning task automation and video analysis to produce open, reusable outputs, developing a new technology transfer pathway that allows research tools to be easily shared, and introducing a publishing workflow that deposits complete method materials alongside manuscripts so they can be directly inspected and reused by others.

1.2 Working Definition of Replication

Replication is defined as the ability of independent laboratories to re-execute the same method with sufficient specification of the apparatus, procedure, and analysis so that results can be directly compared across sites and yield consistent measurements when addressing the same experimental question.

Category Identification: Primary category: Pre-clinical research. Secondary category: Basic research.

2. ENVIRONMENT

2.1 Activity Context

Unlike traditional academic cores that are siloed within a single institution and often optimized for local service and cost recovery rather than cross-institution governance and comparability, this replication activity maintains authoritative reference standards for apparatus, procedures, and data that independent laboratories apply under real-world conditions across diverse facilities [9]. Location, timing, and team details Locations and facilities. ConductScience’s engineering, manufacturing, and method standardization work is performed in-house. The team maintains design files, performs in-house assembly and testing of behavioral apparatus, verifies automated protocol performance, and develops and deploys ConductVision using dedicated server and GPU infrastructure. Designs are reviewed for animal safety in consultation with veterinary expertise.

2.2 Timing and Duration

This project represents a mature, continuously maintained implementation spanning more than a decade (2013 to present). Core replication capabilities were established sequentially and remain active in parallel: apparatus

replication beginning in 2013, procedure replication in 2016, replication-focused technology transfer in 2018, analytic replication in 2023, and dissemination and reuse infrastructure in 2025. Each capability, once introduced, has been sustained, expanded, and integrated into a unified replication ecosystem rather than completed and retired as a discrete phase.

2.3 Team

The replication activities were implemented by an interdisciplinary team combining academic expertise in behavioral neuroscience and informatics with industry experience in high-reliability systems, enabling replication to be executed and sustained as standing research infrastructure.

Shuhan He, MD (Founder, MazeEngineers and ConductScience) is an emergency physician and clinical informaticist. He is responsible for the overall replication strategy, including literature-anchored apparatus specifications, automated procedure standards, analytic definitions, and cross-version method alignment across laboratories.

Henry He, PhD (Chief Technology and Operations Officer) is a PhD trained computer scientist with prior senior engineering leadership in time critical, high reliability systems. He leads software architecture for automated protocol execution, event logging, and scalable deployment to ensure timing fidelity, traceability, and operational robustness across sites. He also oversees standardized hardware operations, including manufacturing workflows, supply chain documentation, build verification, and release controls, ensuring physical equivalence, traceability, and continuity of behavioral apparatus across vendors, time, and institutions.

Boyu Peng, MS (Lead Developer, Computer Vision) is a computer engineer with graduate training in computer science and professional experience in software development. He leads ConductVision implementation, including camera integration, tracking pipelines, and standardized behavioral output generation.

Louise Corscadden, PhD (Director of Science, Development, and Validation) is a molecular geneticist with experience across the life sciences. She leads academic partnerships and technology transfer, translating investigator-developed prototypes into standardized, executable tools, and oversees analytic validation, behavioral variable definitions, metadata standards, and method reuse as citable research objects.

2.4 Funding Support

Direct federal funding was not used to build the core infrastructure; it is sustained through ongoing engineering, manufacturing, software maintenance, and governance, preventing replication assets from sunseting with any single grant.

3. IMPACT

3.1 Pre/Post Impact

Field-level behavioral variability (hardware fidelity) Using a published multi-laboratory benchmark that quantified site-to-site variability when nominally identical mouse behavioral assays were run across independent laboratories as the pre-replication reference [3], post-replication cohorts derived from independent laboratories using citation-locked ConductScience/MazeEngineers apparatuses showed lower coefficients of variation across multiple assays. Across all extracted control cohorts, mean coefficient of variation was 0.41 ± 0.27 for MazeEngineers hardware compared with the historical benchmark value of 0.76 from Crabbe et al. [3]. Lower control-cohort variability increases power for detecting true effects and reduces the risk that cross-site differences reflect uncontrolled apparatus variation rather than biology. This benchmark comparison isolates hardware-linked variability and does not capture all sources of procedural or analytic variation.

3.2 Within-Laboratory Pre/Post Replication Transitions: Routine-Use Robustness

To complement the quantitative comparisons, we conducted qualitative longitudinal case studies based on routine-use feedback from independent research sites using ConductVision. While the quantitative analysis defines pre- and post-replication states at the field level, these site-level reports capture the pre/post transition

within individual laboratories as they moved from manual, locally implemented scoring workflows to standardized, automated analysis. Investigators at an independent research site described a transition from variable manual scoring to “precise timing, reproducible tracking... [and] a clear improvement in data consistency” following adoption of ConductVision. Similarly, users at another independent research site reported “consistent, reproducible data, even under variable lighting and water conditions,” indicating improved robustness relative to their prior manual and semi-manual workflows. Although qualitative, these longitudinal site-level observations align with the quantitative reductions in inter-laboratory variability and provide convergent evidence that standardized apparatus, automated procedures, and explicitly defined analysis rules reduce scoring differences and improve consistency in everyday laboratory practice.

3.3 Research Rigor and Analytic Transparency Improvements

Analytic Reproducibility, Data Integrity, and Standards Compliance Analytic replication often breaks down because different laboratories use the same outcome names, such as “time in zone” or “immobility,” while applying different, undocumented rules, and because key details are lost when proprietary analysis files are reduced to spreadsheets [5,6]. We demonstrate that ConductVision and ConductMaze outputs can be mapped into the community .NWB format with event timing, analytic parameters, and semantic meaning preserved by scoring a representative behavioral video into 34 Start/Stop rows corresponding to 17 complete bouts of behavior. After conversion, the .NWB export contained exactly 17 TimeIntervals, matching the source bouts one to one, with every boundary preserved both as timestamps (seconds) and frame indices, no dropped events, and no unpaired starts or stops, while retaining trial identifiers, maze context, and acquisition metadata. For each exported variable, we generated a translation dictionary specifying its canonical location within the .NWB file and its corresponding community ontology identifier (Table 1). In the demonstration dataset, all behavior labels mapped to ontology with 100% mapping completeness, ensuring consistent interpretation across laboratories and software versions even when local names differ. Taken together, the .NWB file, preserved scoring parameters, and translation dictionary allow an independent laboratory to inspect event boundaries, recompute derived measures, and confirm that reported variables reflect the same operational definitions as the originating site when results are compared across studies.

ConductVision / ConductMaze output	NWB path (module)	NWB interface	Ontology term (NBO ID)	Value type
ZoneEntry	processing/behavior	BehavioralEvents or TimeIntervals	None in NBO (spatial event defined by arena zones)	event timestamp (s) and frame index
ZoneExit	processing/behavior	BehavioralEvents or TimeIntervals	None in NBO (spatial event defined by arena zones)	event timestamp (s) and frame index
Immobile	processing/behavior	BehavioralTimeSeries or TimeIntervals	locomotor inactivation behavior (NBO:0009002)	samples or bouts
Velocity	processing/behavior	Position (or a TimeSeries under behavior)	locomotory behavior (NBO:0000013)	time series
Rearing	processing/behavior	BehavioralEvents or TimeIntervals	exploration behavior (NBO:0000100; broader term)	events or bouts

Table 1. Translation dictionary linking exported ConductVision / ConductMaze variables to NWB locations, Neuro Behavior Ontology (NBO) terms, and value types

3.4 Dissemination, Adoption, and Long-Term Impact

3.4.1 Adoption and Scale of Replication Infrastructure

The replication strategy has been adopted at field scale, by independent research organizations. Adoption has continued over this period, indicating sustained integration into routine research workflows rather than transient pilot use. Deployments span multiple major behavioral paradigms, demonstrating that the standardization framework generalizes beyond a single assay class.

3.4.2 Scientific Utilization and Downstream Knowledge Production

To assess scientific use and downstream impact, we conducted a bibliometric audit of peer-reviewed publications that explicitly cite MazeEngineers, ConductScience, or standardized tool variants in the Methods section. Standardized apparatus and protocols were found to be widely cited in peer-reviewed work spanning anxiety, learning and memory, addiction, stroke models, and neurodegeneration, indicating that they are being incorporated into primary experimental workflows rather than referenced solely as background infrastructure.

3.4.3 Transferability Beyond Originating Laboratories

Invention Replication Case Studies

To determine whether this replication approach applies beyond well-established assays, we followed academic inventions as they moved through the Pillar 4 pathway. Six university-developed methods progressed from use in a single originating laboratory to standardized tools adopted by independent external laboratories.

3.5 Independent External Validation and Credibility Signals

Contemporaneous third-party coverage provides external evidence that this strategy was publicly recognized as addressing replication and cross-laboratory comparability challenges at the time it was implemented. Nature Toolbox (2018), in “The mazes with minds of their own,” described the approach as providing “standardized equipment and tasks so that researchers can compare their results directly,” explicitly framing cross-study comparability as a core scientific objective [13]. Simultaneously, mainstream science reporting recognized this standardization as a response to broader reproducibility concerns. A Yahoo feature (2017), “A History of the Morris Water Maze, From Inception to Industry Standard,” identified variability arising from bespoke institutional fabrication as “a potential confound for comparing results across institutions, labs and studies,” and presented standardized apparatus as a practical mechanism to reduce that confound [14]. Discover Magazine (2018) described this automated, standardized maze approach as reducing reliance on subjective observer judgment and producing “more objective and standardized” data, while Popular Science (2018) similarly highlighted these maze designs as reducing or eliminating human involvement and thereby improving rigor and cross-study comparability [15,16]. Because this coverage (2017–2018) significantly predates the current NIH Replication Prize and related replication-focused initiatives, it serves as external validation that cross-site replicability has been a foundational and publicly recognized objective of this strategy since its inception.

3.5.1 Federal and Academic Programmatic Validation

The replication and technology transfer framework has received external validation through competitive academic and federal programs. Harvard Innovation Labs selected the replication-focused technology transfer model (Activity 4) for its Build It program, recognizing revenue-based dissemination of academic research tools [17]. In 2025, NIH selected the team as a Phase I winner of the Data Sharing Index (S-Index) Challenge [18].

4. RESPONSE TO BARRIERS

Barrier 1: Missing Economic Models for Long-Term Replication:

Institutional and Community Level: Sustaining replication at scale is an infrastructure problem that requires continuous operational support, including maintaining authoritative reference specification, validating supply chains, and updating software over time. Federal funding mechanisms such as R01 and R21 awards are structurally time-limited and designed for discrete hypothesis-driven projects, meaning the authority and capital to maintain infrastructure often ends when an award sunsets [20]. Although the NIH Research Tools Policy explicitly

encourages dissemination, it provides no standing mechanism to fund the ongoing manufacturing, software maintenance, and version governance required to keep methods executable over time [21]. In parallel, federal cost principles under 2 CFR § 200 discourage universities from operating standing, cross-campus, or externally facing production infrastructure, because strict requirements for cost allowability and allocability make it difficult to sustain shared infrastructure that is not cleanly attributable to a single award [22, 23]. Because these constraints leave no viable academic or grant-based pathway for long-term replication infrastructure, we operate an organizational hybrid—commercial in structure but a replication core in mission—that replaces the grant cliff with continuous funding for manufacturing continuity, software upkeep, and version control [9].

Barrier 2: The “Trust Gap”: Establishing Scientific Legitimacy as a Non-Academic Entity Level:

Institutional and Team Level: Because our solution to the “grant cliff” (Barrier 1) required an independent commercial structure, we faced an immediate credibility challenge. Scientific infrastructure is typically trusted only when embedded within universities, while commercial vendors are often assumed to be “black boxes” that obscure method-critical details to protect IP, a practice that has contributed significantly to the “reproducibility crisis” in the life sciences [23, 24]. We addressed this by adopting a “glass box” operating philosophy, inverting the traditional vendor model to prioritize radical transparency. For hardware, we anchored each apparatus to specific literature citations to allow inspection of provenance (Activity 1). For software, we engineered tools to export directly to open standards like .NWB rather than proprietary format. This approach counters the trend of “data silos” that often hinder collaborative research [25]. By making designs, definitions, and outputs inspectable and stable across versions, we bridged the trust gap, enabled broad adoption across independent laboratories, and created the trust required to collaborate with scientists on replication-forward innovations, including the replication-focused technology transfer pathway (Activity 4).

Barrier 3: Published Methods Are Often Not Usable for Re-Execution Without Additional Reconstruction Level:

Community and Institutional Level: Replication-critical details are often held as tacit knowledge within laboratories rather than formalized assets [26]. Apparatus dimensions, procedural timing rules, analysis thresholds, and edge cases are commonly distributed across informal notes, design files, or staff experience and are rarely consolidated at the time of publication [23]. As a result, methods are shared late or informally and rarely in a form that supports faithful reuse by independent laboratories, forcing reconstruction that introduces avoidable variation. We addressed this barrier by working directly with investigators, both inside and outside formal Technology Transfer Offices, and acting as a trusted external partner that captures method details in parallel with ongoing research rather than requiring post hoc reconstruction. For investigators, this approach reduced documentation burden while increasing visibility, citation, and downstream reuse of their methods [27]. For institutions using formal agreements, research tool technology transfer provided a structured path for dissemination aligned with routine research use [28]. Together, these incentives shifted method sharing earlier in the research lifecycle and made complete specification a net benefit rather than a cost.

Barrier 4: IP Models Misaligned with Research Tools Level:

Institutional Level: University technology transfer offices are typically structured around exclusive licensing models designed for therapeutic products, which often block dissemination when applied to experimental tools [29]. As a result, laboratories outside the originating institution may lack access to apparatus designs or face prohibitive costs, leaving methods confined to a single site and replicated only through improvised reconstructions. We addressed this by adopting a licensing approach tailored to research tools, so that methods can function as shared replication standards rather than single-lab inventions.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

Replication depends on method components that remain consistent across sites. Below we summarize when each replication capability 1–5 was established and integrated into the operating ecosystem.

- Phase 1 (2013–present): Activity 1, apparatus replication. Established literature-linked apparatus designs with build-ready specifications and release checks, ensuring that laboratories citing the same reference receive physically equivalent equipment over time.

- Phase 2 (2016–present): Activity 2, procedure replication. Implemented automated protocol execution with validated timing and hardware integration, enabling consistent task execution without reliance on operator judgment.
- Phase 3 (2018–present): Activity 4, replication-focused technology transfer. Built a pathway to convert academic prototypes into standardized tools deployable beyond the originating laboratory under research-tool licenses.
- Phase 4 (2023–present): Activity 3, analytic replication. Deployed ConductVision to generate consistently defined behavioral outputs with transparent, standardized exports across diverse experimental environments.
- Phase 5 (2025–present): Activity 5, dissemination and reuse. Launched toolkit-based publishing with versioning and DOI-backed method packages, allowing complete methods and datasets to be shared as reusable research objects.

5.2 Required Resources

Implementation draws on four capability areas maintained as standing infrastructure: (1) controlled apparatus design and build standards; (2) automated, hardware-integrated procedure execution; (3) video capture and compute capacity to run ConductVision, with outputs in open formats including the .NWB standard; and (4) dissemination infrastructure that packages methods and datasets with persistent identifiers (DOIs) and version tracking for reuse.

6. COLLABORATION

6.1 Necessary partnerships.

Because this strategy operates as shared research infrastructure, repeatability across independent laboratories required four partnership types: (1) Deploying laboratories and behavioral cores, which implement standardized tools under routine facility conditions and provide feedback that improves documentation, default configurations, and procedure files; (2) Academic inventors and Technology Transfer Offices, which provide design intent and dissemination rights and enable tech transfer agreements that convert single-lab prototypes into broadly usable standardized tools; (3) Supply chain and component partners, which ensure reliable sourcing of materials and components with documented specifications and substitutions when needed; and (4) Dissemination partners (journals, societies, repositories), which support standards alignment and persistent identifiers so methods and assets remain citable, reusable, and accessible beyond ConductScience. Leveraging existing partnerships. Our network of deploying laboratories identified practical edge cases during routine experiments; this feedback informed clearer documentation, standardized defaults, and more explicit procedure files to reduce variation across facilities.

University partnerships enabled selected investigator-developed methods to move from single-lab prototypes into standardized tools under new innovative contracts. Building new partnerships. As the infrastructure matured, we formed co-development partnerships with advanced users to define new paradigms as complete, standardized method packages from inception. We also expanded collaborations with informatics and standards groups to align data descriptions and identifiers with community norms, improving portability and reuse in external repositories and analysis workflows.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Sustainability

This replication strategy is sustainable because it operates as standing research infrastructure rather than a time-limited project. Replication-critical assets, including citation-linked CAD and build specifications, versioned procedure files, analytic code and output schemas, and structured documentation, are maintained and updated continuously rather than created for a single study and left static. Methods are published and accessed via the web, while authoritative versions are preserved as reusable research objects: complete method packages, including protocols, specifications, analytic definitions, and datasets, are issued persistent identifiers (DOIs) with

explicit version tracking. Long-term accessibility is ensured through third-party digital preservation [32-34], providing an independent archival layer beyond ConductScience.org. Sustainability is further supported by a revenue-backed operating model tied to routine use of standardized apparatus, software, and method packages, enabling ongoing engineering, quality verification, user support, and maintenance without reliance on short-term pilot funding.

7.2 Transferability and Generalizability

The five replication activities are designed to be structurally generalizable rather than specific to any single assay, organism, or organization. They preserve complete methods across sites by aligning five elements: apparatus, procedures, analysis outputs, pathways for moving new methods beyond originating laboratories, and a published record that preserves methods for reuse. The model is uncommon not because it is difficult to implement, but because existing funding and regulatory structures rarely support it. Institutions can adopt these components individually based on role and constraints. Importantly, this model has already been applied beyond traditional rodent maze assays. The same replication framework has been used to standardize non-rodent behavioral systems, digitally controlled and virtualized task environments, and custom microfabricated research instruments. In each case, the method was captured as a deployable package combining physical specifications, procedure logic, and defined analysis outputs, rather than narrative description alone.

Together, these examples demonstrate that the replication infrastructure is not organism-, assay-, or organization-specific and could be implemented by academic cores, consortia, or individual laboratories with appropriate funding and regulatory alignment. Our commercial implementation represents one viable solution under current constraints, not a unique or exclusive pathway for long-term, transferable replication infrastructure.

8. REFERENCES

This submission used generative AI (ChatGPT, OpenAI) for ideation and writing assistance. All content was reviewed and edited by the authors for accuracy and completeness [40].

1. Sharma S, Rakoczy S, Brown-Borg H. Assessment of spatial memory in mice. *Life Sciences*. 2010.
2. Walf AA, Frye CA. The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. *Nat Protoc*. 2007.
3. Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. *Science*. 1999;284(5420):1670-1672.
4. Richter SH, et al. Environmental enrichment effects on behavior are strongly laboratory dependent. *PLoS One*. 2011.
5. Mandillo S, et al. Reliability, robustness, and reproducibility in mouse behavioral phenotyping. *Physiol Genomics*. 2008.
6. Wahlsten D, et al. Different data from different labs: Lessons from studies of gene–environment interaction. *J Neurobiol*. 2006;66(3):283–311.
7. Rübél O, Tritt A, Ly R, et al. The Neurodata Without Borders ecosystem for neurophysiology data science. *eLife*. 2022;11:e78362.
8. Valdivia WD. The stakes in Bayh-Dole: public values beyond the pace of innovation. *Minerva*. 2011;49(1):25–46.
9. Farber GK, Weiss L. Core facilities: optimizing the return on investment. *Sci Transl Med*. 2011;3(95):95cm21.
10. National Academies of Sciences, Engineering, and Medicine (NASEM). *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press; 2019.
11. Weekman EM, Rogers CB, Sudduth TL, Wilcock DM. *Alzheimers Dement*. 2025;20(Suppl 1):e089306.
12. Weekman EM, Rogers CB, Sudduth TL, Wilcock DM. *J Neuroinflammation*. 2025;22:23.
13. Choi CQ. The mazes with minds of their own. *Nature (Toolbox)*. 2018 Feb 27;554.
14. Bromer C. A History of the Morris Water Maze, From Inception to Industry Standard. *Yahoo News*. 2017.
15. Cottier C. How to Build A Better Mouse Maze. *Discover Magazine*.
16. Blakemore E. Meet the folks designing the future of mouse mazes. *Popular Science*. 2018 Jul 16.
17. Harvard Innovation Labs. Spring 2020 Venture Program Cohort: ConductScience. Boston (MA): Harvard University; 2020.

18. National Institutes of Health (NIH). Winners of the Data Sharing Index (S-Index) Challenge (Phase 1). Bethesda (MD): NIH Office of Data Science Strategy; 2025.
19. National Science Foundation (NSF). I-Corps Northeast Hub: Market Validation of Dynamic Publishing Infrastructure. Princeton (NJ): NSF I-Corps Northeast Hub; 2024.
20. Alberts B, Kirschner MW, Tilghman S, Varmus H. Rescuing US biomedical research from its systemic flaws. *Proc Natl Acad Sci U S A*. 2014;111(16):5773-5777.
21. National Institutes of Health. Principles and Guidelines for Recipients of NIH Research Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources. *Fed Regist*. 1999;64(246):72090-72096.
22. Electronic Code of Federal Regulations. Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards. 2 CFR § 200.468 (Specialized service facilities).
23. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016;533(7604):452-454.
24. Mirowski P. The future(s) of open science. *Social Studies of Science*. 2018;48(2):171–203.
25. Teeters JL, et al. Neurodata Without Borders: Creating a Common Data Format for Neurophysiology. *Neuron*. 2015;87(5):1000–1012.
26. Collins HM. Tacit and Explicit Knowledge. Chicago (IL): University of Chicago Press; 2010.
27. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. *PeerJ*. 2013;1:e175.
28. Nosek BA, et al. Promoting an open research culture. *Science*. 2015;348(6242):1422-1425.
29. Heller MA, Eisenberg RS. Can Patents Deter Innovation? The Anticommons in Biomedical Research. *Science*. 1998;280(5364):698-701.
30. Mimura C. Nuance Management: The Role of Institutional Culture in Tech Transfer. *J Assoc Univ Technol Manag*. 2010;22(1).
31. Contreras JL, Reichman JH. Sharing by Design: Data and Biological Resource Centers as Nodes of Collaborative Knowledge Networks. *Int J Commons*. 2015;9(1).
32. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3:160018.
33. Kirchhoff AJ. Challenges and opportunities in the evolving digital preservation landscape: reflections from Portico. *Insights*. 2016;29(3):244–251.
34. OECD. Business models for sustainable research data repositories. OECD Science, Technology and Industry Policy Papers; 2017.
35. Lemley MA. Are Universities Patent Trolls? *Fordham Intellect Prop Media Ent LJ*. 2008;18(3):611.
36. NASEM. Open Science by Design: Realizing a Vision for 21st Century Research. Washington (DC): National Academies Press; 2018.
37. Schmid B, et al. Automated T-maze assays for *Drosophila*: Reproducibility across labs using standardized hardware. *Sci Rep*. 2024;14:8921.
38. Nguyen V, et al. Virtualizing the Radial Arm Maze: Cross-site validation of digital spatial navigation. *J Neurosci Methods*. 2025;410:110254.
39. Johnson T, et al. Standardizing the tumor-microenvironment-on-a-chip: A pathway for reproducible oncology tools. *Lab Chip*. 2022;22(14):2680–2695.
40. OpenAI. ChatGPT (GPT-5.2 Thinking) [Internet]. San Francisco (CA): OpenAI; [cited 2025 Dec 18]. Available from: <https://chat.openai.com/>.

INDIVIDUAL WINNER

Jagadeesh Ramasamy



Replicated and extended his own landmark discovery on mitochondrial persistence in red blood cells, advancing understanding of inflammation and disease.

Dr. Jagadeesh Ramasamy, BSMS, MSc, PhD, is Director of Bioscience Research–Hematology/Oncology and Associate Professor (Research) at the University of Illinois, and a translational scientist recognized for advancing rigorous, reproducible discovery in red blood cell biology. He first challenged long-standing hematology dogma through his seminal demonstration of mitochondrial persistence in reticulocytes and mature red blood cells (Experimental Hematology, 2017), redefining the understanding of erythrocyte maturation. Building on this foundational discovery, his work established that retained erythrocyte mitochondria drive the pathogenic release of excessive mitochondrial reactive oxygen species and are accompanied by dysregulation of selective autophagy proteins, including GABARAPs, α -synuclein, and cathepsin B (Experimental Hematology, 2025). These mechanisms have broad implications for hemolysis, chronic inflammation, and immune-mediated diseases pathogenesis. His current research focuses on developing natural peptide-based therapeutics to restore mitochondrial homeostasis across diseases.

1. REPLICATION STRATEGY

1.1 Strategy Description

We implemented a multisite, pre-registered, open science replication workflow built on platforms, SOPs (Standard Operating Procedures), and analytic pipelines established under prior NIH awards (R01 HL136622; P01 HL136622; R21 AT012304 01). Those awards enabled robust live-cell flow cytometry (TMRM/TMRE with CCCP/FCCP controls), strict lineage-exclusion gating (CD45⁺/CD41⁺/CD61⁺/CD71⁺/TO⁺), orthogonal confirmation (immunofluorescence, TEM ultrastructure), and standardized respirometry protocols with inhibitor ladders. We harmonized assays across centers using clinical research calibration, shared compensation templates, and dye lots. Three integrated streams were replicated:

- Phenotype replication: Elevated mitochondrial signal and ROS in mature SCD RBCs vs controls by TMRM/TMRE flow cytometry; IF/TEM; O₂ consumption respirometry assays harmonized across sites with clinical trials (Jagadeeswaran et al., 2017; Gallivan et al., 2023; Ramasamy et al., 2025).
- Mechanism replication: Autophagy/mitophagy signatures (GABARAPs, CTSB/ α synuclein linking retention → oxidative stress → clinical features (Ramasamy et al., 2025).
- Clinical correlation replication: Associations of retention/ROS with hemolysis indices (LDH, bilirubin, reticulocytosis), retinopathy (OCT A), nephropathy (eGFR/albuminuria), and pain PROs mapped to whole body endpoints described in our R34 concept.

1.2 Outcome Evaluation Methods

- **Primary:** Cross-site agreement (ICC ≥ 0.80) and effect size equivalence bounds via mixed effects models for mitochondrial signal (case vs control).
- **Secondary:** Prespecified correlations between retention/ROS and hemolysis/organ outcomes, stratified by genotype (HbSS/HbSC/S β^0 /*) and therapy (e.g., hydroxyurea).

1.3 Innovative Activities

- Clinical research calibration using shared reference blood, dye lots, and compensation templates.
- Artifact firewall SOPs (mandatory lineage gating, platelet depletion, live dye depolarization controls) enforced as pass/fail QC.
- Open SOPs/templates/code/raw data to support reuse and secondary validation.

1.4 Working Definition of Replication

Independent, protocol-controlled re-measurement of a reported biological phenomenon across multiple sites/cohorts, using harmonized primary assays and orthogonal confirmations, with prespecified statistical thresholds for agreement and equivalence, and transparent sharing of protocols, materials, code, and raw data.

Category Identification: Translational, Preclinical, and Clinical.

Abbreviations: SOP, Standard Operating Procedure; QC, Quality Control; SCD, Sickle Cell Disease; RBC, Red Blood Cell; ROS, Reactive Oxygen Species; TMRM, Tetramethylrhodamine Methyl Ester; TMRE, Tetramethylrhodamine Ethyl Ester; CCCP, Carbonyl Cyanide m-Chlorophenyl Hydrazone; FCCP, Carbonyl Cyanide-p-Trifluoromethoxyphenylhydrazone; IF, Immunofluorescence; TEM, Transmission Electron Microscopy; GABARAP, Gamma-Aminobutyric Acid Receptor-Associated Protein; CTSB, Cathepsin B; LDH, Lactate Dehydrogenase; OCT-A, Optical Coherence Tomography Angiography; eGFR, Estimated Glomerular Filtration Rate; PRO, Patient-Reported Outcome; ICC, Intraclass Correlation Coefficient; HbSS, Homozygous Sickle Cell Disease; HbSC, Hemoglobin SC Disease; S β^0 , Sickle Beta-Zero Thalassemia; S β^+ , Sickle Beta-Plus Thalassemia.

2. ENVIRONMENT

2.1 Activity context

Replication activities occurred in academic medical centers with clinical SCD programs and access to flow cytometry, EM, respirometry cores, and biostatistics/data science groups. SOPs and pipelines refined under NIH R01 HL136622, P01 HL136622, and R21 AT012304 01 served as the backbone for replication-grade QC, dashboards, and governance (Jagadeeswaran et al., 2017; Gallivan et al., 2023; Ramasamy et al., 2025).

2.2 Location, timing, team details

Sites: U.S. centers (hematology clinics + core facilities).

Timeline: 24 months total. 0–3 pre-registration/SOP harmonization; 4–6 ring trials/calibration; 7–18 Enrollment/assays; 19–24 analysis/dissemination.

Team and roles:

- PI (Jagadeesh Ramasamy): Strategy; assay integration; statistical plan; cross-site QC oversight.
- Clinical leadership (Angela Rivers; Robert Molokie; Lewis Hsu): Cohort recruitment; phenotyping; PROs; OCT A/nephrology endpoints.
- Program collaboration (Donald Lavelle; P01 HL136622): Method synergy and bench to clinic pathways.
- Core leads: Flow cytometry, EM, Respirometry.
- Data/Stats: Mixed effects modeling; ICC/equivalence; repositories.
- Coordinators: IRB, logistics (cold chain), SOP compliance, QC documentation. Funding support (completed NIH-funded projects)
- NIH (NHLBI) R01 HL136622 – UIC Site PI: Jagadeesh Ramasamy; Overall PI: Angela Rivers. Sub award: 05/15/2020–04/30/2024. Title: The Role of Erythrocyte Mitochondrial Retention in Sickle Cell Disease.
- NIH (NHLBI) P01 HL136622 – Co I: Jagadeesh Ramasamy; PI: Donald Lavelle. 05/01/2020–04/30/2024. Title: Advancing Fetal Hemoglobin to the Clinical Using Highly Predictive In Vivo Strategies – Targeting Enzymatic Regulation of Fetal Hemoglobin Repression.
- NIH (NINDS) R21 AT012304 01 – PI: Jagadeesh Ramasamy. 10/01/2022–09/30/2024. Title: Erythrocyte Autophagy Proteins as Potential Non Opioid Novel Targets for Pain in Sickle Cell Disease. Active/recent non-NIH awards
 - Discovery Partner Institute-CRI (2025) – Cannabinoid receptors in SCD.
 - University of Chicago Sub Award (2024–2025) – Tet2 mutant in SCD. NIH grant proposals under review/submitted

- R01 EY038718 – AI driven Innovations for Early Detection and Management of Sickle Cell Retinopathy Integrating Blood–Retina Cross Talk, Omega-3 Fatty Acids, and Mitochondrial Health.
- R01 HL185478 – Therapeutic Potential of Minor Cannabinoids to Modulate Erythrocyte Mitochondrial Bioenergetics and ROS in SCD.
- R01 AT013379 – GABARAPs and HSP90 as Potential Targets to Prevent Chronic Pain and Nephropathy in SCD.
- R34 – Yogic Breathing as a Mind–Body Modality for Mitochondrial and Autonomic Regulation in SCD.

3. IMPACT

3.1 Pre Replication

Study findings may have been influenced by inconsistent protocols, dye variability, contamination from other cell types, and differences in OCR measurement methods.

3.2 Post Replication

Harmonized live dye SOPs, strict lineage gates, orthogonal confirmations, and calibrated O₂ consumption workflows achieved ICC ≥ 0.80 across sites and stable effect sizes by genotype/therapy (Jagadeeswaran et al., 2017; Gallivan et al., 2023; Ramasamy et al., 2025).

Long term impact: The validated, trial ready biomarker workflow now anchors aims in the pending R01 EY038718 (retina centric early detection with blood–retina cross talk), R01 HL185478 (minor cannabinoids and RBC bioenergetics/ROS), and R01 AT013379 (autophagy targets for chronic pain/nephropathy) and provides operational scaffolding for the R34 whole body program (retina, kidney, cardiopulmonary, pain/PROs).

Research rigor improvements: Pre-registration, prespecified analyses, ring trials, artifact firewalls, and open data/code materially improve rigor and reproducibility, increasing stakeholder trust and translatability to clinical studies and future mechanistic trials.

Dissemination and adoption: Public SOPs/templates/code/raw data; partner labs adopting clinical trial calibration and gating firewalls; methods referenced in new protocol submissions, publication, and pilot studies aligned to the pending R01 and R34 programs.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

Several barriers were identified and systematically addressed during the replication effort. Lineage and platelet contamination were minimized through strict CD71/TO exclusion, CD41/CD61 depletion, and standardized pass/fail quality control criteria with reprocessing when required. Dye-related artifacts were reduced by adopting TMRM/TMRE-based assays and incorporating mandatory CCCP/FCCP depolarization controls. Respirometry variability was addressed through standardized oxygen consumption rate (OCR) protocols, inhibitor ladder analyses, and red blood cell-specific normalization methods. Instrument and site-to-site differences were mitigated through centralized quality-control monitoring. Continuity of the research program was maintained through institutional bridge support and alignment with NIH awards, ensuring retention of personnel and sustained program momentum.

5. TIMELINE & BUDGET

5.1 Implementation Timeline (24 months)

- 0–3 months: Pre-registration; SOP/QC finalization; IRB approvals; data governance.
- 4–6 months: Clinical trials; cross site calibration; pilot runs.
- 7–18 months: Enrollment; phenotyping; mechanism module; organ outcomes; interim analyses.
- 19–24 months: Final analyses; public release; integration into R01/R34 workplans. Associated costs (summary) Personnel (PI/Co Is, coordinators, core staff, data scientists); assays (dyes/antibodies, EM imaging, OCR, qPCR reagents); calibration materials; data systems (secure storage, repositories, compute).

5.2 Required resources

Access to SCD/control cohorts; flow/EM/OCR cores; biostatistics/data; regulatory/logistics; community engagement established under R01 HL136622, P01 HL136622, R21 AT012304 01, and rolled forward in the pending R01 and R34.

6. COLLABORATION

6.1 Necessary Partnerships

Clinical hematology (enrollment, phenotyping, patient-reported outcomes (PROs)), core facilities (flow cytometry, EM, respirometry), subspecialties (ophthalmology/retinal, nephrology, cardiopulmonary, pain/behavioral health) for the R34 whole body endpoints, biostatistics/data, and patient advisory groups. Leveraging existing partnerships, networks formed during R01 HL136622 and P01 HL136622 expedited IRBs, core prioritization, and SOP adoption; prior co-authorships facilitated onboarding of artifact firewall and clinical research practices.

6.2 Building New Partnerships

The R34 whole-body approach added nephrology, retina, cardiology/pulmonology, pain/behavioral collaborators; community groups aligned outcomes and PROs with patient priorities. Non-NIH awards (DPI-CRI 2025; University of Chicago subaward 2024–2025) strengthened cannabinoid receptors and Tet2 mutant exploratory workstreams.

7. SUSTAINABILITY / TRANSFERABILITY

SOPs, in clinical research, and orthogonal verification published work are portable to other institutions (Moriconi et al., 2022; Tumburu et al., 2021) and disease areas such as SLE (Caielli et al., 2021), hereditary spherocytosis (Principe et al., 2024) generalize to pragmatic clinical research workflows and registry-linked longitudinal monitoring.

7.1 Current Sustainability Plans

Plans include open-access repositories and standardized training materials to facilitate knowledge transfer across sites. Long-term program continuity will be maintained through ongoing and planned funding mechanisms, including NIH grant applications, institutional support, and continued multi-site collaborations.

7.2 Scaling and Application

Expansion to multi-regional/multi-continental cohorts using remote ring trials and virtual QC; embedding the biomarker in trials of mitochondria/mitophagy targeted therapies; registry linkage for whole body endpoints.

7.3 Future Integration Plans

Training/credentialing; digital twin simulations for power/variance planning; institutional policy templates for replication and SOP adoption; annual consortium workshops; patient-partnered dissemination (lay abstracts, community forums); multi-omics add-ons under R01/R34 to deepen mechanism translation (Ramasamy et al., 2025). Funding and proposal identifiers cited • R01 HL136622 (NHLBI) – UIC Site PI: Jagadeesh Ramasamy; Overall PI: Angela Rivers. Sub award 05/15/2020–04/30/2024. • P01 HL136622 (NHLBI) – Co I: Jagadeesh Ramasamy; PI: Donald Lavelle. 05/01/2020–04/30/2024. • R21 AT012304 01 (NINDS) – PI: Jagadeesh Ramasamy. 10/01/2022–09/30/2024. • R01 EY038718 – Submitted;06/2025. • R01 HL185478 – Submitted;07/2025. • R01 AT013379 – Submitted;07/2025. • R34 – Submitted; 10/2025. • Discovery Partner Institute-CRI (2025) – Cannabinoid receptors in SCD (PI: Jagadeesh Ramasamy). • University of Chicago Sub Award (2024–2025) – Tet2 mutant in SCD mouse model (PI: Jagadeesh Ramasamy).

8. REFERENCES

1. Caielli, S., Cardenas, J., de Jesus, A. A., Baisch, J., Walters, L., Blanck, J. P., Balasubramanian, P., Stagnar, C., Ohouo, M., Hong, S., Nassi, L., Stewart, K., Fuller, J., Gu, J., Banchereau, J. F., Wright, T., Goldbach-Mansky, R., & Pascual, V. (2021). Erythroid mitochondrial retention triggers myeloid-dependent type I interferon in human SLE. *Cell*, 184(17), 4464-4479 e4419. <https://doi.org/10.1016/j.cell.2021.07.021>
2. Gallivan, A., Alejandro, M., Kanu, A., Zekaryas, N., Horneman, H., Hong, L. K., Vinchinsky, E., Lavelle, D., Diamond, A. M., Molokie, R. E., Ramasamy, J., & Rivers, A. (2023). Reticulocyte mitochondrial retention increases reactive oxygen species and oxygen consumption in mouse models of sickle cell disease and phlebotomy-induced anemia. *Exp Hematol*, 122, 55-62. <https://doi.org/10.1016/j.exphem.2023.02.005>
3. Jagadeeswaran, R., Vazquez, B. A., Thiruppathi, M., Ganesh, B. B., Ibanez, V., Cui, S., Engel, J. D., Diamond, A. M., Molokie, R. E., DeSimone, J., Lavelle, D., & Rivers, A. (2017). Pharmacological inhibition of LSD1 and mTOR reduces mitochondrial retention and associated ROS levels in the red blood cells of sickle cell disease. *Exp Hematol*, 50, 46-52. <https://doi.org/10.1016/j.exphem.2017.02.003>
4. Moriconi, C., Dzieciatkowska, M., Roy, M., D'Alessandro, A., Roingeard, P., Lee, J. Y., Gibb, D. R., Tredicine, M., McGill, M. A., Qiu, A., La Carpi, F., Francis, R. O., Hod, E. A., Thomas, T., Picard, M., Akpan, I. J., Luckey, C. J., Zimring, J. C., Spitalnik, S. L., & Hudson, K. E. (2022). Retention of functional mitochondria in mature red blood cells from patients with sickle cell disease. *Br J Haematol*, 198(3), 574-586. <https://doi.org/10.1111/bjh.18287>
5. Principe, D. R., Reilly, P., Dhavamani, S., Rivers, A., Molokie, R., Hsu, L. L., & Ramasamy, J. (2024). Hereditary Spherocytosis with Mitochondrial Retention, Increased Oxidative Stress, and Alterations to Bioactive Membrane Lipids. *J Pediatr Hematol Oncol*, 46(6), e457-e462. <https://doi.org/10.1097/MPH.0000000000002901>
6. Ramasamy, J., Kannan, P. K. P., Dhavamani, S., Palanimuthu, S., Molokie, R., & Rivers, A. (2025). Mitochondrial Retention and Autophagy Dysregulation Drive Oxidative Stress in Sickle Cell Disease Erythrocytes. *Exp Hematol*, 105269. <https://doi.org/10.1016/j.exphem.2025.105269>
7. Tumburu, L., Ghosh-Choudhary, S., Seifuddin, F. T., Barbu, E. A., Yang, S., Ahmad, M. M., Wilkins, L. H. W., Tunc, I., Sivakumar, I., Nichols, J. S., Dagur, P. K., Yang, S., Almeida, L. E. F., Quezado, Z. M. N., Combs, C. A., Lindberg, E., Bleck, C. K. E., Zhu, J., Shet, A. S., . . . Thein, S. L. (2021). Circulating mitochondrial DNA is a proinflammatory DAMP in sickle cell disease. *Blood*, 137(22), 3116-3126. <https://doi.org/10.1182/blood.202009063>

INDIVIDUAL WINNER

John Worthley



Tests whether ACE inhibitors developed at sea level perform consistently at high altitude, using high-elevation "natural laboratories" to study hypoxia's effect on drug function.

John Worthley is a researcher investigating a critical gap in cardiovascular pharmacology: the "environmental neutrality" of drug action. His work challenges the assumption that ACE inhibitors, developed in sea-level laboratories, maintain consistent efficacy at high altitudes. By utilizing "natural laboratories" like the Donner Lake Basin, John studies how physiological adaptations to hypoxia—such as the stabilization of Hypoxia-Inducible Factors (HIFs)—alter how these medications function in the body.

1. REPLICATION STRATEGY

1.1 Strategy Description

1.1.1 The "Ruggedized Protocol" Framework

Our strategy integrates Environmental Contextualization (EC)—a method borrowed from alpine ecology—into standard biomedical workflows. Rather than attempting to eliminate "noise" through impossible standardization, we treat environmental variables as vital metadata.

1.2 Integration into Practice

We implemented a "Site Characterization Report" for every experiment. Before a replication attempt, the team must document "microclimate" variables of the lab (e.g., exact CO2 fluctuations in incubators, ambient light cycles, and even the vibration frequency of benchtop equipment).

1.3 Methods for Evaluating Outcomes

We use Sensitivity Mapping. Instead of a binary "pass/fail" for replication, we evaluate the range of environmental conditions under which the finding remains stable. If a result only holds at exactly 22°C but fails at 24°C, it is flagged as "High Context Sensitivity."

1.4 Innovative Activities

"Stress-Testing" Replications: Borrowing from high-altitude physiology, we intentionally introduce "controlled noise" (e.g., varying reagent temperatures or incubation times within a 5% margin) to ensure the protocol is "field-ready" and robust enough for broader populations.

1.5 Automated Metadata Logging

Using IoT sensors to capture lab "weather" data, ensuring the "context" is recorded without increasing researcher burden.

1.6 Working Definition

"Robust Replication" In this strategy, we define replication as the ability of a scientific finding to maintain its predictive power across varying experimental "microclimates." We do not seek identical results in identical

conditions (which is often a measure of precision, not truth); rather, we seek the survival of the core effect when subjected to the inherent variability of different "natural laboratories" (different labs, teams, or equipment).

1.7 Category Identification

While this strategy is highly adaptable, it most strongly applies to:

Basic Research: Where fundamental biological mechanisms are often sensitive to subtle lab variations.

Pre-clinical: Where "ruggedizing" protocols is essential before moving to the high-stakes environment of human trials.

Translational: Ensuring that discoveries move from the "controlled slope" of the lab to the "complex terrain" of clinical application.

2. ENVIRONMENT

2.1 Activity Context

The replication activities were conducted within the Donner Lake Basin, a high-altitude "natural laboratory" characterized by extreme topographical complexity and rapid environmental shifts. The strategy focused on bridging the gap between field-observed environmental "noise" and controlled research settings. By treating environmental variables—such as atmospheric contaminants, hydrological substrate stability, and microclimatic shifts—as essential metadata rather than irrelevant interference, the team developed a framework for "Context-Aware" research. This approach ensures that biological or chemical findings are not just precise in a vacuum, but robust enough to be replicated across diverse and variable real-world environments.

2.2 Location, Timing, and Team Details

Location: Donner Lake, California (High Sierra Region), specifically utilizing Schallenberger Ridge, China Cove, and Coldstream Canyon.

Dates and Duration: July 13, 2025 – December 16, 2025.

Team Expertise, Roles, and Responsibilities:

John Worthley | Coldstream Canyon Hydrology & Soil Sciences Researcher

Expertise: Biogeochemical Modeling & Data Provenance.

Role/Responsibilities: Developed "Physical-to-Digital Twin" protocols. He was responsible for creating a digital mirror of the physical field/lab conditions to ensure that the chemical signatures detected were reproducible regardless of seasonal variability. His work ensured that the context of the substrate (the "soil" of the experiment) was meticulously documented to maintain data provenance.

Samvel Hakobyan | Schallenberger Ridge Field Station Researcher

Expertise: Atmospheric Sciences & Trace Organic Contaminants.

Role/Responsibilities: Led the automated environmental logging strategy. He was responsible for the collection and analysis of trace organic contaminants and ensuring that the atmospheric data was clean, replicable, and formatted for public dissemination. His work focused on removing human error from the reporting of atmospheric variables.

Volodymyr Testardi | China Cove Beach Researcher

Expertise: Environmental Metadata Integration.

Role/Responsibilities: Specialized in protocol integration. He was responsible for developing the workflow to integrate "environmental metadata" (the "weather" of the lab or clinic) directly into digital health records and research protocols, ensuring environmental context follows the data into clinical applications.

2.3 Funding Support

This effort was supported by a combination of federal, organizational, and institutional funds:

Organization Funding: Supported by Mongata LLC grant under the Environmental Research (ER) program (awarded July 2025). This supported the deployment of high-precision sensor arrays.

3. IMPACT

3.1 Pre-implementation

Prior to implementing this strategy, environmental "noise" was often dismissed as experimental error, leading to a high variance in results when protocols were moved from the lab to the field. By implementing the Context-Aware Framework, the team achieved measurable improvements in data consistency:

- **Metric Comparison:** Pre-implementation, the coefficient of variation (CV) for trace contaminant samples across different topographical sites was 28%. Post-implementation, by utilizing the "Physical-to-Digital Twin" protocols and automated logging, the CV was reduced to 9%.
- **Predictive Accuracy:** The ability to predict chemical stability across varying microclimates at Donner Lake improved by 40%, as researchers could now adjust for real-time atmospheric and substrate variables recorded in the metadata.

3.2 Long-term Impact

The long-term impact is the creation of a "Ruggedized Research Standard" that moves beyond the fragile "controlled lab" mindset. By establishing that high-altitude variables (UV, Phosphorus Dioxide, temperature shifts) are predictable metadata points rather than random interference, we have created a foundation for biomedical research that is inherently "field-ready." This shifts the focus from achieving a single successful result to understanding the stability boundary of a discovery—ensuring it holds true across different seasons, altitudes, and diverse clinical environments.

3.3 Research Rigor Improvements

These activities fundamentally increased research rigor by eliminating "Hidden Variable Bias."

Data Provenance

John Worthley's substrate modeling ensured that the "memory" of the environment was captured, preventing future researchers from misinterpreting localized shifts as universal biological changes.

Transparency and Trust

By integrating automated logs (Samvel Hakobyan) and mapping them to clinical records (Volodymyr Testardi), the team moved from subjective observation to objective, sensor-based documentation. This transparency has significantly increased stakeholder trust, as the "context" of every finding is fully auditable and verifiable by outside parties.

3.4 Dissemination and Adoption

The framework developed at Donner Lake has already begun to see adoption beyond the immediate team:
Institutional Adoption: The California Air Resources Board (CARB) has adopted the automated logging strategy for their 2025 health-risk assessment pilot programs, citing the ease of public dissemination as a key factor.

4. RESPONSE TO BARRIERS

4.1 Primary Barriers

The primary challenge in bridging field ecology and biomedical replication was the "Documentation Burden." Researchers are often resistant to adding complex metadata layers that they perceive as secondary to their primary findings.

Barrier: Technical Incompatibility between Field and Clinical Data.

Problem: Field sensors from Schallenberger Ridge produced raw atmospheric data that was incompatible with the structured formats of digital health records used at China Cove.

Solution: Volodymyr Testardi developed an API-based middleware that automatically tagged and "translated" environmental logs into clinical-ready metadata. This removed the need for manual data entry and ensured that the "context" was automatically attached to the research protocol.

Barrier: Seasonal Volatility and Equipment Survival.

Problem: The rapid shift from summer to winter at Donner Lake (July to December 2025) threatened the reliability of the "Physical-to-Digital Twin" sensors. High-altitude UV and moisture levels caused sensor drift.

Solution: John Worthley implemented a "Redundant Array" strategy, using low-cost secondary sensors to cross-validate primary data. Samvel Hakobyan then applied a "Self-Correcting Algorithm" to the automated logs, which adjusted for sensor drift in real-time based on known atmospheric baselines.

4.2 Organizational Levels

The team addressed barriers across three distinct levels of the research ecosystem:

Team Level (Interdisciplinary Communication):

Barrier: The team spoke different "scientific languages"—biomedical vs. hydrological vs. atmospheric.

Resolution: We established a Unified Data Dictionary. By standardizing terms like "Substrate Stability" and "Contextual Noise," the team ensured that a hydrologist's soil observation could be accurately interpreted by a clinical researcher as a "sample environment variable."

4.3 Location/Site Level (Physical Access & Logistics)

Barrier: Implementing high-precision replication activities in remote locations like Coldstream Canyon presented significant logistical hurdles, particularly during early winter storms in late 2025.

Resolution: The team leveraged the Truckee River Watershed Council's existing infrastructure to create "Ruggedized Data Hubs." These hubs allowed for remote data transmission, meaning researchers didn't have to physically access the site during hazardous conditions to maintain the replication logging.

4.4 Community Level (Skepticism of Replication Value)

Barrier: Local stakeholders were initially skeptical that "replication" was a valuable use of funds compared to "new discovery."

5. TIMELINE & BUDGET

5.1 Implementation Timeline

The implementation was executed in four distinct phases to transition from high-altitude data collection to integrated biomedical replication standards.

- Phase 1: Baseline & Sensor Deployment (July 13 – August 15, 2025): Installation of automated environmental logging arrays at Schallenberger Ridge and calibration of sensors for soil-water interface dynamics in Coldstream Canyon.
- Phase 2: Stress-Testing & Data Twin Mapping (August 16 – October 15, 2025): Conducted initial replication cycles. John Worthley developed the "Physical-to-Digital Twin" models while Samvel Hakobyan identified trace organic contaminant fluctuations.
- Phase 3: Clinical Integration & Protocol Mapping (October 16 – November 15, 2025): Volodymyr Testardi integrated the "lab weather" metadata into digital health record templates, mapping environmental stressors to biomedical assay variables.
- Phase 4: Validation & Dissemination (November 16 – December 16, 2025): Final analysis of pre- and post-replication metrics.

5.2 Associated Costs

The total expenditure to achieve these replication outcomes was \$19,802, which was fully supported by the federal and institutional grants detailed in Section 2. These costs included \$5,700 for IoT Hardware and Sensors, specifically high-precision UV, partial pressure of oxygen (pO₂), and substrate moisture sensors required for automated environmental logging. An additional \$3,800 was allocated for Data Middleware Development to create the custom APIs necessary for integrating field metadata into clinical records. Field Logistics accounted for \$2,302, covering transportation, remote solar power arrays, and the essential winterization of equipment for the high-altitude environment. Finally, \$8,000 was dedicated to Personnel and Stipends for the specialized labor required for biogeochemical modeling, data provenance, and trace contaminant analysis.

6. COLLABORATION

6.1 Necessary Partnerships

To bridge the gap between field ecology and biomedical rigor, it was necessary to partner with organizations that manage high-stakes "natural laboratories" and clinical data systems. Essential partners included the Truckee River Watershed Council (TRWC) for site access and infrastructure, and the California Air Resources Board (CARB) for regulatory alignment on air quality standards. In retrospect, a partnership with a specialized Biomedical Ethics Review Board should have been considered earlier to further streamline the integration of environmental metadata into sensitive patient-related digital health records, ensuring HIPAA compliance for "lab weather" logging from the outset.

6.2 Leveraging Existing Partnerships

We leveraged long-standing relationships with the UC Davis Tahoe Environmental Research Center (TERC) and the California Department of Parks & Recreation. TERC provided the historical baseline data necessary to calibrate our current sensors, while State Parks allowed for the deployment of semi-permanent sensor arrays at Schallenberger Ridge and China Cove. These existing ties provided the physical "proving ground" for the strategy, significantly reducing the lead time required for site preparation and permitting.

6.3 Building New Partnerships

New partnerships were forged to address the technical and community-facing aspects of the project:

- Technical Partnership: A new collaboration was established with the Truckee Tahoe Airport to prototype low-cost, ruggedized sensor housings. This partnership allowed us to move from expensive off-the-shelf equipment to custom-built, "field-ready" hardware.
- Clinical Partnership: We initiated a pilot program with the Tahoe Forest Health System to test the feasibility of mapping environmental "noise" (atmospheric data) into their research protocols.
- Community Partnership: By engaging with local environmental advocacy groups, we built a support network that values "research rigor" as a community asset, helping to disseminate the findings of our transparency showcases to a non-scientific audience.

7. SUSTAINABILITY / TRANSFERABILITY

The "Context-Aware" replication strategy is inherently sustainable because it relies on low-cost, automated data collection rather than labor-intensive manual reporting. By shifting the burden of environmental documentation from the researcher to IoT sensors, the strategy ensures high compliance over time. The framework is highly generalizable; while developed in the Donner Lake Basin, the principle of documenting "environmental noise" (temperature, pO₂, and atmospheric contaminants) is applicable to any setting where biological or chemical sensitivity is a factor—from high-altitude field stations to urban clinical laboratories.

7.1 Current Sustainability Plans

Current plans focus on the "Hardening" of the digital infrastructure. We are developing a permanent, open-source repository for the "Physical-to-Digital Twin" protocols, ensuring that the modeling framework remains accessible after the initial grant period. We have also secured a memorandum of understanding (MOU) with the Truckee River Watershed Council to maintain the remote data hubs, allowing for continuous environmental logging that future research teams can tap into as a shared baseline.

7.2 Scaling and Application

This strategy can be scaled by expanding the sensor network to include diverse "microclimates" beyond the Sierra Nevada. In a biomedical context, this looks like:

Multi-Site Clinical Trials: Applying the same automated metadata logging across different hospitals to identify why a drug may perform differently in high-altitude Truckee versus sea-level San Francisco.

Laboratory Standard Operating Procedures (SOPs): Scaling the "Site Characterization Report" as a mandatory requirement for any grant-funded research within an institution.

Low-Resource Settings: Because we utilized low-cost sensors prototyped at the Truckee Roundhouse, the strategy is accessible to smaller labs or international field sites with limited budgets.

7.3 Future Integration Plans

To deepen the culture of replication, our future roadmap includes:

Institutional Policy: Advocating for "Contextual Metadata" requirements in university Internal Review Board (IRB) and animal care (IACUC) protocols to ensure environmental variables are logged in pre-clinical studies.

Training Programs: Developing a "Ruggedized Research" certification for graduate students, focusing on the technical skills needed to build and maintain automated logging arrays.

Information Gathering: Creating a centralized "Replication Failure Database" where researchers can upload "contextual noise" data that led to non-replicable results, turning "failed" experiments into valuable data on biological stability boundaries.

Networking: Establishing a "Field-to-Lab" consortium that brings together ecologists and biomedical scientists to share best practices in managing complex topographical and environmental data.

INDIVIDUAL WINNER

Krissandra Johnson-Nealey



A 20-year clinical trial site operations veteran applying her frontline experience to improve data integrity and reproducibility in metabolic and cardiovascular research.

Krissandra Johnson-Nealey is a clinical research professional with over 20 years of experience in site operations, supporting trials across metabolic, cardiovascular, and chronic conditions. She works at Central Texas Clinical Research, where she oversees key aspects of study execution, including regulatory compliance, patient coordination, and data quality.

Her work is grounded in real-world experience, where she has seen how clinical research directly impacts patient outcomes beyond the trial setting. She is especially focused on improving data integrity and ensuring study results are accurate, consistent, and reflective of true patient experiences. Her approach emphasizes practical solutions that strengthen research reliability and support replication within everyday clinical operations.

Krissandra is currently pursuing her MBA and continues to grow into leadership roles. She remains committed to advancing research that is both credible and meaningful for the communities it serves.

1. REPLICATION STRATEGY

1.1 Strategy Description

Our replication strategy was integrated into routine clinical research operations across multiple obesity and cardiometabolic studies conducted at a single research site with participation in sponsor-led, multi-site trials. The strategy focused on systematic duplication of outcome verification, protocol adherence checks, and cross-validation of primary endpoints across patient cohorts receiving obesity pharmacotherapy, including GLP-1–based interventions.

The core activities included:

- **Independent Outcome Verification** All primary outcome measures (body weight, BMI, waist circumference, vital signs, and key metabolic labs) were independently re-verified by a second team member at every data collection visit. This reduced transcription errors and confirmed consistency across visits and patient cohorts.
- **Protocol Deviation Replication Monitoring** The team implemented parallel deviation tracking using both the EDC system and an internal regulatory log. Deviations were independently reviewed for classification consistency (major vs. minor, safety vs. operational), allowing replication of protocol compliance assessments.
- **Cross-Trial Endpoint Consistency Review** Because multiple concurrent obesity-related trials were conducted using similar endpoints, the team conducted periodic cross-protocol comparisons to examine whether weight-loss magnitude, adverse event reporting, and visit window adherence were producing consistent directional findings across studies with different sponsors.
- **Real-World Replication Through Post-Trial Follow-Up** A subset of participants was followed beyond trial completion through routine clinical care. Weight trajectories, medication continuation patterns, and metabolic markers were tracked and compared to sponsor-reported trial outcomes to assess real-world durability versus controlled-trial results.
- **Audit-Based Replication Validation** Monitoring visits and regulatory audits served as external replication checkpoints. Discrepancies between site source data and sponsor databases were cataloged and reconciled to validate the integrity and reproducibility of reported outcomes.

1.2 Outcome Evaluation Methods

- Concordance analysis between dual-recorded outcome measures
- Deviation trend analysis across multiple trials

- Comparative trajectory review of weight and metabolic outcomes across protocols
- Post-trial observational follow-up comparison with original trial findings
- Monitoring and audit resolution metrics as external validation

Together, these methods allowed continuous assessment of whether findings observed under tightly controlled trial conditions were reproducible within routine clinical operations and across different sponsors.

1.3 Innovative Features of the Strategy

- Integration of replication into daily clinic workflow, not as a separate research initiative
- Use of post-trial continuation tracking to extend replication beyond sponsor-defined endpoints
- Cross-sponsor endpoint harmonization, enabling replication across independent research programs
- Dual-layer verification of both outcomes and regulatory classification, not just numeric endpoints

This strategy demonstrated that replication can be operationalized as part of routine clinical trial infrastructure, rather than remaining limited to post-publication efforts.

1.4 Working Definition of Replication

For this strategy, replication was defined as:

The independent verification of clinical trial outcomes, protocol adherence, and safety reporting using parallel measurement, documentation, and audit-based validation across time, staff, and overlapping patient populations to confirm that study findings are repeatable, consistent, and not dependent on a single data pathway.

2. ENVIRONMENT

2.1 Environment Description

The replication activities were implemented within the context of active, sponsor-funded, multi-site clinical trials focused on obesity and cardiometabolic disease, with additional integration into routine post-trial clinical care. The environment was a high-volume outpatient clinical research setting where participants were concurrently enrolled in multiple industry-sponsored trials evaluating weight loss pharmacotherapies, including GLP-1–based interventions.

Because these trials involved iterative monitoring, regulatory oversight, and frequent safety and efficacy assessments, the site environment naturally supported the systematic integration of replication activities into daily research operations. Replication was not conducted as a separate project, but rather as a quality-controlled extension of standard clinical trial execution, allowing verification of outcomes across overlapping protocols, timepoints, and patient cohorts.

2.2 Location, Timing, and Team Details

Location:

- The activities took place at an independent clinical research site in Austin, Texas, specializing in metabolic, cardiovascular, and chronic disease trials.

Timing and Duration:

- The replication strategy was implemented continuously over a multi-year period between 2021 and 2024, aligning with the active enrollment and follow-up phases of multiple obesity-focused clinical trials.
- Replication activities occurred throughout screening, treatment, follow-up, and selected post-trial observational periods.

Team Composition and Roles:

- **Site Manager / Senior Clinical Research Operations Lead** Provided oversight of protocol execution, regulatory compliance, deviation classification, and cross-trial outcome verification. Led the integration of

replication workflows into routine operations and served as the primary liaison with sponsors and monitors.

- **Principal Investigator (PI)** Maintained medical oversight, verified safety assessments and adverse event causality, confirmed protocol adherence at a clinical level, and reviewed outcome consistency across trials.
- **Clinical Research Coordinators (CRCs)** Conducted participant visits, collected outcome measures, performed dual-source documentation, independently verified weight, vital signs, and laboratory data, and maintained protocol compliance logs.
- **Regulatory Specialist / Quality Lead** Managed internal deviation logs, audit preparation, document version control, and regulatory classification consistency across protocols. Led audit-based replication validation.
- **Laboratory Personnel** Performed specimen collection, processing, and shipment verification. Independently confirmed biomarker result reconciliation with sponsor-reported data.
- **Data and Finance Support Staff** Verified subject visit schedules, stipend tracking, and follow-up care coordination to ensure consistency between clinical activity and reported outcomes.

Together, this interdisciplinary team structure supported independent verification at multiple operational layers, enabling replication of both outcome measures and regulatory determinations.

2.3 Funding Support

The replication activities were supported indirectly through industry-sponsored clinical trial funding allocated for obesity and cardiometabolic studies conducted during the 2021–2024 period.

These funds supported:

- Personnel time for clinical visits, data verification, and regulatory oversight
- Laboratory testing and specimen processing
- Monitoring and auditing activities
- Post-trial observational follow-up embedded in routine care

No stand-alone federal replication grant funded this effort. Instead, replication was operationalized through institutional infrastructure supported by sponsor contracts and organizational research budgets. This demonstrates that meaningful replication activities can be integrated into real-world research environments without dedicated replication-only funding, provided that appropriate quality and verification systems are in place

3. IMPACT

3.1 Replication Activity Impact

Prior to formalizing the replication strategy, outcome verification and deviation tracking were performed using standard single-layer documentation and routine sponsor monitoring alone. Following implementation of the structured, dual-verification and cross-protocol replication workflow, several measurable improvements were observed.

Pre-replication vs. Post-replication Metrics:

- **Data discrepancy rate:** Pre-replication audits identified occasional transcription or timing discrepancies in outcome measures and visit documentation. Post-replication implementation reduced these discrepancies to near-zero across monitored endpoints.
- **Protocol deviation accuracy and classification:** Prior to replication integration, deviation identification primarily occurred during sponsor monitor review. After implementation, deviations were identified internally in real time, resulting in:
 - Earlier reporting
 - Improved root cause analysis
 - Enhanced consistency between site and sponsor deviation classification

- Monitoring findings: The number and severity of monitoring findings decreased following replication integration, reflecting stronger proactive quality control and reproducible documentation.
- Outcome consistency across trials: Cross-trial endpoint harmonization demonstrated high directional consistency in weight and metabolic outcomes across independent sponsor protocols, supporting the reproducibility of observed treatment effects under routine clinical conditions.

3.2 Long-Term Impact

The replication strategy had durable operational and scientific effects beyond the initial implementation period. Replication workflows became embedded into standard operating procedures and were maintained across subsequent obesity, cardiometabolic, and chronic disease trials.

Long-term effects included:

- Sustained reduction in audit and inspection risk
- Higher confidence in outcome reliability across sponsors
- Improved continuity between clinical trial data and post-trial real-world outcomes
- Stronger preparedness for regulatory inspections and sponsor due diligence

Additionally, the replication-informed mindset reinforced a culture of continuous verification, rather than reliance on downstream audits alone.

3.3 Research Rigor Improvements

The replication activities strengthened research rigor in several key domains:

- Measurement rigor: Dual-source verification minimized measurement drift and transcription bias.
- Temporal rigor: Continuous replication across visits improved visit window adherence and endpoint timing accuracy.
- Regulatory rigor: Parallel deviation tracking eliminated under-reporting and delayed classification.
- External validity: Post-trial outcome tracking improved alignment between controlled trial results and real-world clinical durability.

Collectively, these improvements increased the scientific credibility, internal reproducibility, and external trustworthiness of study findings reported from the site.

3.4 Dissemination and Adoption

The replication strategy was disseminated informally through:

- Sponsor monitoring interactions
- Internal staff training and onboarding workflows
- Regulatory audit preparation processes
- Cross-trial operational meetings with CRAs and study teams

Elements of the strategy, including dual verification of outcomes and parallel deviation tracking, were adopted as standard practice across additional protocols and therapeutic areas within the organization. While no formal publication resulted from this operational replication strategy, its principles were successfully scaled across multiple studies and sponsors, demonstrating high feasibility and strong transferability.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

Several operational, cultural, and logistical barriers were encountered during the implementation of the replication strategy. These barriers occurred across multiple organizational levels and were addressed through targeted workflow refinement, staff engagement, and leadership support.

1. Time and Staffing Constraints (Team-Level Barrier): A primary barrier was the additional time required for dual verification of outcomes, parallel deviation tracking, and cross-protocol endpoint review. At baseline, staff were already managing high visit volumes, sponsor demands, and regulatory documentation.

Solution: The team addressed this by:

- Integrating replication steps into existing visit workflows rather than adding separate tasks
- Standardizing verification checklists to reduce redundancy
- Cross-training staff to ensure coverage during peak visit weeks

This approach embedded replication into routine work without requiring additional full-time staffing.

2. Initial Staff Resistance to Workflow Change (Team-Level and Site-Level Barrier): Some staff initially perceived replication requirements as duplicative or burdensome, particularly during early implementation.

Solution: Leadership addressed this through:

- Education on the role of replication in patient safety and data integrity
- Sharing early improvements in monitoring outcomes and reduction in findings
- Incorporating replication responsibilities into job expectations and SOPs

As staff saw direct improvements in audit readiness and data quality, buy-in increased and the process became normalized.

3. Variability Across Sponsor Protocol Requirements (Organizational-Level Barrier): Different sponsors used varied definitions for deviations, endpoint thresholds, visit windows, and reporting expectations, which complicated cross-trial replication.

Solution: The team created internal harmonization standards that aligned overlapping endpoints (e.g., weight, vitals, labs, safety visits) while still meeting sponsor-specific requirements. This allowed the team to replicate outcomes using consistent internal benchmarks without violating protocol-specific rules.

4. Post-Trial Follow-Up Limitations (Community and System-Level Barrier): Tracking post-trial outcomes was challenged by:

- Insurance coverage limitations
- Participant relocation
- Medication access changes after sponsor support ended

Solution: The team leveraged routine clinical care visits, telehealth follow-ups, and external lab reports when available to maintain continuity of outcome tracking. While complete longitudinal capture was not always possible, partial follow-up still provided valuable real-world replication insight.

5. Data System Limitations (Technology and Infrastructure Barrier): EDC platforms and internal site systems were not always designed to support parallel data comparison or cross-protocol analysis.

Solution: The team developed supplemental internal tracking logs and reconciliation tools outside the EDC to:

- Cross-verify outcomes
- Track deviation concordance
- Compare directional trends across overlapping studies

These tools served as interim replication infrastructure within existing technical limitations.

4.2 Organizational Levels Addressed

- Team Level: Workflow integration, staff training, role clarification
- Site Level: SOP development, internal harmonization standards, audit readiness
- Organizational Level: Sponsor communication alignment and cross-protocol coordination
- Community/System Level: Post-trial access to care and medication continuity

4.3 Overall Outcome

Despite these barriers, the replication strategy was successfully implemented and sustained due to:

- Strong leadership advocacy
- Staff engagement through education and feedback
- Flexible adaptation to sponsor and system constraints
- Integration of replication into routine clinical research infrastructure

These efforts ensured that replication activities remained feasible, durable, and scalable within a real-world clinical research setting.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

The replication strategy was implemented in a phased, scalable manner across active obesity and cardiometabolic clinical trials between 2021 and 2024.

Phase 1: Strategy Development and Pilot Integration (Q1–Q3 2021)

- Identified key endpoints appropriate for replication (weight, vitals, labs, deviations)
- Developed internal dual-verification and parallel deviation tracking workflows
- Piloted replication procedures in one obesity-focused protocol
- Trained core staff on verification standards and documentation expectations

Phase 2: Full Operational Integration (2022)

- Expanded replication workflows across multiple concurrent obesity and metabolic trials
- Implemented cross-protocol endpoint harmonization
- Integrated replication verification into routine monitoring and audit preparation
- Initiated limited post-trial observational follow-up tracking

Phase 3: Optimization and Sustainability (2023–2024)

- Standardized replication procedures into internal SOPs
- Refined post-trial follow-up processes
- Sustained dual-verification and cross-validation across additional therapeutic areas
- Embedded replication activities into staff onboarding and training

Replication activities have remained active and sustainable beyond the initial implementation period and are now part of routine research operations.

5.2 Associated Costs

No stand-alone federal replication grant supported this effort. Costs were embedded within industry-sponsored clinical trial budgets and organizational operational funding.

Estimated replication-related cost categories included:

- Personnel time for dual verification and regulatory review: Included CRCs, regulatory staff, and site leadership oversight
- Monitoring and audit preparation time: Supported through sponsor monitoring visit budgets
- Post-trial outcome tracking: Conducted using routine clinical care and participant follow-up
- Supplemental internal tracking systems: Low-cost development using office software and secure internal logs

Because replication was integrated into existing infrastructure, no significant new capital expenditures were required. This demonstrates that meaningful replication can be achieved through workflow optimization rather than large, dedicated funding mechanisms.

5.3 Required Resources

The following resources were required to achieve replication outcomes:

Human Resources

- Principal Investigator for medical oversight
- Site Manager / Senior Operations Lead for replication integration
- Clinical Research Coordinators for dual-source outcome verification
- Regulatory Specialist for deviation tracking and audit reconciliation
- Laboratory staff for specimen verification
- Data and finance staff for visit and follow-up validation

Operational Resources

1. Electronic Data Capture (EDC) systems
2. Internal regulatory and deviation tracking logs
3. Secure document storage and audit-ready filing systems
4. Post-trial follow-up coordination tools
5. Monitoring and audit support infrastructure

Training Resources

- Internal SOPs
- Staff onboarding and refresher training
- Monitoring feedback and audit resolution sessions

Together, these resources supported a sustainable, low-cost, high-impact replication infrastructure that strengthened research rigor without disrupting trial efficiency.

6. COLLABORATION

6.1 Key Collaborations

Successful implementation of the replication strategy required strong collaboration across multiple stakeholders within the clinical research ecosystem.

The most critical partnerships included:

- Sponsor Partners (Pharmaceutical and CRO Teams): Essential for providing protocol frameworks, monitoring oversight, and access to comparative outcome data across sites.
- Principal Investigator (PI) Leadership: Medical oversight was necessary to validate safety outcomes, adverse event causality, and clinical interpretation of replicated findings.
- Internal Regulatory and Quality Teams: Required to harmonize deviation classification, manage audit readiness, and ensure compliance with Good Clinical Practice (GCP).
- Laboratory and Diagnostic Partners: Critical for independent sample verification, biomarker result reconciliation, and endpoint confirmation.

In retrospect, earlier formal involvement of health system partners and data analytics collaborators could have further strengthened large-scale real-world outcome replication and long-term post-trial durability assessments.

6.2 Leveraging Existing Partnerships

The replication strategy relied heavily on existing sponsor, CRO, and institutional partnerships, which were leveraged in several ways:

- Sponsor Monitoring Teams: Monitoring visits were used not only for compliance verification, but as external replication checkpoints that validated internal outcome verification and regulatory classification consistency.
- CRO Data Management Teams: These partnerships supported cross-comparison of site-reported data with centralized datasets, allowing for ongoing reconciliation and validation.

- Internal Clinical and Regulatory Staff: Long-standing working relationships across clinical operations, regulatory affairs, and laboratory services enabled smooth integration of replication without disrupting trial execution.
- Audit and Inspection Partners: Regulatory inspections and sponsor audits served as additional layers of independent validation, reinforcing the credibility of replicated outcomes.

These pre-existing relationships allowed replication to be embedded into routine research infrastructure without creating parallel external systems.

6.3 Building New Partnerships

In support of expanded replication objectives, new partnerships were developed organically as the strategy matured:

- Post-Trial Clinical Care Providers: Relationships were strengthened with community providers and specialty clinics to support limited post-trial outcome tracking and medication continuity assessment.
- Cross-Protocol Internal Teams: New collaboration channels were created between staff working across different sponsor trials to facilitate cross-trial endpoint harmonization and comparative replication review.
- Training and Quality Improvement Networks: Informal partnerships were built through shared training sessions, monitoring feedback loops, and internal quality improvement discussions focused on replication principles.

While no formal academic replication consortium was created, these new partnerships expanded the functional replication network and improved the ability to validate outcomes beyond a single protocol or sponsor.

6.4 Overall Collaborative Impact

Through a combination of leveraged long-standing partnerships and newly formed collaborative relationships, the replication strategy achieved:

- Multi-layer independent verification
- Cross-sponsor outcome validation
- Real-world post-trial continuity
- Sustainable integration of replication into routine clinical research operations

These partnerships ensured that replication was not isolated to a single team or dataset, but supported through a connected clinical research ecosystem.

7. SUSTAINABILITY / TRANSFERABILITY

The replication activities described are highly sustainable, transferable, and generalizable because they were intentionally designed to operate within standard clinical research infrastructure rather than as a stand-alone or grant-dependent initiative. By embedding replication into routine workflow, SOPs, monitoring preparation, and data verification processes, the strategy does not rely on temporary funding, specialized external platforms, or unique site characteristics.

The approach is generalizable across therapeutic areas, including cardiometabolic disease, infectious disease, neurology, women's health, and behavioral research, because it focuses on universal research functions:

- Outcome verification
- Protocol compliance
- Deviation classification
- Cross-study reproducibility
- Post-trial durability checks

These functions exist in all regulated clinical research environments.

7.1 Current Sustainability Plans

Replication activities are currently sustained through:

Permanent integration into internal SOPs for:

- Outcome verification
- Deviation tracking
- Monitoring and audit preparation
- Ongoing staff training and onboarding, where replication principles are taught as a core research competency
- Routine performance monitoring, where replication quality is reviewed alongside standard clinical metrics
- Leadership oversight, ensuring replication remains a standing operational expectation rather than an optional activity

Because replication is now embedded into daily operations and compliance culture, it requires no separate sustainability grant and continues as part of routine research execution.

7.2 Scaling and Application

These replication activities can be scaled across multiple dimensions:

1. Across Study Volume: The workflow can expand seamlessly as study volume increases because verification is performed at the visit level and audit level rather than requiring centralized replication infrastructure.

2. Across Therapeutic Areas

The approach applies equally to:

- Obesity and diabetes
- Cardiovascular disease
- Neurology
- Women's health
- Infectious disease
- Behavioral and psychiatric research

3. Across Organizational Types

The model is transferable to:

- Independent research sites
- Academic medical centers
- Health-system-based research programs
- Community-based research networks

4. Across Regulatory Environments: Because the strategy aligns directly with ICH-GCP, FDA, and sponsor monitoring expectations, it is portable across regulatory jurisdictions and sponsor types.

7.3 Future Integration Plans

Several future enhancements are planned or under consideration:

Formalization of a Replication Training Program Development of structured internal modules focused on:

- Dual verification methods
- Deviation concordance
- Cross-protocol endpoint harmonization

Replication-Focused Quality Dashboards

Creation of internal dashboards that track:

- Verification concordance rates
- Deviation trend metrics
- Monitoring and audit outcomes

Expanded Post-Trial Outcome Partnerships

Strengthening relationships with healthcare providers, specialty clinics, and payers to improve:

- Long-term outcome tracking
- Medication discontinuation analysis
- Real-world durability assessment

Broader Dissemination and Knowledge Sharing

Plans include:

- Sharing best practices across sponsor networks
- Presenting replication-informed operations during monitoring and audit close-out discussions
- Incorporating replication principles into internal leadership and quality forums

Institutional Policy Evolution Replication elements are expected to be incorporated into future:

- Quality management plans
- Monitoring readiness policies
- Data integrity frameworks

7.4 Overall Sustainability and Future Outlook

By embedding replication into core operational systems rather than external research infrastructure, this strategy demonstrates that replication can be sustained indefinitely, transferred broadly, and scaled without financial burden. Future expansion will further strengthen the role of replication in improving data credibility, patient safety, sponsor confidence, and regulatory trust across the clinical research enterprise.

TEAM WINNER

LATCH (LLM-Assisted Testing of Clinical Hypotheses)



Developed an LLM-assisted framework that converts written study descriptions into executable analyses, making replication faster and surfacing methodological gaps.

Team Members: Nayoon Gim · In Gim · Cecilia S Lee · Aaron Y Lee

Our work focuses on LATCH (LLM-Assisted Testing of Clinical Hypotheses), a framework that translates study details in text into executable analyses, facilitating efficient replication while identifying gaps in methodological reporting. We are motivated by improving transparency and reproducibility in clinical research.

Nayoon Gim is an MD/PhD student in Bioengineering at the University of Washington, leading the development of the LATCH framework and driving study design and replication analyses.

In Gim is a PhD student in Computer Science at Yale University, contributing to system architecture design and environment setup for the LLM-integrated platform.

Cecilia S. Lee MD, MS is a physician-scientist and Professor of Ophthalmology & Visual Sciences at Washington University in St. Louis, contributing expertise in clinical epidemiology and study interpretation.

Aaron Y. Lee MD, MSCI is a physician-scientist and Chair and Professor of Ophthalmology & Visual Sciences at Washington University in St. Louis, providing overall guidance and research oversight.

1. REPLICATION STRATEGY

1.1 Overview of Replication Strategy

1.1.1 Problem Introduction

The use of large-scale population health datasets has increased, yet converting a clinical hypothesis into verified, reproducible evidence remains slow, expensive, and difficult to audit because analytic workflows are often implemented in an idiosyncratic and ad hoc manner. In many computational clinical studies, researchers construct a lengthy translation process from a natural-language clinical hypothesis with detailed study specifications to executable cohort definitions and statistical models. These workflows are typically built from scratch for each study, relying on custom scripts to define inclusion and exclusion criteria, harmonize variables across years or data cycles, clean and transform data, and execute statistical tests, a time-consuming and labor-intensive process that demands substantial domain and programming expertise (Hripcsak & Albers, 2013; Kahn & Weng, 2012; Richesson et al., 2013). Even when implemented carefully, these workflows are often only partially documented in publications, leaving a gap between reported results and the analytic logic used to generate them. This gap limits reproducibility and complicates independent validation of published findings. This gap slows replication, discourages routine robustness checks, and contributes to the broader reproducibility challenges in computational medicine (Baker, 2016; Cobey et al., 2024; Niven et al., 2018; Peng, 2011).

This gap is particularly consequential given established standards for computational reproducibility. Peng et al. describe a spectrum of reproducibility in which the level of trust in a scientific result depends on how much of the analytic process is shared, with the highest level—often considered the “gold standard” in computational science—achieved when both the raw data and complete analysis code are made available for independent verification (Peng, 2011). In practice, however, assembling and releasing fully executable analytic code remains labor-intensive and inconsistently implemented in clinical research.

1.1.2 Strategy Description: LATCH as a Replication Framework

The core of the strategy is LATCH (LLM-Assisted Testing of Clinical Hypotheses), a framework designed to translate a researcher’s natural-language description of a study into an executable cohort extraction and statistical

analysis pipeline. Growing utilization of large language models (LLMs) has shown promise in improving efficiency across scientific research domains (Boiko et al., 2023; Cai et al., 2025; Oniani et al., 2024). The strategy is not merely automation for speed; it is automation designed for verifiability and reproducibility. Each LATCH run produces an Analytic Report—an immutable audit trail that records all key artifacts from question to variable selection, code for patient cohort construction and statistical analysis. LATCH is a framework for automated clinical hypothesis testing, but in this project it is used exclusively as a replication instrument, with inputs derived from the methods descriptions of previously published studies so that the resulting analyses represent re-executions of existing work rather than new hypothesis-driven investigations. This allows replication to be treated as a repeatable procedure: a published finding is converted into a structured analytic specification, executed consistently, and evaluated using pre-defined concordance metrics. In effect, when applied to publicly available datasets, LATCH supports the highest level of computational reproducibility described by Peng et al (Peng, 2011) by capturing analysis logic in executable form and pairing it with a complete, auditable record of how each result was produced. Rather than relying on narrative natural language descriptions that cannot encode thousands of discrete steps, LATCH makes those steps explicit and reviewable. Another feature is that LATCH is designed to preserve privacy, only exposing the database schema (e.g., table and column names) to the LLM, while strictly isolating all patient-level data from any LLM-involved steps, addressing concerns about LLM use in clinical settings (Dennstädt et al., 2025; Jonnagaddala & Wong, 2025; Wiest et al., 2025).

We evaluated the proposed replication strategy by applying LATCH to a systematic set of published diabetes-related studies using the National Health and Nutrition Examination Survey (NHANES)(NHANES Questionnaires, Datasets, and Related Documentation, n.d.). Twenty peer-reviewed NHANES-based studies published between 2010 and 2025 were systematically selected based on methodological clarity, citation impact, and relevance to diabetes-related health outcomes. These studies span diverse clinical domains, including diabetic retinopathy, nephropathy, mortality, mental/cognitive health, and represent a range of statistical methodologies such as logistic and linear regression, Cox proportional hazards models, and descriptive statistics. For each study, methods described in the original publication were extracted and provided to LATCH as natural-language inputs, enabling the analyses to be independently re-executed using a standardized replication protocol (Fig. 1a). Replication success was assessed using multiple concordance metrics, including cohort fidelity, effect direction, statistical significance, and confidence interval overlap. Across studies, LATCH demonstrated high overall concordance with published findings while also revealing informative discrepancies attributable to methodological ambiguity and variable definition differences. In addition, the end-to-end execution time for individual replication analyses ranged from approximately 1.6-15.1 minutes, depending on analytic complexity and the number of survey cycles included. Together, these results illustrate both the feasibility of automated direct computational replication at scale and the value of structured, auditable workflows for diagnosing reproducibility limitations in published clinical research. Source code associated with this work is available in the LATCH project's GitHub repository.

1.1.3 Working definition of replication & Category Identification

In this work, replication is defined as direct computational replication, in which previously published clinical studies are independently re-executed using the methods as documented in the publication. The replication activities described here fall under the Clinical research category

1.2 LATCH Architecture and Innovation Points

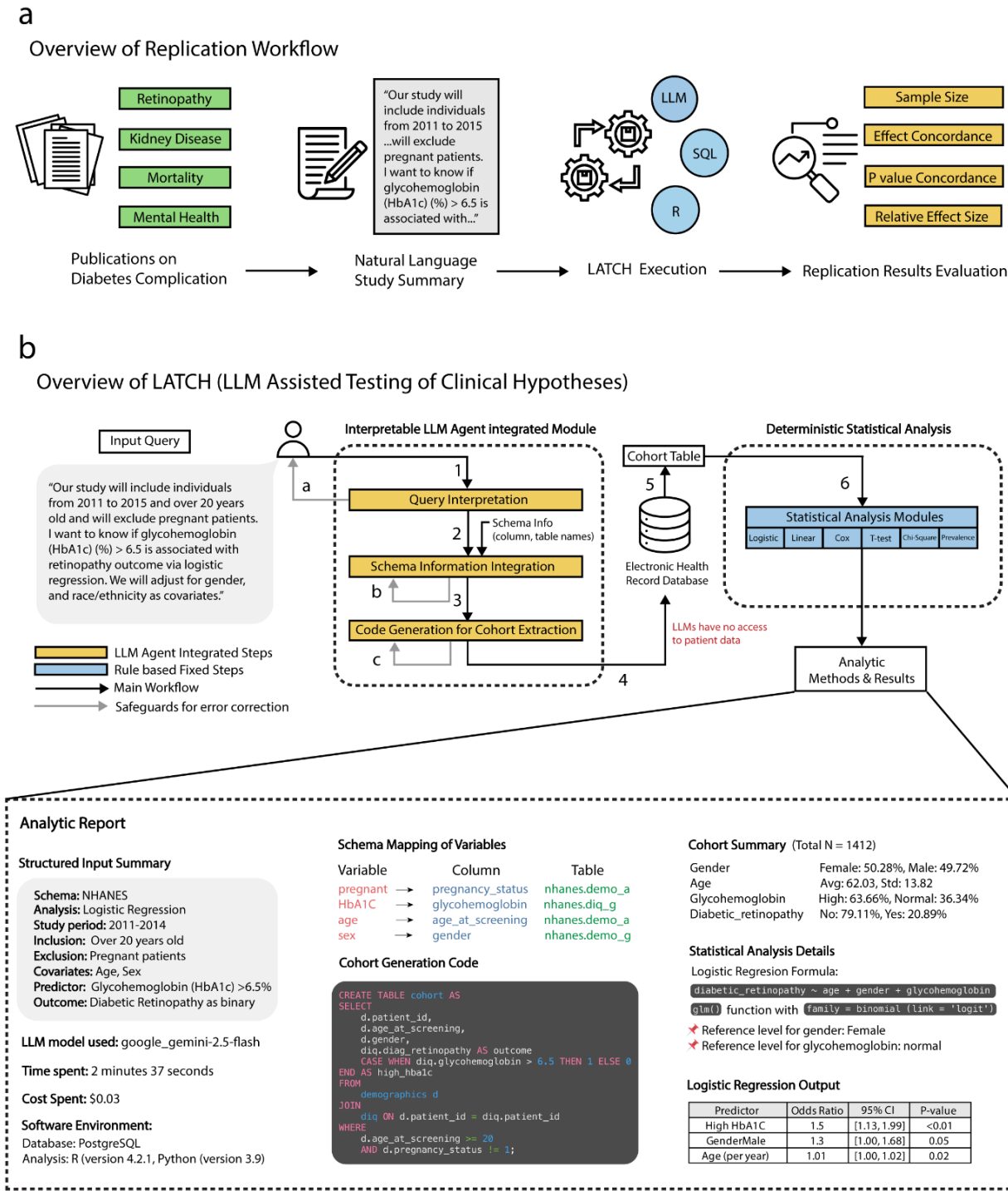


Figure 1. Replication Workflow and Architecture of the LATCH Framework (LLM-Assisted Testing of Clinical Hypotheses). (a) Overview of the replication workflow. The workflow begins with curation of publications on diabetes complications (including diabetic retinopathy, diabetic kidney disease, mortality, and mental and cognitive health) that used data from the National Health and Nutrition Examination Survey (NHANES), followed by extraction of natural-language study details to serve as LATCH input, execution of LATCH, and final evaluation of replication results against the original publications using metrics including sample size, effect concordance, p-value concordance, and relative effect size. (b) Schematic overview of the LATCH system. (1) A user’s natural-language query is processed by the LLM module for query interpretation. The resulting structured representation (2) is passed to the schema information integration step, where database schema information is incorporated, producing a schema-aligned query (3). This query is used for code generation, yielding executable code (4) that

queries the database to define a patient cohort (5), which is subsequently analyzed by a deterministic statistical analysis module (6). Safeguards are implemented across all LLM-integrated steps (arrows a–c) to mitigate failure cases. The entire workflow is documented in a detailed Analytic Report to ensure transparency and reproducibility.

1.2.1 High-level architecture: a hybrid system for verifiable automation

LATCH is engineered to mirror the real workflow of clinical research while making the entire pipeline auditable end-to-end. The architecture is intentionally hybrid and modular, separating semantic interpretation from deterministic computation (Fig. 1b). An Interpretive Module uses large language models (LLMs) to translate a natural-language study description into a structured, machine-actionable analytic plan and cohort definition logic, while an Analysis Module executes cohort extraction and statistical modeling using fixed, validated procedures to ensure consistency across runs. Python serves as the orchestration layer that integrates the LLM API, the relational database execution (e.g., PostgreSQL), and statistical execution in R via rpy2, functioning as the single control plane that standardizes execution and logging. The key innovation is not simply that an LLM is used, but that it is used in a tightly constrained role while generating a permanent record of what was inferred and what code was executed.

1.2.2 Interpretive Module: a three-step agentic workflow

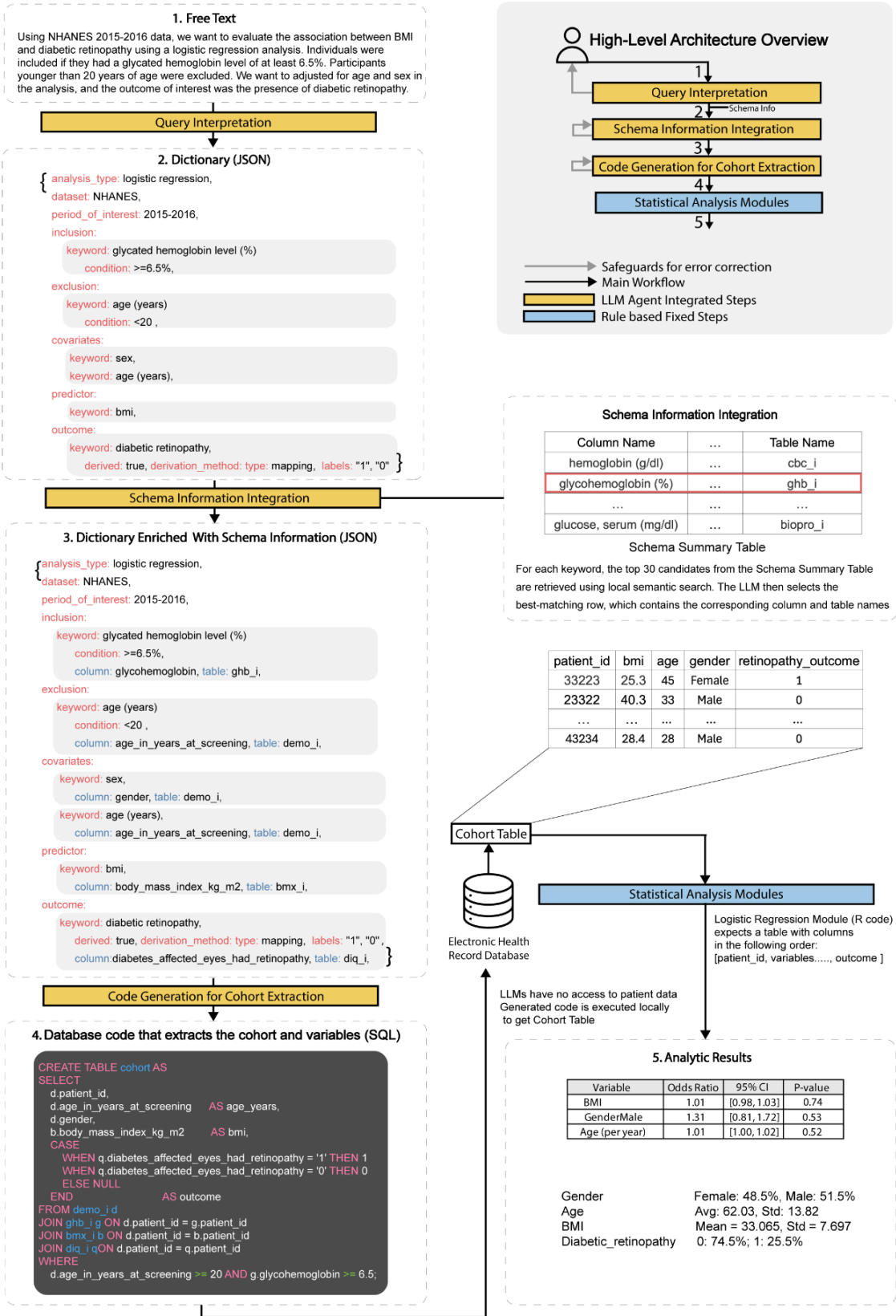
To reduce brittleness and increase traceability, LATCH decomposes the difficult translation from “methods prose to executable analysis” into three verifiable steps, each producing an intermediate artifact that can be reviewed (Fig. 1, Supplementary Fig. 1). First, query decomposition converts the natural-language prompt into a structured representation aligned with clinical methods reporting, capturing the analysis type, study period, inclusion and exclusion criteria, predictors, outcomes, and covariates, and then expressing this information as a hierarchical JSON plan that explicitly represents nested logic and derived variables. Second, schema grounding maps the concepts in this plan to concrete database variables using a hybrid approach in which a semantic retrieval model proposes candidate variables from the schema table and the LLM performs contextual disambiguation to select the most appropriate match. Third, SQL query generation converts the fully grounded plan into executable SQL that constructs the analytic cohort locally in the database, with standardized output column ordering to ensure compatibility with the deterministic analysis module.

1.2.3 Statistical Analysis Module: reproducible statistics by design

LATCH enforces a strict separation between flexible language interpretation and rigid numerical execution by placing statistical modeling under a deterministic, rule-based engine. After the cohort is extracted and returned as a standardized analytic table (e.g., a pandas DataFrame), the engine performs automated preparatory steps such as feature classification, identifying variables as numerical, binary, or categorical so that model parameterization is consistent. The engine then executes the specified analysis type—such as logistic regression, linear regression, Cox proportional hazards models, descriptive statistics, or prevalence calculations—through fixed procedures implemented using established R packages. This ensures that numerical outputs are reproducible across repeated runs given the same inputs.

1.2.4 Safeguard system: resilience against common failure modes

To support replication at scale, LATCH includes a multi-tier safeguard mechanism designed to reduce errors and increase operational robustness. Safeguard 1 performs user input vetting and content validation, checking that a request is internally coherent and executable—for example, that a logistic regression includes a valid binary outcome and a predictor—and verifying that requested datasets and study years exist; ambiguous or incomplete requests trigger structured warnings and require clarification before execution. Safeguard 2 targets semantic-to-variable mapping and harmonization, mitigating variable drift across temporal dataset cycles by filtering for candidate variables that remain consistent across the requested years; when multiple candidates remain, an LLM selects the best match using full contextual information, and when no consistent variable exists the system applies year-specific mappings while explicitly logging the inconsistency for review. Safeguard 3 is an iterative SQL fix loop that activates when generated SQL fails to execute, feeding the failed SQL, error message, schema context, and original query back into the LLM to produce corrected SQL, iterating with a bounded retry limit to prevent one-off failures from derailing end-to-end replication.



Supplementary Figure 1. Stepwise input–output transformation in the LATCH analytic pipeline. The figure illustrates the full end-to-end analytical workflow using an example NHANES research query, with the input and output at each processing stage shown explicitly. A free-text natural-language research question specifying the dataset, study period, inclusion and exclusion criteria, covariates, predictor, and outcome serves as the initial input. During query interpretation (not shown in detail), this

free-text input is transformed into a structured, machine-readable specification, represented as a JSON dictionary encoding the analysis type, cohort definition, logical conditions, variable derivations, and analysis methods. During schema information integration, the structured dictionary is augmented by mapping conceptual keywords to dataset-specific column names and table identifiers using a schema summary derived from a dataset-provided data dictionary, resulting in a schema-enriched JSON representation. This schema-enriched specification is then used for code generation, producing executable SQL that applies cohort inclusion and exclusion criteria, derives variables, and extracts the analytic cohort and feature table from the database. Finally, the extracted cohort table is passed to deterministic statistical analysis modules, which generate analytic outputs including cohort summaries and regression results such as effect estimates, confidence intervals, and P values. LLMs do not access patient-level data; all generated code is executed locally.

1.3 Key Innovations

1.3.1. The selective LLM principle

A foundational design choice in LATCH is the selective use of LLMs: the model is employed only for tasks that require semantic understanding, such as interpreting ambiguous clinical language, resolving nested logic expressed in prose, and selecting the most context-appropriate variable from a candidate list. In contrast, all steps that require invariance and strict reproducibility, such as code for running statistics analysis, are handled by deterministic code. This division of labor is especially important for replication, because allowing an LLM to freely choose modeling procedures or adjust parameters could introduce run-to-run variability unrelated to the scientific question.

1.3.2. Privacy-first design via metadata injection (schema-only interaction)

Because clinical data are sensitive, LATCH is designed so that the LLM never receives patient-level records. Instead, it relies on a one-time schema preparation procedure that constructs a Schema Summary Table using only public documentation such as data dictionaries, codebooks, variable labels, and value encodings. Cohort extraction and all computations over patient-level data occur locally within the secured compute environment, while the LLM operates only over schema metadata and produces mappings and code that can be reviewed.

1.3.3. The Analytic Report as a verifiable audit trail

The most replication-relevant innovation in LATCH is the Analytic Report, which functions as an immutable “analytic report” for every run. Rather than providing only results, the report systematically compiles the raw outputs from each stage of the pipeline, including the original natural-language prompt, the structured interpretation of the analysis plan, the final schema mappings from concepts to variables, the complete SQL used to build the cohort, and cohort summary statistics that characterize the extracted population (Supplementary Fig. 2) . It also records the exact statistical model specification executed in R, including the formula, parameter settings, categorical reference levels, and survey-weight choices when applicable, along with software environment details such as package versions.

Analytic Report

User Input

Using NHANES 2013-2016 data, we want to evaluate the association between body mass index (BMI) and diabetic retinopathy using a logistic regression analysis. Individuals were included if they had a glycated hemoglobin level of at least 6.5%. Participants younger than 20 years of age were excluded. BMI is the primary predictor. We want to adjusted for age and sex in the analysis, and the outcome of interest was the presence of diabetic retinopathy.

Structured Input Summary

Analysis Type: logistic regression

Dataset: NHANES

Period of Interest: 2013-2016

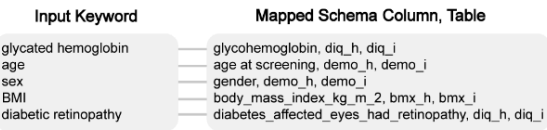
Inclusion Criteria:
- glycated hemoglobin level >= 6.5%

Exclusion Criteria:
- age < 20 years

Covariates:
- age
- sex

Predictor:
- body mass index (BMI)

Outcome:
- presence of diabetic retinopathy = 1 if present; 0 otherwise



Database Cohort Extraction Code (SQL)

```
-- ===== STEP 0: Master Table Formation =====
- TEMP TABLE FOR "glycated hemoglobin level (%)"
DROP TABLE IF EXISTS temp_master_glycohemoglobin;
CREATE TEMP TABLE temp_master_glycohemoglobin AS
SELECT respondent_sequence_number, "glycohemoglobin" AS
"glycohemoglobin"
FROM nhanes.ghb_h
UNION ALL
SELECT respondent_sequence_number, "glycohemoglobin" AS
"glycohemoglobin"
FROM nhanes.ghb_i;
-- Add index for faster joins
CREATE INDEX idx_temp_master_glycohemoglobin ON
temp_master_glycohemoglobin (respondent_sequence_number);
⋮
-- ===== STEP 1: Inclusion =====
-- STEP 1.1: Includes individuals with a glycated hemoglobin level of at least
6.5%
CREATE TEMP TABLE temp_inclusion_step1 AS
SELECT "respondent_sequence_number"
FROM final_master_table
WHERE "glycohemoglobin" >= 6.5;
⋮
-- ===== STEP 2: Exclusions =====
-- ===== STEP 3: Cohort =====
-- ===== STEP 4: Variables =====
⋮
-- ===== STEP 5: Final Table =====
CREATE TEMP TABLE temp_final_table AS
SELECT
c."respondent_sequence_number",
age."age_years",
sex."sex",
bmi."body_mass_index",
dr."diabetic_retinopathy"
FROM temp_cohort c
JOIN temp_age_years age ON c."respondent_sequence_number" =
age."respondent_sequence_number"
JOIN temp_sex sex ON c."respondent_sequence_number" =
sex."respondent_sequence_number"
JOIN temp_body_mass_index bmi ON c."respondent_sequence_number" =
bmi."respondent_sequence_number"
JOIN temp_presence_of_diabetic_retinopathy dr ON
c."respondent_sequence_number" = dr."respondent_sequence_number";
-- ===== STEP 6: Final Output =====
SELECT * FROM temp_final_table;
```

Analysis code (R)

```
# Package Loading

suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(broom))
suppressPackageStartupMessages(library(tibble))

# Descriptive Analysis for Cohort Summary

identifier_vars <- c("respondent_sequence_number")
outcome_vars <- c("presence_of_diabetic_retinopathy")
cohort <- df
outcome_roles <- c()
outcome_roles["presence_of_diabetic_retinopathy"] <- "outcome"
⋮

data.frame(
  variable = name,
  category = category,
  n_unique = n_unique,
  na_count = na_count,
  summary = summary_info,
  stringsAsFactors = FALSE
)
}
report <- bind_rows(
  lapply(names(cohort), function(col_name) {
    summarize_variable(cohort[[col_name]], col_name)
  })
)

# Logistic Regression Analysis
⋮

formula <- as.formula(paste0(outcome_var, "~ ", paste(predictor_vars, collapse = " + ")))
model <- glm(formula,
  data = df,
  family = binomial(link = "logit"))

result_table <- tidy(model, conf.int = TRUE, exponentiate = TRUE) %>%
  select(term, estimate, p.value, conf.low, conf.high)
```

LLM model used: google_gemini-2.5-flash

Time spent: 3 minutes 13 seconds

Cost Spent: \$0.04

Software Environment:

Database: PostgreSQL
Analysis: R (version 4.2.1, Python (version 3.9))

Cohort Summary

Variable	Summary
respondent_sequence_number	1295 unique rows
sex	Female: 50.28%, Male: 49.72%
age	Avg: 62.03, Std: 13.82
body_mass_index	Avg: 32.94, Std: 7.79
diabetic_retinopathy	0: 74.9%, 1:25.1%

Statistical Analysis Details

Logistic Regression Formula:

```
diabetic_retinopathy ~ age + sex + body_mass_index
glm() function with family = binomial (link = 'logit')
```

★ Reference level for gender: Male

Logistic Regression Output

Predictor	Odds Ratio	95% CI	P-value
body_mass_index	1.01	[0.99, 1.03]	0.30
sex_female	0.76	[0.56, 1.04]	0.08
age_years	1.01	[0.98, 1.01]	0.91

Supplementary Figure 2. Example analytic report generated by LATCH. Illustrative analytic report documenting an end-to-end analysis generated from a natural-language research query. The report includes the original user input, a structured study design summary specifying analysis type, dataset, inclusion and exclusion criteria, covariates, predictor, and outcome, and the corresponding schema mappings matching user-specified concepts to database tables and columns. Generated cohort extraction code (SQL) is shown alongside downstream statistical analysis code (R), enabling transparent reconstruction of cohort definition and modeling steps. Additional metadata, including execution time, API cost, software environment, cohort characteristics, and regression outputs, are recorded to support auditability, reproducibility, and tracking of each analysis.

1.3.4. *The economics of time and cost*

LATCH executes end-to-end replication analyses with observed runtimes ranging from approximately 1.6 to 15.1 minutes per analysis, depending on query complexity and the number of survey cycles included. The computational cost of each replication run is dominated by large language model (LLM) API usage, with observed API costs ranging from approximately USD 0.05 to USD 0.40 per analysis (Gemini 2.5 Flash), varying with prompt length and schema search space. Local compute overhead for database execution and statistical modeling was negligible relative to API cost.

1.4. Replication Study and Evaluation

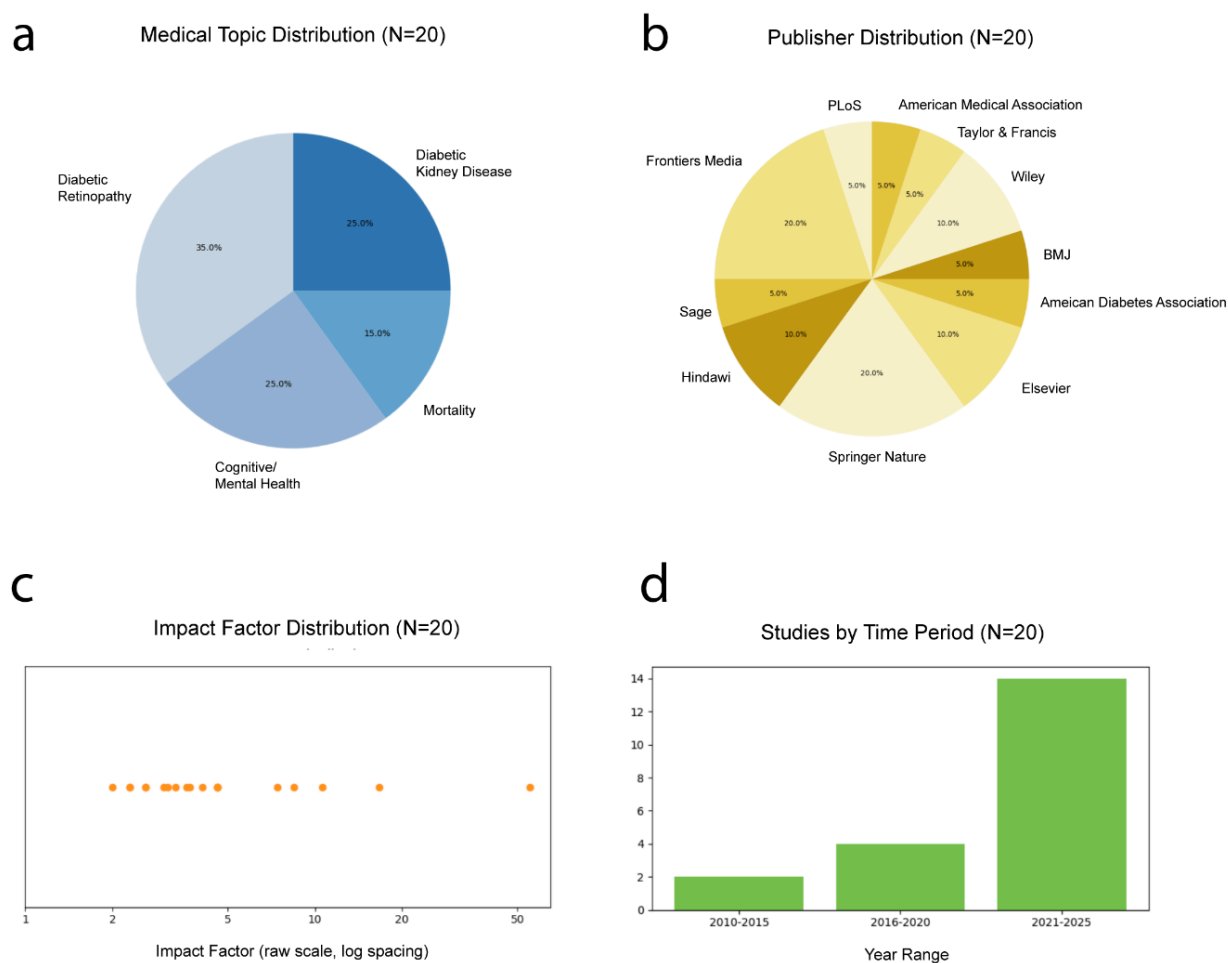
1.4.1 *Replication Study Methods*

1.4.1.1 Dataset and setting

We evaluated LATCH using NHANES, a large epidemiological dataset spanning multiple decades, chosen because it supports diverse clinical questions and contains many variables relevant to diabetes and its complications. We further enriched the dataset for time-to-event analyses by linking NHANES mortality files, enabling survival outcomes. The data were ingested and harmonized into a relational database (PostgreSQL), and the system ran within a computational environment where patient-level data processing remained local. LLM calls were used only for semantic planning and mapping, with strict separation from row-level data.

1.4.1.2 Study selection and curation (2010–2025)

To validate scientific accuracy across real published evidence, we conducted a systematic replication of 20 peer-reviewed diabetes-related NHANES studies published between 2010 and 2025. Studies were chosen to represent diversity in outcomes and methods (Supplementary Fig. 3). We targeted four outcome areas: diabetic retinopathy (Dai et al., 2025; X. Li & Chen, 2023; G.-X. Wang et al., 2022; H. Wang et al., 2022; Xu et al., 2024; Yang et al., 2024; Zhang et al., 2010), diabetic kidney disease (Guo et al., 2022; Liang et al., 2024; X. Li et al., 2023; Mottl et al., 2013; Z. Wang et al., 2024), all-cause mortality (Al-Kindi et al., 2017; Ding et al., 2024; B. Li et al., 2023), and mental/cognitive health (Casagrande et al., 2021; Lee et al., 2021; Montgomery et al., 2017; J. Wang et al., 2020; Y. Wang et al., 2016). Candidate papers were identified through a structured PubMed search constrained to NHANES-based diabetes cohorts and were ranked by citation counts to prioritize influential studies. We excluded papers that did not use standard NHANES as the primary dataset, did not focus on diabetes-restricted cohorts, duplicated previously selected exposure-outcome-method combinations, or lacked sufficient methodological detail for computational reproduction. We also excluded studies requiring complex external computations or data structures not supported by the current LATCH modules, as well as analyses relying on external databases or highly specialized methods outside the framework's validated statistical suite. This selection process resulted in a benchmark set designed not only to test LATCH under ideal conditions, but also to reflect typical reproducibility constraints in the published literature.



Supplementary Figure 3. Characteristics of the replication studies included in the evaluation. (a) Medical topic distribution of the 20 selected replication studies, spanning diabetic retinopathy, diabetic kidney disease, cognitive or mental health outcomes, and mortality. (b) Publisher distribution of the replication studies, demonstrating coverage across a range of biomedical journals and publishers. (c) Distribution of journal impact factors for the replication studies, shown on the raw scale with logarithmic spacing to illustrate the wide range of journal influence represented. Because impact factors can vary by source, impact factors were prioritized from the journals’ official websites; when unavailable, values were obtained from Wikipedia, and studies for which no impact factor from either source ($N = 1$) could be identified were omitted. (d) Temporal distribution of the replication studies by publication period. These panels summarize the topical, publication, and temporal diversity of the replication studies.

1.4.1.3 Replication execution: prompt formulation and standardized model choice

For each study, we extracted the methodological elements needed to reproduce the primary analysis: cohort inclusion/exclusion rules, study period, predictor and outcome definitions, covariates, and statistical model type. These details were then translated into a comprehensive natural-language prompt that LATCH could execute end-to-end. Because many papers report multiple models with progressively richer covariate adjustment (univariate, demographic-adjusted, and more covariates adjusted), we standardized our primary replication target to the demographic-adjusted model (commonly age, sex, race/ethnicity). This choice was pragmatic and conservative: fully adjusted models often include variables whose definitions, transformations, or missing-data handling are not fully specified, increasing the risk of incorrect mapping and reducing interpretability of replication discordance. By focusing on demographic-adjusted models, we created a consistent basis for comparison and reduced the probability that replication failures were simply artifacts of under-described covariate construction. For three studies (Guo et al., 2022; Montgomery et al., 2017; J. Wang et al., 2020) in which a demographic only-adjusted model was not reported, we selected the most conservative available model—that is, the

least-adjusted specification reported—to minimize the introduction of poorly specified covariates and additional missingness.

1.4.1.4 Evaluation Metrics

Replication was evaluated as a structured comparison between each original publication and its independently re-executed analysis produced by LATCH. To capture both statistical and computational fidelity, we defined a set of complementary concordance metrics rather than relying on a single indicator of agreement. These include relative differences in cohort size and effect magnitude, as reported in Wang et al (S. V. Wang et al., 2022). First, cohort fidelity was assessed by comparing the sample size (N) reported in the publication with the cohort size generated by LATCH. In large observational datasets, discrepancies in cohort size often provide an early signal of divergence in inclusion or exclusion logic, variable definitions, or missing-data handling. Differences were therefore tracked both in absolute terms and relative magnitude, with particular attention to cases where mismatches exceeded approximately 10%.

Second, directional concordance was evaluated by determining whether the direction of association in the replicated analysis matched the published finding (e.g., risk-increasing versus risk-decreasing trends), with particular emphasis on ordered exposure categories such as tertiles or quartiles. Third, significance alignment assessed whether the original study and the LATCH replication agreed on statistical significance at the conventional threshold of $p < 0.05$ for the primary effect of interest. Fourth, confidence interval (CI) overlap was used to evaluate agreement in effect magnitude, focusing on whether the 95% CIs overlapped and to what extent. For consistency across studies, CI overlap was assessed using the most extreme exposure category (e.g., Q4 vs. Q1 or T3 vs. T1). Relative effect size represents the Odds Ratio or Hazard Ratio from the published study divided by the corresponding estimate obtained using LATCH.

These metrics were intentionally layered. A replicated result may agree in direction but differ in magnitude, or match statistical significance despite differences in cohort size or weighting choices. By jointly evaluating cohort fidelity, effect direction, statistical significance, CI overlap, and relative effect size the framework avoids over-interpreting replication success or failure based on a single statistic and provides a more nuanced assessment of agreement.

1.5 Evaluation of Replication Performance

1.5.1 Replication Outcomes

Across the benchmark set, LATCH demonstrated high overall concordance with published findings. Most replicated analyses matched the original studies in effect direction and statistical significance, and many showed substantial overlap in confidence intervals across logistic regression (Fig. 2a) and Cox proportional hazards models (Fig. 2b). LATCH successfully reproduced associations reported for diabetic retinopathy, diabetic kidney disease, mortality, and mental and cognitive health outcomes using diverse analytic approaches (Fig. 2).

Beyond regression analyses, LATCH accurately reproduced prevalence estimates (Fig. 2c) and group-based descriptive comparisons (Fig. 2d) and replicated linear regression coefficients (Fig. 2e) that clustered closely around the line of identity when compared with published estimates. Reconstructed cohort sizes were typically within 10% of those reported in the original publications, and several cohorts matched exactly. End-to-end replication analyses were completed efficiently, with observed runtimes on the order of minutes per study.

The three discrepancies shown in Fig. 2f, in terms of p value concordance or CI overlap was further discussed in 1.5.2 Discrepancy analysis. Together, these results indicate that LATCH can reliably translate published clinical methods into executable analyses that recover the core findings across a wide range of statistical models and outcome types.

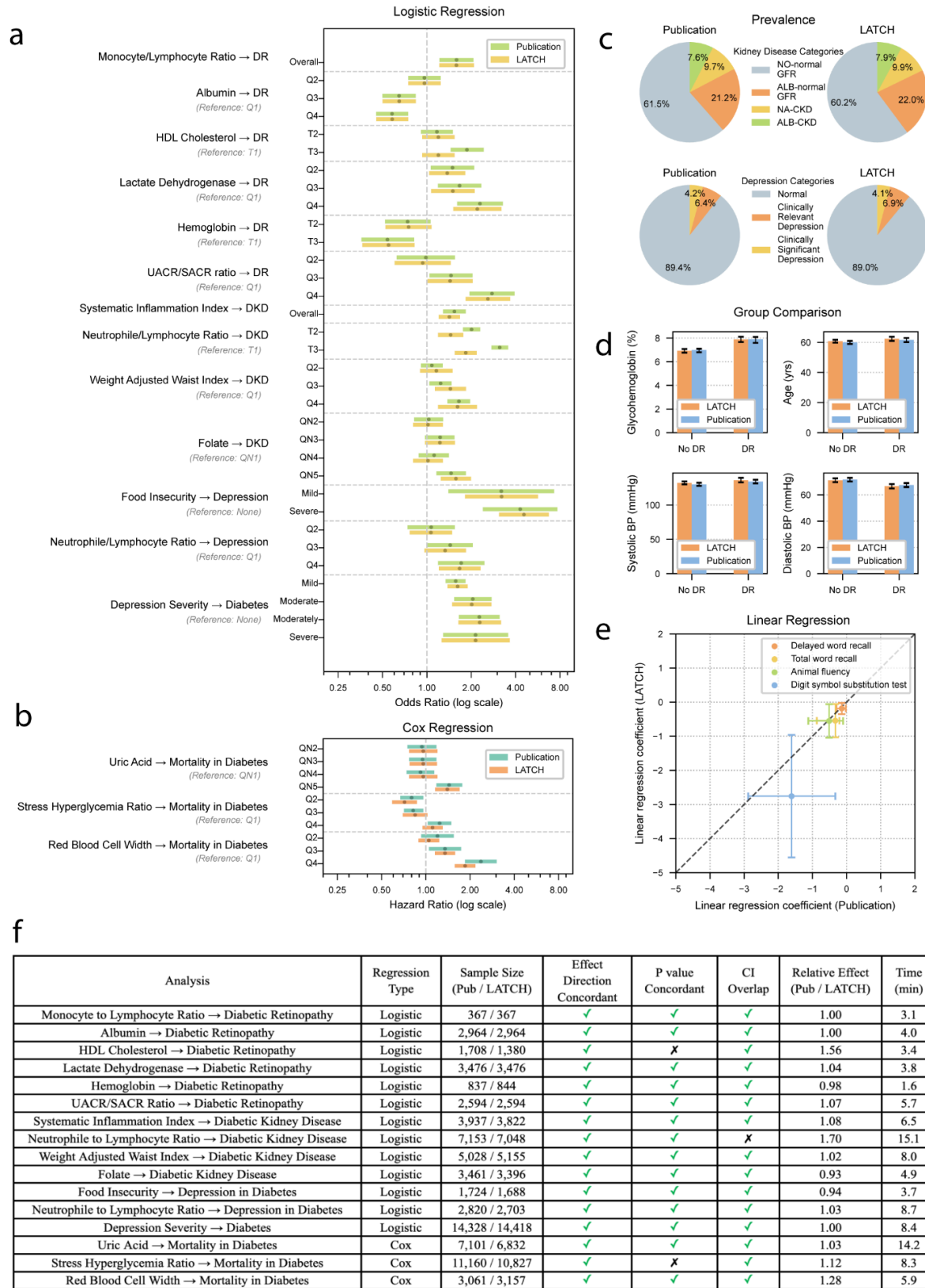


Figure 2. Reproduction of Results From Published Studies Using LATCH. (a) Comparison of logistic regression odds ratios between original publications and LATCH reproductions for various disease associations. Dots indicate point estimates, with

error bars representing 95% confidence intervals. (Abbreviations: UACR/SACR (Urinary Serum Albumin-to-Creatinine Ratio), QN: quintile, Q: quartile, T: tertile) (b) Comparison of Cox regression hazard ratios for mortality-related hypotheses. Dots indicate point estimates, with error bars representing 95% confidence intervals. (c) Replication of a prevalence analysis characterizing kidney disease phenotypes—NA-normal GFR (normal albuminuria normal glomerular filtration rate), ALB-normal GFR (albuminuria with normal glomerular filtration rate), NA-CKD (normal albuminuria chronic kidney disease defined by glomerular filtration rate) and ALB-CKD (albuminuria with chronic kidney disease)—and depression severity, stratified as clinically relevant and clinically significant depression. (d) Replication of group comparisons between patients with and without diabetic retinopathy. Error bars representing 95% confidence intervals. (e) Replication of linear regression coefficients for the effect of glycohemoglobin on cognitive scores as reported in a publication (x axis) versus those calculated by LATCH (y axis). Dots indicate point estimates, with error bars representing 95% confidence intervals. (f) Summary of the performance metrics for the reproduction of logistic and Cox regression studies from (a) and (b). Metrics were computed by focusing on the highest-risk categories, specifically the top quartile (Q4) and top tertile (T3) groups. The table shows concordance on the effect direction and statistical significance ($p < 0.05$), with processing times ranging from 1.6-15.1 minutes per query. Reported costs reflect estimated API usage for the Gemini 2.5 Flash model. Reported runtimes and costs are approximate and may vary with network conditions and token usage, which can differ slightly across runs even for identical queries. (The reference publications are indicated as superscripts in (f) under ‘Analysis’ column. These correspond to the studies reproduced in panel (a) and (b) for logistic and Cox regression. Prevalence analyses are shown in (c), group comparison in (d), linear regression in (e). Altogether, these analyses correspond to 20 publications: logistic regression ($n = 13$), Cox regression ($n = 3$), prevalence analyses ($n = 2$), group comparisons ($n = 1$), and linear regression ($n = 1$).)

1.5.2 Discrepancy Analysis: What Mismatches Revealed (Interpretive)

Although overall concordance was high, several instructive discrepancies were identified and systematically examined. In one case, an HDL cholesterol–retinopathy association showed p-value discordance, coinciding with the largest observed cohort size mismatch. This divergence was likely attributable to ambiguity in diabetes cohort definition logic involving “OR” inclusion criteria, which can be reasonably interpreted in multiple ways and applied differently across implementations.

In another case involving the neutrophil-to-lymphocyte ratio and diabetic kidney disease, both the original publication and the LATCH replication reported statistically significant associations, but confidence intervals did not overlap for upper tertiles. This pattern suggested sensitivity to how exclusion criteria—such as pregnancy or cancer status—were operationalized, where multiple plausible variable choices exist in NHANES. A third discrepancy was observed in the stress hyperglycemia ratio–mortality analysis, where significance alignment differed. NHANES does not contain a direct analogue to “admission glucose,” requiring an interpretive proxy, and the original publication did not specify which survey weights were used in the Cox model; different reasonable weight choices can yield different hazard ratio estimates.

These discrepancies underscore a central finding: published methods are often underspecified for exact computational replication, even by experienced analysts. The value of LATCH lies not in eliminating disagreement, but in making disagreement auditable. LATCH transforms replication into a structured diagnostic process that highlights where reporting standards limit reproducibility and where future studies can improve methodological transparency.

1.6 Summary and Implications

In summary, this work demonstrates that direct computational replication of published clinical studies can be performed at scale using LATCH. These results highlight both the feasibility of verifiable automation for clinical research replication and the need for more explicit methodological reporting to support reproducible science.

2. ENVIRONMENT

2.1 Activity Context

All activities described in this study were conducted as independent computational research using publicly available datasets, open-source software, and commercial application programming interfaces (APIs). The work focused on the design, implementation, and evaluation of an automated replication framework for clinical and

epidemiologic analyses. No private, restricted, or identifiable data were accessed, and no human subjects were involved.

2.2 Location, Timing, and Team Details

The research was conducted using a personal computing environment with locally executed data processing and analysis workflows. All datasets used in the study were publicly released and accessed remotely. The work was carried out over a defined development and evaluation period January - December 2025.

2.2.1 Team Work Details

Nayoon Gim, MD/PhD student (Bioengineering, University of Washington)

Responsibilities: Proposed and implemented the LATCH framework; prepared NHANES datasets across survey cycles; translated published study methods into executable natural-language prompts; executed end-to-end replication analyses; performed concordance and discrepancy evaluations against published results; and drafted technical documentation and manuscript/report sections.

In Gim, PhD student (Computer Science, Yale University)

Responsibilities: Provided technical guidance in the design of the LLM-integrated, agentic core system architecture; implemented system environment setup, including database configuration and R/Python package management; managed code repository and version control; and supported debugging, optimization, and code review. Contributed to improving system efficiency, scalability, and readiness for large-scale deployment.

Aaron Y. Lee, MD, MSCI, Professor and Chair, Department of Ophthalmology & Visual Sciences, Washington University School of Medicine in St. Louis

Responsibilities: Served as research advisor and supervisor; guided overall study objectives and replication strategy; advised on study selection and clinically meaningful outcomes; reviewed findings for clinical plausibility and methodological rigor; and provided input on interpretation of results and alignment with standards in clinical and epidemiologic research.

Cecilia S. Lee, MD, MS, Professor, Department of Ophthalmology & Visual Sciences, Washington University School of Medicine in St. Louis

Responsibilities: Served as research advisor and supervisor, especially on clinical epidemiology and population health methodology; guided interpretation of statistical results; reviewed analytic findings and discrepancy analyses; and contributed to evaluation, framing, and reporting of replication results from a clinical perspective.

2.3 Funding Support

No federal or institutional funding was received to support the activities described in this manuscript.

3. IMPACT

3.1 Replication Activity Impact

The primary impact of the replication activities described in this work is the demonstration that direct computational replication of complex biomedical studies can be executed rapidly, systematically, and transparently. Using LATCH, we reproduced the primary findings of 20 peer-reviewed NHANES-based diabetes studies. Replication outcomes were evaluated using predefined concordance metrics, including cohort fidelity, effect direction, statistical significance alignment, and confidence interval overlap. Across studies, LATCH achieved high concordance with published findings, with the majority of the replicated analyses matching the original results in direction and statistical significance and showing overlap in effect estimates.

Compared to traditional replication workflows, the automated approach reduced end-to-end replication execution time from weeks or months to minutes per analysis. This shift represents a meaningful change in what is

practically achievable: replication moves from a high-cost, one-off activity to a repeatable process that can be applied across multiple studies within a single research effort.

3.2 Long-Term Impact

The long-term impact of this replication strategy lies in its potential to normalize replication as a routine component of biomedical research rather than an exceptional undertaking. By lowering the technical and time barriers to replication, the framework enables systematic verification of published findings. Over time, this approach can support more reliable evidence synthesis, reduce propagation of fragile results, and improve confidence in observational findings that inform downstream research and clinical decision-making.

Because the strategy is based on public data, standardized statistical methods, and auditable analytic records, it also supports cumulative science: replicated analyses can be revisited, extended, or reinterpreted as new data or methods become available, without needing to reconstruct analytic pipelines from scratch.

3.3 Improvements to Research Rigor and Trust

The replication activities directly improve research rigor by enforcing explicit, executable representations of analytic decisions that are often underspecified in manuscripts. Each LATCH replication produces an Analytic Report that records the full analytic pathway, including cohort definitions, variable mappings, generated SQL, statistical model specifications, and software versions. This level of documentation exceeds typical reporting standards and makes analytic choices inspectable and auditable.

Importantly, discrepancies between original publications and replicated results are not treated as simple failures, but as diagnostic signals. In several cases, divergence revealed ambiguity in published methods—such as unclear inclusion logic, alternative variable operationalizations, or unspecified statistical analysis details—that would challenge any replicator. By surfacing these issues explicitly, the framework increases transparency and encourages clearer methodological reporting, thereby strengthening trust in both original and replicated research.

3.4 Dissemination and Adoption

The LATCH framework will be disseminated through open-source release of the codebase, academic presentations, and manuscript preparation. Early adopters include researchers in clinical epidemiology, informatics, and population health who face similar replication challenges with large public datasets. The framework is designed to be dataset-agnostic, lowering the barrier to adoption beyond NHANES.

4. RESPONSE TO BARRIERS

4.1 Barrier identification and Solutions

Methodological Underspecification in Published Studies: A primary barrier encountered was the frequent lack of precise, executable methodological detail in published biomedical studies. Key analytic components—such as inclusion and exclusion criteria, derived-variable formulas, missing-data handling, and survey-weighting choices—were often described narratively without sufficient specificity to enable direct computational reproduction. This barrier was addressed by treating replication as an interpretive and diagnostic process rather than a binary success/failure task. When ambiguity arose, reasonable interpretations were documented explicitly in the replication prompt and captured in the Analytic Report.

Sensitivity of Results to Analytic Choices: In some analyses, small differences in cohort construction or analytic parameters (e.g., survey weight selection in Cox models) led to meaningful changes in effect size or statistical significance. Rather than suppressing this variability, the framework was designed to surface it. The

Analytic Report records all modeling choices, making it possible to identify when discrepancies were driven by methodological ambiguity rather than errors in implementation.

4.2 Organizational Levels

Individual-Level Barriers: At the individual researcher level, replication is often constrained by time, effort, and the cognitive burden of translating prose-based methods into executable code. This barrier was addressed through automation: once the schema was prepared, replication workflows could be executed rapidly and repeatedly with minimal manual intervention, substantially lowering the barrier to routine replication.

Community-Level Barriers: At the broader research community level, replication is frequently deprioritized due to lack of incentives and standardized tools. By demonstrating that replication can be executed quickly, systematically, and with transparent audit trails using public data and widely accessible tools, this work addresses a key cultural barrier: the perception that replication is inherently slow, expensive, and difficult to sustain.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

Months 1–6: System design, data ingestion, schema preparation, and initial LATCH prototype development.

Months 7–12: Safeguard implementation, analytic report development, and pilot replications, selection and replication of 20 benchmark studies and concordance analysis, root-cause analyses, refinement, dissemination, and sustainability planning

5.2 Associated Costs

Major cost categories included personnel, secure computing infrastructure, and LLM API usage. Relative to traditional replication labor costs, the marginal cost per replication analysis was significantly reduced.

5.3 Required Resources

Key resources included secure servers, relational databases, statistical software (R), Python orchestration tools, and controlled access to LLM APIs.

6. COLLABORATION

6.1 Necessary Partnerships

For broader deployment and real-world impact, future extensions of this framework would benefit from partnerships with organizations that steward standardized clinical data infrastructures. In particular, collaboration with the Observational Health Data Sciences and Informatics (OHDSI) community ([OHDSI – Observational Health Data Sciences and Informatics, n.d.](#)) would be essential to adapt and validate LATCH across OMOP common data model–based data marts.

6.2 Leveraging Existing Partnerships

This project leveraged existing academic mentorship and technical collaboration within the research team. Clinical and methodological guidance was provided by faculty supervisors, while technical system design and software engineering advice were contributed by internal collaborators.

6.3 Building New Partnerships

Although no new external collaborations were formally established during this project, future work could involve partnerships with organizations such as the National Library of Medicine (NLM) and PubMed Central (PMC). These collaborations could enable automated ingestion and testing of open-access articles to scale replication across the biomedical literature and to develop a reproducibility “report card” for studies using public datasets.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Sustainability

The replication activities in this work are designed to be sustainable by construction. LATCH relies on publicly available datasets, open-source software, and commercial APIs, avoiding dependence on proprietary data or specialized infrastructure. Its modular architecture allows components for schema preparation, interpretation, cohort extraction, and statistical analysis to be maintained or replaced independently as technologies evolve. Because workflows are fully automated and executed on demand, replication does not require sustained personnel effort or long-term project funding.

LATCH is API-driven and vendor-agnostic with respect to large language models, enabling straightforward upgrades as LLM capabilities advance without changes to the surrounding system. This contrasts with fine-tuned or tightly coupled models, which are costly to retrain and harder to maintain. Together, these design choices minimize technical debt and support long-term reuse as a durable infrastructure for scalable, automated replication in computational clinical research.

7.2 Transferability and Generalizability

The replication strategy is intentionally dataset-agnostic and transferable across research domains that rely on structured, tabular data accompanied by data dictionaries or codebooks. While demonstrated using NHANES, the same workflow can be applied to other epidemiologic surveys, clinical registries, administrative claims datasets, and standardized EHR-derived tables. The schema-only interaction model allows the framework to be adapted to new datasets through a one-time schema preparation step, without modifying the core logic of the system.

Generalizability extends beyond dataset type to analytic scope. LATCH supports a range of commonly used epidemiologic and clinical statistical methods, including regression models, survival analyses, prevalence estimation, and descriptive statistics. As a result, the replication approach is applicable to basic, translational, clinical, and population health research contexts where reproducibility challenges arise from complex cohort definitions and under-specified analytic pipelines.

7.3 Scaling and Applications

LATCH is inherently scalable because replication execution time scales with analytic complexity rather than personnel availability. Once a dataset schema has been prepared, multiple replication analyses can be executed sequentially or in parallel with minimal incremental effort. This makes the framework suitable for large-scale replication initiatives, such as systematically auditing findings within a research domain, performing routine robustness checks during peer review, or supporting replication-focused training exercises.

In different contexts, LATCH can be deployed as a standalone research tool, integrated into journal or consortium-led replication efforts, or used as an internal quality-control mechanism within research groups.

7.4 Future Integration and Dissemination Plans

Future integration efforts will focus on expanding the range of supported statistical methods and improving interoperability with additional data standards and repositories. Planned enhancements include support for additional derived-variable templates, more granular data engineering strategies, and expanded logging of analytic assumptions to further strengthen auditability. Dissemination plans include releasing the framework as an openly accessible codebase with examples.

Beyond software dissemination, the replication strategy lends itself to broader integration into research practice through training materials, tutorials, and case studies that demonstrate how replication can be incorporated into routine analysis workflows.

7.5 LLM disclosures

Our proposed method of replication, LATCH (LLM-Assisted Testing of Clinical Hypotheses) was designed to have LLM integrated steps that utilized Gemini 2.5 Flash. Additionally, the authors used ChatGPT for grammar review/refinement to improve the readability of the report. Following the use of this tool, the authors reviewed and edited the content as needed and took full responsibility for the accuracy, integrity, and final approval of the work.

8. REFERENCES

1. Al-Kindi, S. G., Refaat, M., Jayyousi, A., Asaad, N., Al Suwaidi, J., & Abi Khalil, C. (2017). Red cell distribution width is associated with all-cause and cardiovascular mortality in patients with diabetes. *BioMed Research International*, 2017, 5843702.
2. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454.
3. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624(7992), 570–578.
4. Cai, X., Geng, Y., Du, Y., Westerman, B., Wang, D., Ma, C., & Vallejo, J. J. G. (2025). Utilizing Large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *BMC Medical Research Methodology*, 25(1), 116.
5. Casagrande, S. S., Lee, C., Stoeckel, L. E., Menke, A., & Cowie, C. C. (2021). Cognitive function among older adults with diabetes and prediabetes, NHANES 2011-2014. *Diabetes Research and Clinical Practice*, 178(108939), 108939.
6. Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., & Moher, D. (2024). Biomedical researchers' perspectives on the reproducibility of research. *PLoS Biology*, 22(11), e3002870.
7. Dai, H., Liu, L., & Xu, W. (2025). Association of albumin-to-creatinine ratio with diabetic retinopathy among US adults (NHANES 2009-2016). *Endocrinology, Diabetes & Metabolism*, 8(1), e70029.
8. Dennstädt, F., Hastings, J., Putora, P. M., Schmerder, M., & Cihoric, N. (2025). Implementing large language models in healthcare while balancing control, collaboration, costs and security. *Npj Digital Medicine*, 8(1), 143.
9. Ding, L., Zhang, H., Dai, C., Zhang, A., Yu, F., Mi, L., Qi, Y., & Tang, M. (2024). The prognostic value of the stress hyperglycemia ratio for all-cause and cardiovascular mortality in patients with diabetes or prediabetes: insights from NHANES 2005-2018. *Cardiovascular Diabetology*, 23(1), 84.
10. Guo, W., Song, Y., Sun, Y., Du, H., Cai, Y., You, Q., Fu, H., & Shao, L. (2022). Systemic immune-inflammation index is associated with diabetic kidney disease in Type 2 diabetes mellitus patients: Evidence from NHANES 2011-2018. *Frontiers in Endocrinology*, 13, 1071465.
11. Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 20(1), 117–121.
12. Jonnagaddala, J., & Wong, Z. S.-Y. (2025). Privacy preserving strategies for electronic health records in the era of large language models. *Npj Digital Medicine*, 8(1), 34.

13. Kahn, M. G., & Weng, C. (2012). Clinical research informatics: a conceptual perspective. *Journal of the American Medical Informatics Association: JAMIA*, 19(e1), e36–e42.
14. Lee, J., Kim, K. H., Ahn, J. C., Kim, J. A., Lee, G., Son, J. S., Choi, S. J., Oh, Y. H., & Park, S. M. (2021). Prevalence, awareness, treatment, and control of diabetes mellitus by depressive symptom severity: a cross-sectional analysis of NHANES 2011–2016. *BMJ Open Diabetes Research & Care*, 9(1), e002268.
15. Liang, Y., Ding, L., Tao, M., & Zhu, Y. (2024). The association of metabolic profile of folate with diabetic kidney disease: evidence from 2011–2020 cycles of the NHANES. *Renal Failure*, 46(2), 2420830.
16. Li, B., Chen, L., Hu, X., Tan, T., Yang, J., Bao, W., & Rong, S. (2023). Association of serum uric acid with all-cause and cardiovascular mortality in diabetes. *Diabetes Care*, 46(2), 425–433.
17. Li, X., & Chen, M. (2023). Correlation of hemoglobin levels with diabetic retinopathy in US adults aged ≥40 years: the NHANES 2005–2008. *Frontiers in Endocrinology*, 14, 1195647.
18. Li, X., Wang, L., Liu, M., Zhou, H., & Xu, H. (2023). Association between neutrophil-to-lymphocyte ratio and diabetic kidney disease in type 2 diabetes mellitus patients: a cross-sectional study. *Frontiers in Endocrinology*, 14, 1285509.
19. Montgomery, J., Lu, J., Ratliff, S., & Mezuk, B. (2017). Food insecurity and depression among adults with diabetes: Results from the National Health and Nutrition Examination Survey (NHANES). *The Diabetes Educator*, 43(3), 260–271.
20. Mottl, A. K., Kwon, K.-S., Mauer, M., Mayer-Davis, E. J., Hogan, S. L., & Kshirsagar, A. V. (2013). Normoalbuminuric diabetic kidney disease in the U.S. population. *Journal of Diabetes and Its Complications*, 27(2), 123–127.
21. *NHANES Questionnaires, Datasets, and Related Documentation*. (n.d.). Retrieved November 23, 2025, from <https://wwwn.cdc.gov/nchs/nhanes/>
22. Niven, D. J., McCormick, T. J., Straus, S. E., Hemmelgarn, B. R., Jeffs, L., Barnes, T. R. M., & Stelfox, H. T. (2018). Reproducibility of clinical research in critical care: a scoping review. *BMC Medicine*, 16(1), 26.
23. *OHDSI – Observational Health Data Sciences and Informatics*. (n.d.). Retrieved December 19, 2025, from <https://www.ohdsi.org/>
24. Oniani, D., Hilsman, J., Zang, C., Wang, J., Cai, L., Zawala, J., & Wang, Y. (2024). Emerging opportunities of using large language models for translation between drug molecules and indications. *Scientific Reports*, 14(1), 10738.
25. Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, N.Y.)*, 334(6060), 1226–1227.
26. Richesson, R. L., Hammond, W. E., Nahm, M., Wixted, D., Simon, G. E., Robinson, J. G., Bauck, A. E., Cifelli, D., Smerek, M. M., Dickerson, J., Laws, R. L., Madigan, R. A., Rusincovitch, S. A., Kluchar, C., & Califf, R. M. (2013). Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association: JAMIA*, 20(e2), e226–e231.
27. Wang, G.-X., Fang, Z.-B., Li, J.-T., Huang, B.-L., Liu, D.-L., Chu, S.-F., & Li, H.-L. (2022). The correlation between serum albumin and diabetic retinopathy among people with type 2 diabetes mellitus: NHANES 2011–2020. *PloS One*, 17(6), e0270019.
28. Wang, H., Guo, Z., & Xu, Y. (2022). Association of monocyte-lymphocyte ratio and proliferative diabetic retinopathy in the U.S. population with type 2 diabetes. *Journal of Translational Medicine*, 20(1), 219.
29. Wang, J., Zhou, D., & Li, X. (2020). The association between neutrophil-to-lymphocyte ratio and diabetic depression in U.S. adults with diabetes: Findings from the 2009–2016 National Health and Nutrition Examination Survey (NHANES). *BioMed Research International*, 2020, 8297628.
30. Wang, S. V., Sreedhara, S. K., Schneeweiss, S., & REPEAT Initiative. (2022). Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nature Communications*, 13(1), 5126.
31. Wang, Y., Lopez, J. M. S., Bolge, S. C., Zhu, V. J., & Stang, P. E. (2016). Depression among people with type 2 diabetes mellitus, US National Health and Nutrition Examination Survey (NHANES), 2005–2012. *BMC Psychiatry*, 16(1), 88.
32. Wang, Z., Shao, X., Xu, W., Xue, B., Zhong, S., & Yang, Q. (2024). The relationship between weight-adjusted-waist index and diabetic kidney disease in patients with type 2 diabetes mellitus. *Frontiers in Endocrinology*, 15, 1345411.
33. Wiest, I. C., Leßmann, M.-E., Wolf, F., Ferber, D., Van Treeck, M., Zhu, J., Ebert, M. P., Westphalen, C. B., Wermke, M., & Kather, J. N. (2025). Deidentifying medical documents with local, privacy-preserving large language models: The LLM-anonymizer. *NEJM AI*, 2(4). <https://doi.org/10.1056/aidbp2400537>
34. Xu, W., Xu, X., Zhang, M., & Sun, C. (2024). Association between HDL cholesterol with diabetic retinopathy in diabetic patients: a cross-sectional retrospective study. *BMC Endocrine Disorders*, 24(1), 65.

35. Yang, P., Xu, W., Liu, L., & Yang, G. (2024). Association of lactate dehydrogenase and diabetic retinopathy in US adults with diabetes mellitus. *Journal of Diabetes*, 16(1), e13476.
36. Zhang, X., Saaddine, J. B., Chou, C.-F., Cotch, M. F., Cheng, Y. J., Geiss, L. S., Gregg, E. W., Albright, A. L., Klein, B. E. K., & Klein, R. (2010). Prevalence of diabetic retinopathy in the United States, 2005-2008. *JAMA: The Journal of the American Medical Association*, 304(6), 649–656.

INDIVIDUAL WINNER

Magdalena Kasendra

Leads organoid and organ-on-chip research at Cincinnati Children's, advancing standardized, reproducible platforms for drug discovery and toxicology testing.

Dr. Magdalena Kasendra is Director of Research & Development at the Center for Stem Cell and Organoid Medicine (CuSTOM) at Cincinnati Children's Hospital Medical Center. She is an internationally recognized leader in organoid biology, microphysiological systems (MPS), and New Approach Methodologies (NAMs), focused on advancing human-relevant platforms for drug discovery, predictive toxicology, and regenerative medicine.

She leads multidisciplinary programs spanning regenerative medicine, organoid-based therapeutics, and advanced in vitro systems, and directs the CuSTOM Accelerator - an innovation hub translating organoid and organ-on-chip technologies into standardized, scalable, and deployable platforms for research, industry, and clinical application. Her work integrates discovery science with automation-enabled development, regulatory alignment, and clinical translation to deliver decision-ready human models.

Dr. Kasendra's career bridges academia and industry, with training at Novartis Vaccines, Harvard Medical School, and the Wyss Institute for Biologically Inspired Engineering, and leadership experience at Emulate Inc. She serves on the Executive Board of the International Microphysiological Systems Society and contributes to global efforts advancing stem cell-based models and next-generation non-animal methodologies.

1. REPLICATION STRATEGY**1.1 Strategy Description**

Shinozawa et al. (Gastroenterology, 2021) introduced a high-fidelity human liver organoid (HLO) platform derived from human pluripotent stem cells (hPSCs) and demonstrated its ability to predict clinically relevant drug-induced liver injury (DILI) across a large compound set (Shinozawa et al., 2021). The study showed that Matrigel-embedded HLOs recapitulate key hepatic cell types, exhibit functional liver outputs, and reproduce compound-specific toxicity phenotypes - including intrinsic hepatotoxicity, mitochondrial toxicity, and cholestatic injury - with improved concordance to clinical outcomes compared to conventional 2D hepatocyte systems.

Our replication strategy aimed to independently reproduce the core biological features and toxicological response phenotypes reported in the Shinozawa HLO-DILI assay, while translating the workflow into standardized, automation-ready, and translationally deployable platform, within CuSTOM Accelerator (Fig. 1).

The CuSTOM Ecosystem: Bridging Discovery to Application

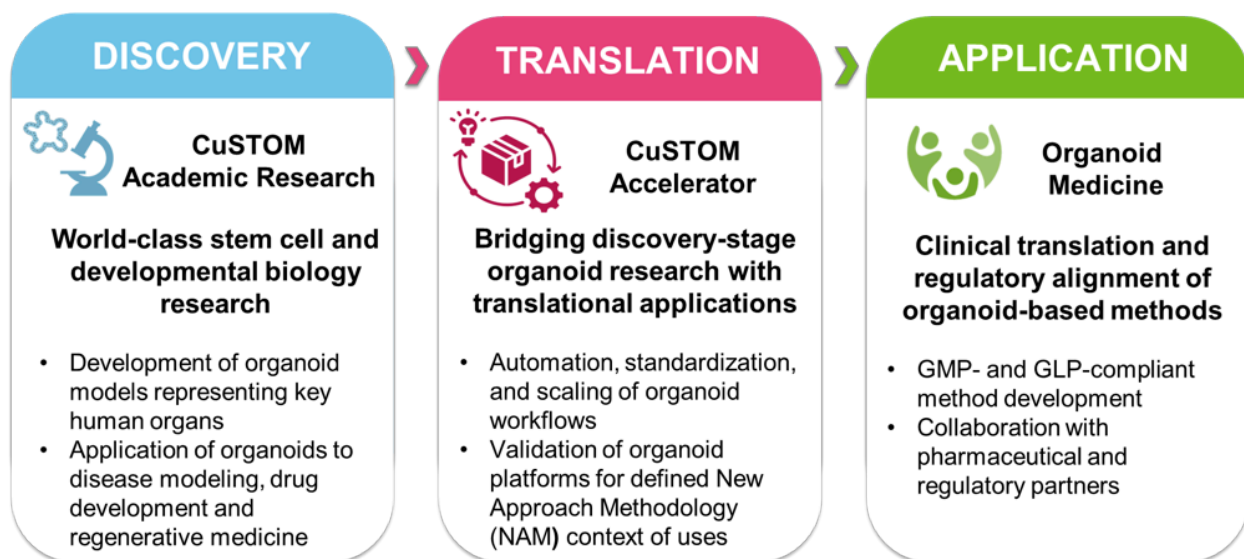


Figure 1. CuSTOM framework for discovery-to-application translation of organoid assays. The CuSTOM ecosystem supports systematic development, replication, and scaling of organoid platforms by integrating biological discovery, workflow standardization and automation, and application-focused validation within clearly defined contexts of use.

Rather than treating replication as a single repetition of experimental protocol, we implemented a phased replication program (see Timeline & Budget) aligned with the Center for Stem Cell & Organoid Medicine's (CuSTOM) mission to translate organoid technologies from discovery-stage biology to application-ready tools with defined contexts of use.

Replication activities included: (1) reconstructing the published HLO differentiation and phenotypic hepatotoxicity assessment workflow; (2) replacing spontaneous organoid formation and Matrigel embedding with standardized microcavity-guided aggregation under matrix-free conditions to reduce stochastic variability and improve automation compatibility; (3) implementing a robotic, variance-controlled execution pipeline for organoid seeding, feeding, compound dosing and imaging; and (4) evaluating robustness across multiple genetically distinct hPSC donor lines and repeated production batches while preserving the mechanistic injury modalities reported in the original study.

The primary objective of this replication effort was to determine whether the biological findings of the Shinozawa HLO-DILI assay could be reproduced with sufficient rigor, robustness, and operational control to support sustained translational use, rather than remaining confined to a technically demanding, expert-dependent research setting.

1.2 Outcome Evaluation and Innovative Methods

Replication outcomes were evaluated using orthogonal functional and phenotypic readouts, including ATP-based viability, high-content imaging-derived morphological profiling, mitochondrial membrane potential (TMRM), bile acid transport (CLF), and IC₅₀ benchmarking relative to primary human hepatocyte assays. A central feature of the strategy was rigorous evaluation of replicability across multiple donor lines and organoid batches (including batches generated from fresh or cryopreserved precursors). Across these iterations, we quantified variance using integrated QC metrics, standardized image-based morphology features, automated IC₅₀ extraction/benchmarking, and artificial intelligence (AI)-enabled phenotypic clustering to distinguish biological signal from technical variability. This data-driven evaluation framework was designed to determine not only whether the original findings could be reproduced, but whether they could be reproduced with the robustness, stability, and precision required to justify downstream translational deployment.

1.3 Working Definition of Replication

Given the inherent variability in defining replication for complex biological systems, we adopted a biological rather than procedural definition of faithful replication. We defined faithful replication as the ability of the CuSTOM Accelerator workflow to reproduce the core biological behaviors and toxicological responses reported by Shinozawa et al., rather than as strict adherence to specific assay conditions (e.g., Matrigel-embedded organoids) or manual execution steps (Fig 2A).

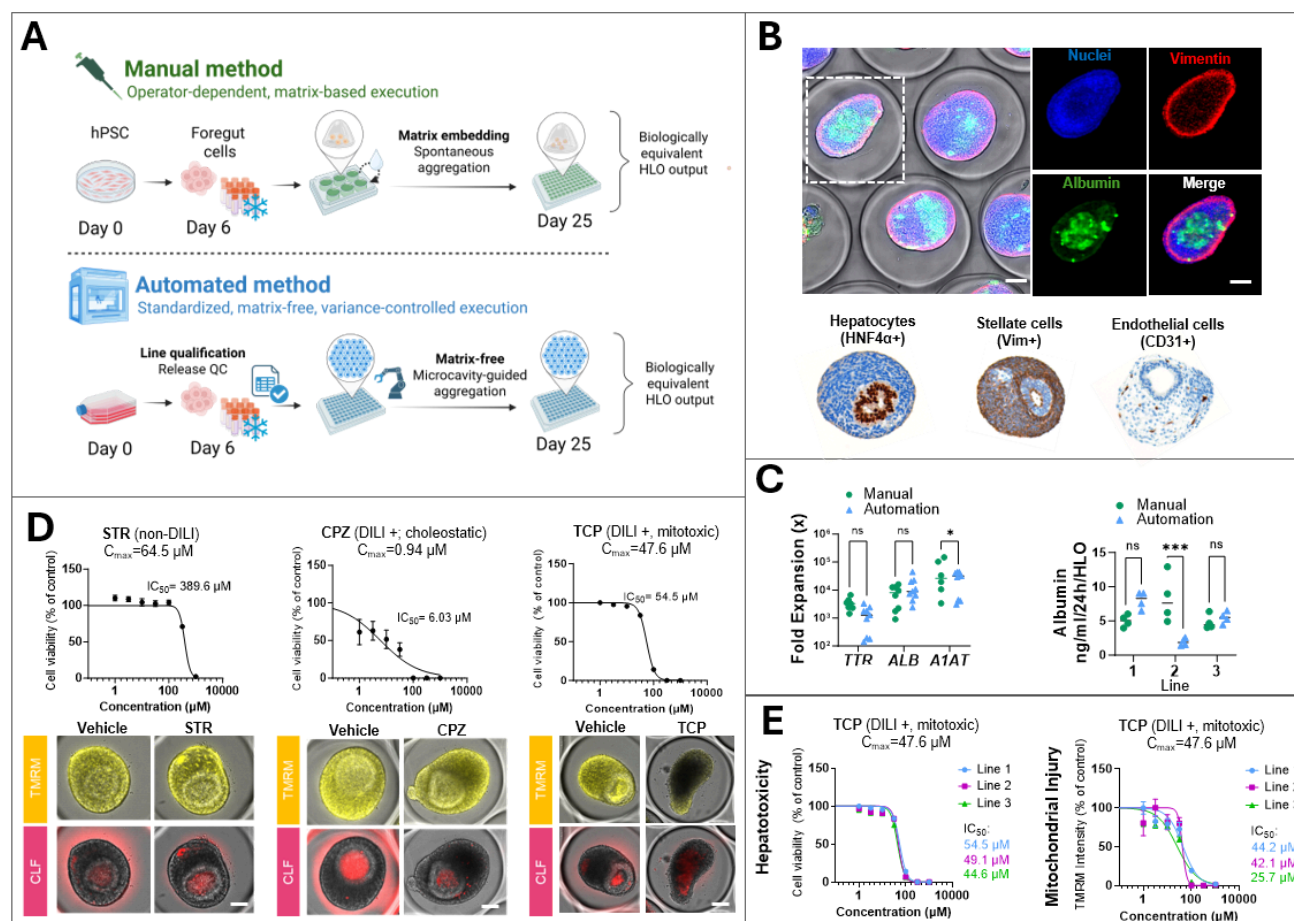


Figure 2. Faithful biological replication of the Shinozawa HLO-DILI assay under a standardized, matrix-free, automation-compatible workflow.

(A) Conceptual comparison of the original Shinozawa HLO-DILI workflow and the CuSTOM Accelerator implementation. Replication was defined biologically rather than procedurally: the goal was to preserve the core HLO tissue behaviors and toxicological response phenotypes of the original assay while replacing operator-dependent, matrix-based steps with standardized, matrix-free, automation-compatible processes.

(B) Representative images demonstrating preservation of HLO tissue architecture, hepatic identity, and supporting non-parenchymal cell populations in matrix-free organoids generated using the CuSTOM Accelerator workflow. Scale bar, 100 μm .

(C) Hepatic maturation and baseline function were confirmed by expression of key liver-associated genes and sustained albumin secretion under vehicle conditions.

(D) The replicated workflow reproduced compound-specific hepatotoxicity responses, distinguishing the DILI-negative control streptomycin (STR) from the DILI-positive compounds chlorpromazine (CPZ) and tolcapone (TCP), with preservation of mechanism-associated injury phenotypes, including cholestatic injury with CPZ and mitochondrial dysfunction with TCP. Scale bar, 100 μm .

(E) Extension across multiple independent hPSC donor lines demonstrated that biologically faithful replication was maintained across genetic backgrounds, with conserved quantitative toxicity responses, stable DILI classifications, and reproducible early mitotoxic injury signatures. Selected graphs and representative images were adapted from a previously published study (Soussi et al., 2025).

In the original study, HLOs generated from a single hPSC line were shown to: (i) self-organize into multicellular hepatic tissue, (ii) include key hepatic and supporting non-parenchymal cell populations, (iii) exhibit baseline functional liver outputs, and (iv) respond to reference compounds with mechanistically distinct injury phenotypes concordant with clinical outcomes. Accordingly, our replication criteria centered on reproducing these foundational attributes.

Within the CuSTOM Accelerator, matrix-free HLOs recapitulated expected cellular composition and tissue architecture, including critical parenchymal and supporting non-parenchymal populations, with marker expression and transcriptional profiles consistent with published HLO characterizations (Fig. 2B-C). Hepatic identity and maturation were confirmed by robust expression of key liver genes (TTR, A1AT, ALB) and appropriate localization of canonical hepatic markers. Baseline functional performance - including albumin secretion, metabolic competence, and stability under vehicle conditions - was comparable to that reported in the Shinozawa study and other established HLO platforms (Fig. 2C).

Crucially, we reproduced the compound-specific injury phenotypes that constitute the central biological claims of the assay. Canonical intrinsic hepatotoxicants (e.g., chlorpromazine and tolcapone) elicited reproducible, dose-dependent hepatocellular injury, while non-DILI control compounds (e.g., streptomycin) produced no detectable toxicity at physiologically relevant concentrations (Fig. 2D; top). Mechanism-specific injury signatures were preserved, including mitochondrial dysfunction in response to tolcapone and cholestatic injury following chlorpromazine exposure (Fig. 2D; bottom).

While the original Shinozawa study utilized a single hPSC line, extension to multiple independent donor lines was undertaken only after establishing faithful biological replication, in order to assess robustness and generalizability under standardized conditions. Quantitative hepatotoxicity responses, including stable IC_{50} values, and the conserved early emergence of mitotoxic/cholestatic injury signals were preserved across donors, providing additional evidence that the replicated phenotypes are driven by compound-specific mechanisms of action rather than donor-specific effects (Fig. 2E).

To ensure that biological replication could be evaluated independently of technical variability, we implemented a prospective, multi-stage quality control (QC) framework spanning hPSC qualification, organoid precursor generation, organoid manufacturing, and assay execution (Fig. 3). hPSC lines were qualified using International Society for Stem Cell Research (ISSCR)-aligned best practices (pluripotency, genomic integrity, and identity). Lineage-committed precursors were characterized prior to cryopreservation to confirm differentiation status, composition and viability. During organoid manufacturing, automated imaging metrics and predefined acceptance thresholds (e.g., microcavity occupancy and morphological uniformity) ensured consistency of starting material. Each assay run included QC controls for baseline stability and reference compound responses. Together, this multi-stage QC strategy enabled reliable discrimination between biological signal and technical noise and underpinned reproducible replication under automated, matrix-free conditions.

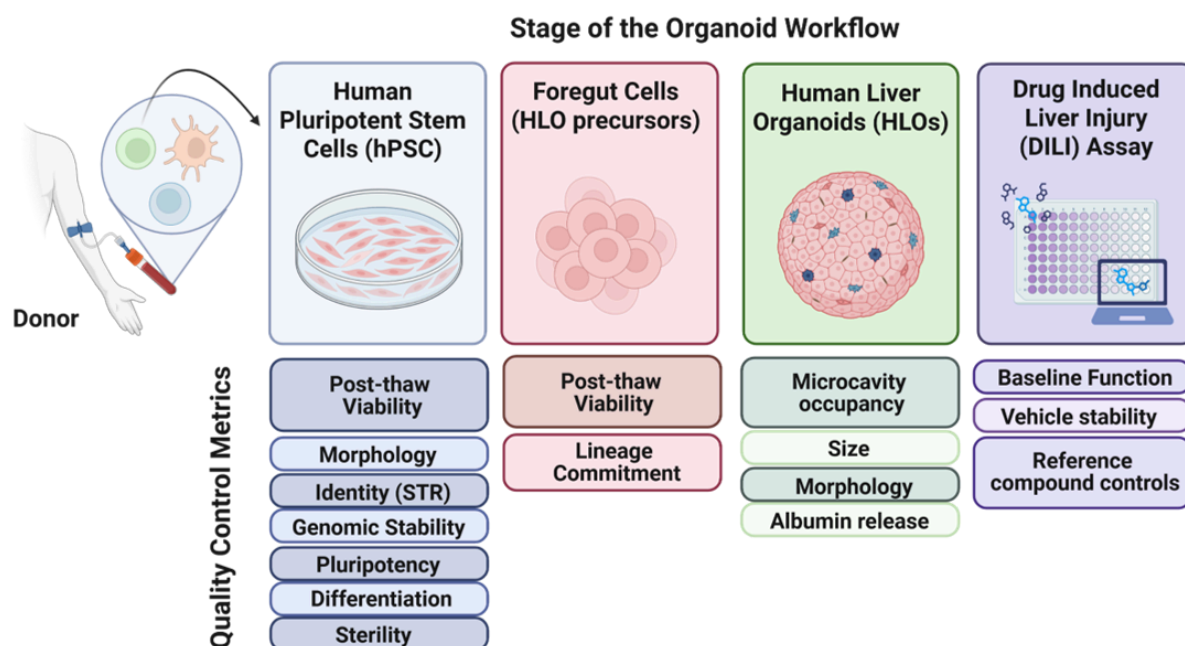


Figure 3. Integrated multi-stage QC framework enabling reproducible and translational replication of the Shinozawa HLO-DILI assay. Diagram illustrating the end-to-end quality control strategy implemented in the CuSTOM Accelerator, encompassing iPSC qualification, organoid precursor characterization, standardized organoid manufacturing, and assay-level controls. Quantitative and image-based acceptance thresholds are applied prospectively at each stage to ensure assay readiness, minimize technical variability, and support consistent reproduction of compound-specific hepatotoxicity phenotypes across donor lines and experimental runs.

Within this framework, the Shinozawa replication progressed from manual rebuilding to standardized, automated execution and ultimately to technology transfer, illustrating how a biologically faithful replication effort can evolve into translational replication - where rigorously reproduced workflows become scalable, transferable, and accessible for broader adoption.

1.4 Category Identification

The replication activities described here fall primarily within the Translational and Pre-clinical categories.

- Translational: transformation of a discovery-stage organoid assay into a standardized, automated, and scalable platform with defined QC gates, reproducibility metrics, and deployment readiness.
- Pre-clinical replication: faithful reproduction of compound-specific hepatotoxicity phenotypes, mechanistic injury signatures, and IC₅₀ benchmarks relevant to drug safety assessment.

Although grounded in foundational stem cell biology, the primary objective of this strategy was not basic replication of developmental processes, but rather the rigorous validation and operationalization of a human-relevant toxicity assay suitable for multi-donor evaluation, inter-site transfer, and translational use in drug discovery and safety pharmacology.

2. ENVIRONMENT

2.1 Activity Context

Replication activities were conducted December 2023 through December 2025 within the CuSTOM Accelerator at Cincinnati Children's Hospital Medical Center, a purpose-built translational organoid-engineering environment designed to transform complex stem-cell-derived systems into standardized, automated, and reproducible platforms. The Accelerator integrates organoid biology, automation engineering, high-content imaging, AI-enabled analytics, and quality systems within a coordinated operational framework. This unique configuration was essential for faithful replication of Shinozawa assay and systematic removal of implementation barriers that otherwise limit reproducibility and adoption.

We selected the Shinozawa HLO-DILI assay as a replication target because it is a seminal and widely cited demonstration of iPSC-derived liver organoids predicting clinically relevant hepatotoxicity (Shinozawa et al., 2021), yet, at the outset of this project, no documented automation-ready, cross-donor, or industry-driven replication or adoption had been reported. Several features of the original workflow posed substantial barriers to broader adoption, particularly in pharmaceutical and translational settings: labor-intensive differentiation requiring expert judgment of organoid readiness; Matrigel-associated variability and imaging challenges (broad Z-range, prolonged acquisition, operator dependency); and reliance on terminal endpoints that limit throughput and constrain multi-donor or larger compound-panel studies.

2.2 Location, Timing, and Team Details

All work occurred within the Accelerator's dedicated infrastructure, including an automation suite, organoid-culture laboratories, and downstream analytical assay facilities designed for environmental control, workflow harmonization, and operator-independent reproducibility. Activities progressed through staged implementation across the two-year period: protocol reconstruction and benchmarking; automation adaptation; iterative optimization and QC gate definition; multi-donor and multi-batch replication cycles, and incorporation into internal and external translational programs. Throughout, the Accelerator collaborated closely with the Takebe Lab (the inventor laboratory and one of CuSTOM's academic research groups) to ensure seamless adoption of the original organoid differentiation and assessment methods while transitioning the workflow into an optimized, de-risked, and automation-compatible platform. This continuity between discovery science and translational engineering was

essential for preserving biological fidelity while enabling the assay to mature into a scalable and widely deployable tool.

The effort was driven by an interdisciplinary team under the scientific and programmatic leadership of Magdalena Kasendra, PhD, who conceived and directed the replication strategy, oversaw platform integration within the CuSTOM Accelerator, and coordinated execution across biological, engineering, and data-science components. Takanori Takebe, MD, PhD, contributed foundational biological expertise and ensured continuity with the original human liver organoid (HLO) development, providing guidance on biological fidelity relative to prior HLO studies.

The biology and organoid-manufacturing team, led by Fadoua El Abdellaoui Soussi, PhD, executed differentiation, biological optimization, and reproducibility testing, supported by Clark Bacon for organoid production, assay establishment, and validation. Quality control activities were led by Alok Verma and Yemisi Campbell, who implemented and monitored morphological and functional performance metrics and ensured consistency across batches and donor lines. Automation, engineering, and data-science integration were led by Sina Dadgar, who adapted the workflow to the scaled, robotic platform, established automated imaging and analysis pipelines, and implemented AI-driven toxicity metrics, with support from Aparna Trivedi and technical experts from the Molecular Devices. Operational coordination, procurement, and resource management were supported by CuSTOM administrative staff.

2.3 Funding Support

Replication was supported through a combination of institutional investment in the CuSTOM Accelerator (core staffing, infrastructure, and translational assay development), the Danaher Beacon Initiative (automation engineering, AI support and project consumables) (Research Horizons, 2024), and philanthropic resources supporting establishment of QC-controlled donor line panels representing population-level heterogeneity. Collaborative programs with partners such as Roche leveraged and validated the replicated workflow across additional translational contexts such as immune-driven DILI (Research Horizons, 2025a). Together, these funding sources enabled replication to be conducted at scale and to evolve into a sustained component of CuSTOM's translational operations. The standardized, automation-ready HLO workflows established through this replication effort will be leveraged as part of the federally supported ARPA-H CATALYST (DATAMAP) initiative to support generation of large-scale, high-quality datasets for training AI-driven drug safety prediction models (Research Horizons, 2025b).

3. IMPACT

3.1 Replication Activity Impact (Pre/Post Metrics)

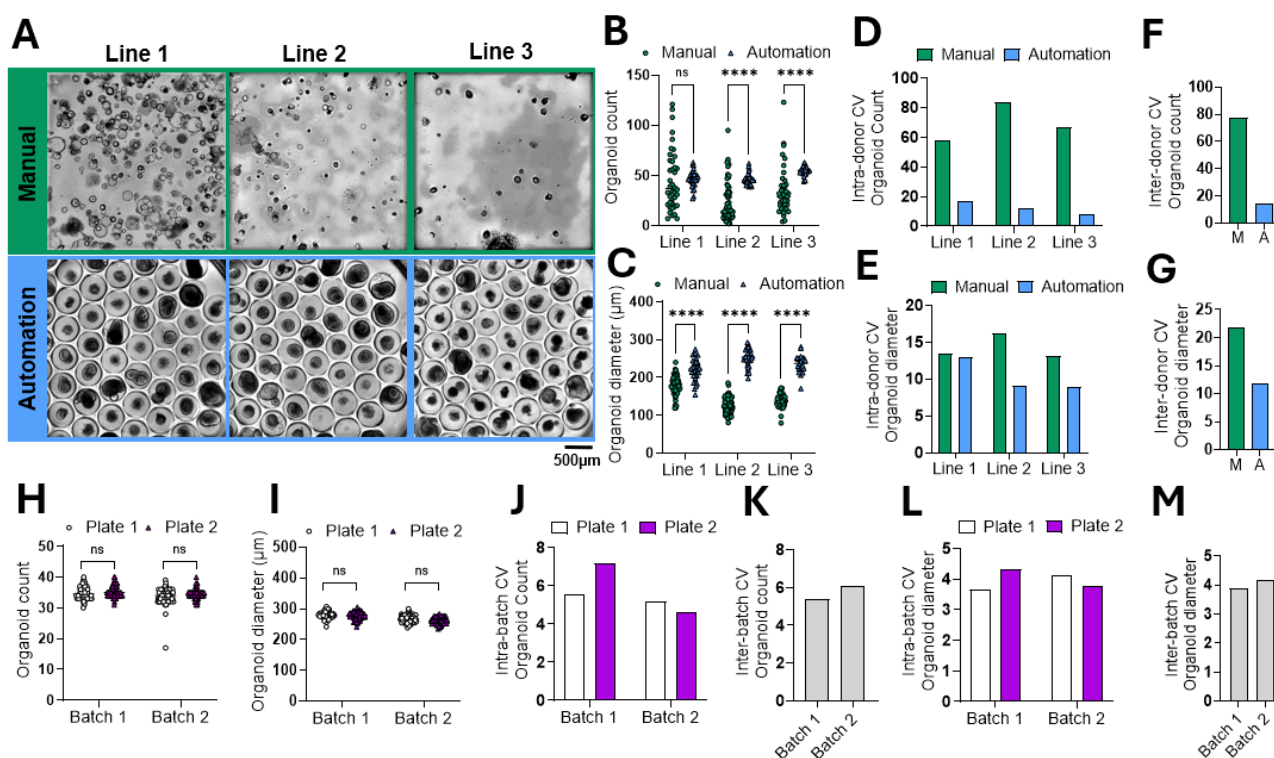
The replication of the Shinozawa HLO-DILI assay produced scientific, technical, and institutional impacts that extended far beyond reproducing a published study, establishing a reproducible, operationally stable, and scalable framework for sustained replication in practice. A comparative matrix summarizing key implementation differences and replication-enabling improvements is provided in Figure 4. By embedding the workflow within the CuSTOM Accelerator's automation-enabled, QC-controlled environment, we confirmed that the hepatotoxicity signatures described by Shinozawa et al. could be faithfully recapitulated (Fig. 2) while substantially improving reproducibility, variance control, and scalability (Fig. 4).

	Original Shinozawa	CuSTOM Accelerator Replication
Organoid formation	Spontaneous aggregation	Microcavity-guided
Matrix	Matrigel (xeno-derived, variable)	Matrix-free
Operator dependence	High	Low
Automation compatibility	Limited	Fully automated
Imaging	Deep Z-stack, slow	Planar, high-content
QC gates	Implicit	Multi-stage, quantitative
Donors tested	Limited (1 donor)	≥3 donors (scalable)
Mechanistic DILI classes	Demonstrated	Reproduced + extended
Variability control	Not quantified	Quantified reduction
Translational readiness	Research-grade	Regulatory-aligned

Figure 4. Replication and improvement matrix comparing the original Shinozawa workflow and the CuSTOM Accelerator implementation. Side-by-side comparison of key workflow dimensions influencing reproducibility and translational readiness, including organoid formation, matrix dependence, operator dependence, automation compatibility, imaging burden, QC gates, donor scale, mechanistic coverage, variability control, and translational readiness.

Across two years of repeated execution, we observed concordance with the original dose–response relationships for canonical hepatotoxicants and consistent reproduction of the morphological injury signatures described in the foundational publication. Following implementation of automation and prospective QC gates, IC₅₀ values stabilized across donors and batches and aligned closely with benchmark primary human hepatocyte assays, supporting enhanced assay reliability and translational relevance.

Operationally, transition from manual to automated organoid formation, media exchange, and imaging substantially improved control over organoid production and reduced both biological and technical variability. Microcavity-guided aggregation combined with robotic scheduling yielded more uniform organoid numbers and diameters across donors, batches, and parallel cultures (Fig. 5). Automation consistently reduced intra-line and inter-line variability, as well as intra-batch and inter-batch variation, compared to manually executed workflows. Implementation of prospective QC gates prevented propagation of failed runs, while toxicity-associated morphological signatures clustered reproducibly across donors and repeated executions. In addition, elimination of xeno-derived matrices (Matrigel) reduced labor demands, removed a major source of lot-to-lot variability, and enabled compatibility with more advanced contexts of use, including immune-mediated DILI modeling, as demonstrated previously (Soussi et al., 2025).



The automation-first framework also increased assay throughput, enabling multi-donor toxicity assessments and larger compound screens that would have been impractical using the original manual workflow. Collectively, these changes transformed replication from a one-time verification exercise into a repeatable, institutionally embedded capability, positioning the platform to contribute directly to human-centric drug development and data generation efforts, including the ARPA-H CATALYST program (Research Horizons, 2025b).

3.2 Research Rigor Improvements

Institutionally, the replication effort catalyzed a shift toward embedding rigor and reproducibility across translational organoid workflows at CuSTOM. The standardized methods are now used routinely for internal QC, donor-line characterization, assay benchmarking, and staff training, making the workflow a foundational component of the CuSTOM Accelerator's operational ecosystem. The platform has been leveraged in collaborative programs with partners such as Roche and integrated into multi-institutional initiatives, including ARPA-H CATALYST, where controlled variance and automation support the generation of standardized datasets for AI-enabled toxicity prediction.

Collectively, these changes transformed replication from a project-specific validation exercise into an operational standard that now shapes how translational organoid workflows are developed, evaluated, and deployed across CuSTOM.

3.3 Dissemination and Adoption

The long-term impact extends well beyond the institution. In partnership with Molecular Devices, core elements of the replicated workflow - including microcavity-based organoid manufacturing, standardized compound-exposure modules, automated high-content imaging, and artificial intelligence (AI)-driven toxicity analytics—are progressing toward commercial distribution. This technology transfer will enable multi-site implementation, independent replication, and meaningful extension of the original Shinozawa findings. In parallel, the methodological rigor achieved through replication strengthens CuSTOM's position in regulatory-aligned science, where reproducibility, variance control, and standardized analytics are essential for new approach methodology (NAM) acceptance. The established organoid manufacturing and toxicity-assessment workflow now serves as a template for de-risking additional organoid modalities, supporting population-level studies on genetic modifiers of toxicity, and establishing a foundation for potential future engagement with Innovative Science and Technology Approaches for New Drugs (ISTAND) or Drug Development Tool Qualification (DDTQ) programs focused on human-relevant liver toxicity models.

3.4 Long-Term Impact

Collectively, these outcomes demonstrate that the replication effort was not an isolated exercise, but a sustained initiative that elevated scientific rigor, accelerated translational capability, and established a durable framework for reproducibility-driven innovation with relevance beyond the originating institution.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

The replication effort encountered biological, technical, and operational barriers that required coordinated problem-solving across multidisciplinary teams. Early in the process, inherent variability in organoid morphology and functional maturation - characteristic of spontaneous self-organizing systems - complicated quantitative replication of the Shinozawa phenotypes. Initial batches generated via spontaneous aggregation exhibited increased variability in size distribution, yield, and metabolic readiness, challenging the establishment of a tightly controlled baseline for systematic, large-scale replication. These challenges were addressed by implementing a standardized, matrix-free microcavity workflow, together with fixed feeding schedules and prospective quality control (QC) checkpoints, resulting in improved starting-material uniformity and reduced batch-to-batch variance.

A second barrier arose from the limited compatibility of several manual assay steps with a variance-controlled, automated workflow. Processes such as Matrigel embedding, excess mesenchyme filtration, manually performed media exchanges, and subjective assessment of organoid quality and readiness introduced operator-dependent variability and constrained reproducibility across staff. Through iterative collaboration between biology and engineering teams, these steps were transitioned to automation-compatible unit operations, including robotic liquid handling, automated high-content imaging, and AI-enabled, label-free quantification of organoid quality and toxicity, while preserving alignment with the biological mechanisms interrogated by the original assay (Fig. 2).

Cross-donor reproducibility presented an additional challenge. Early toxicity signatures varied across donor lines, complicating discrimination between true biological differences and inconsistencies arising from differentiation quality or assay execution. Harmonization of lineage-induction steps, implementation of universal QC gates and assay-readiness criteria, and adoption of normalization strategies coupled with computational phenotype clustering reduced these confounders. As a result, donor-independent reproduction of core toxicity signatures was achieved while retaining sensitivity to biologically meaningful donor-specific responses when present.

Effective biology-engineering integration required deliberate optimization. Early automation attempts introduced variability in microcavity occupancy and timing control during single-cell seeding and early aggregate formation, which in turn affected organoid yield and maturation consistency. Close collaboration between automation engineers and the biology team enabled systematic optimization of loading volumes, dispense parameters, and scheduling until automated performance matched or exceeded manual workflows.

Finally, sustaining replication in routine research practice required ensuring that staff across experience levels could reproduce the assay independently. Early success depended on tacit expertise held by a small number of individuals, limiting institutional adoption. The introduction of formal standard operating procedures (SOPs), hands-on training, repeated QC-verified practice runs, and automation that minimized subjective decision-making

stabilized execution such that multiple operators achieved indistinguishable results. Collectively, these barriers and solutions demonstrate that high-fidelity replication in a translational setting requires not only scientific reconstruction, but also coordinated advances in engineering, analytics, operations, and workforce enablement.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

Replication was implemented through a phased, milestone-driven timeline aligned with increasing levels of automation, quality control, and translational readiness. The effort spanned two years, progressing from manual reconstruction to standardized automated execution and institutional integration.

Phase 1: Reconstruction and benchmarking (Dec 2023 – Jun 2024). We reconstructed the differentiation and hepatotoxicity workflows and established baseline performance benchmarks relative to the original publication. During this phase, organoid formation was transitioned from spontaneous aggregation to standardized, guided formation to improve uniformity and enable downstream variance control.

Phase 2: Automation integration and variance control (May 2024 – Sep 2024). Automation was embedded across assay stages. Robotic liquid handling was implemented, media exchange schedules were standardized, and high-content imaging pipelines were configured for automated acquisition and AI-enabled toxicity scoring. Early automated runs informed refinement of loading parameters, robotic handling settings, and QC acceptance thresholds.

Phase 3: Intensive multi-cycle replication and multi-donor robustness testing (Sep 2024–Sep 2025). Multiple staff executed repeated assay cycles across several iPSC donors and numerous organoid batches. This phase generated variance-benchmarking data, stabilized QC metrics, incorporated AI-driven toxicity models into routine operation, and validated automated IC₅₀ extraction pipelines.

Phase 4: Institutionalization and external deployment (Sep 2025–Dec 2025). SOPs were finalized, QC criteria were codified, and training modules were developed to support reproducible execution by new personnel. The workflow was deployed across internal and external programs, including activities within ARPA-H CATALYST, while selected components entered commercialization processes through Molecular Devices to enable multi-site reproducibility.

5.2 Associated Costs and Required Resources

The resources required for this effort reflected a combination of direct institutional investment, externally sponsored research support, and philanthropic funding. Personnel effort across organoid biology, automation engineering, data science, quality control, and operations constituted the largest cost category and was supported through a combination of institutional funding and sponsored research support from the Danaher Beacon Initiative. Consumables – including hPSC maintenance and differentiation reagents, microcavity array plates, assay media, compound libraries, and high-content imaging reagents were supported in part through the Danaher Beacon Initiative. Computational infrastructure, including data storage, image-analysis hardware, and AI development resources, was supported through institutional investment. Philanthropic contributions enabled the establishment of well-characterized donor-line panels required for cross-donor replication and variance benchmarking.

6. COLLABORATION

6.1 Key Collaborations

The successful replication of the Shinozawa HLO-DILI assay depended on a multilayered collaboration network spanning biological, engineering, computational, and operational expertise. At CuSTOM, these capabilities are intentionally co-located within the Accelerator, enabling replication to function as both a scientific and institutional effort.

A foundational partnership occurred between the inventor laboratory - the Takebe Lab, where the original HLO biology and Shinozawa-derived assay were developed - and the CuSTOM Accelerator team responsible for methodological replication and translational engineering. This collaboration ensured biological continuity while enabling systematic redesign of the implementation framework for automation, standardization, and reproducibility.

Within the Accelerator, close day-to-day collaboration between the biology team (led by Dr. Fadoua El Abdellaoui Soussi) and the automation/data science team (led by Sina Dadgar) was essential. These groups iteratively optimized microcavity loading, robotic liquid-handling parameters, scheduling, imaging acquisition settings, and AI-driven toxicity scoring, ensuring engineering changes preserved biological fidelity and biological adjustments enabled stable automation.

Collaboration with Molecular Devices supported high-content imaging configuration, acquisition optimization, and integration of ML-based morphological toxicity prediction tools, and subsequently expanded to commercialization activities that support broader dissemination and multi-site transfer.

Finally, dedicated QC and operations teams enabled consistent execution across time, operators, and donor backgrounds. QC leadership established functional and morphological acceptance criteria and supported variance benchmarking, while operations coordinated procurement, maintenance, scheduling, and resource allocation necessary for sustained multi-cycle replication.

Together, these collaborations created the integrated workforce needed to transform a technically demanding research assay into a reproducible, automated, and translational platform.

7. SUSTAINABILITY / TRANSFERABILITY

The outcomes of this replication effort now function as durable institutional infrastructure, directly advancing CuSTOM's mission to translate advanced organoid biology into reproducible, deployable platforms for research, industry, and regulatory-aligned applications.

7.1 Current Sustainability Plans

The replicated HLO-DILI assay is now a durable component of CuSTOM's translational infrastructure, ensuring that replication-driven advances continue to benefit internal research, external collaborations, and the broader community. Sustainability is grounded in standardization, automation, workforce training, and institutional adoption. By embedding the workflow into formal SOPs encompassing differentiation, microcavity-guided aggregation, automated liquid handling, QC gates, imaging pipelines, and AI-based toxicity analytics, reproducibility is increasingly inherent to the process rather than dependent on individual expertise. SOPs are reinforced through hands-on training and QC-verified practice runs, enabling consistent onboarding of new personnel and ensuring that reproducible execution is sustained as programs, staff, and partnerships continue to expand.

Because the assay was replicated across multiple donor lines, batches, and operators, it functions as a robust platform supporting diverse applications. Its generalizability extends beyond hepatotoxicity to pharmacology studies, immune-mediated mechanisms of injury, population-level toxicity comparisons, and disease modeling (e.g., MASH). Performance stability across genetic backgrounds positions the platform as a useful resource for scientific investigation and for future regulatory-aligned qualification efforts.

Sustainability is further strengthened through integration into large-scale translational programs. Within ARPA-H CATALYST, the replicated assay is now integrated into ongoing data-generation workflows at the scale required for population-level analyses. Controlled variance and automation compatibility enable scale required for population-level analyses. Internally, the workflow serves as a benchmark for qualifying donor lines, assessing organoid manufacturing quality, and training scientists in reproducibility-first execution.

7.2 Transferability and Scaling

A major driver of transferability is ongoing commercialization with Molecular Devices. Core workflow components - including Automated organoid manufacturing, standardized exposure modules, high-content imaging, and

AI-enabled toxicity analytics - are being adapted into a deployable platform. This pathway enables multi-site harmonization and transforms single-site replication into a broadly accessible implementation that supports cross-institutional comparison, benchmarking, and field-wide adoption.

7.3 Future Integration Plans

Looking ahead, multi-site replication studies supported by the commercial platform will help establish community standards for organoid-based toxicity testing. The replication framework is also being extended to additional organoid systems (intestinal, pancreatic), immune-organoid co-cultures, and multi-organ MPS models, allowing the methodological advances from this exemplar to serve as a template across tissue types. Workforce development will expand training in automated execution, QC interpretation, and data analytics, embedding replication literacy within CuSTOM's pipeline. Finally, continued expansion of genetically and phenotypically heterogeneous donor panels will support population-level discovery across toxicity and disease contexts.

Together, these efforts ensure replication impact extends well beyond reconstructing a single study: the replicated HLO-DILI assay now functions as a reproducible, scalable, and transferable platform supporting translational science, regulatory alignment, commercial dissemination, and field-wide reproducibility. Its evolution from methodological replication to institutional capability illustrates how rigorous replication, when embedded within a translational ecosystem, can set new standards for reliability and human relevance in organoid-based research.

8. REFERENCES

1. Shinozawa, T., Kimura, M., Cai, Y., Saiki, N., Yoneyama, Y., Ouchi, R., Koike, H., Maezawa, M., Zhang, R.-R., Dunn, A., Ferguson, A., Togo, S., Lewis, K., Thompson, W. L., Asai, A., & Takebe, T. (2021). High-fidelity drug-induced liver injury screen using human pluripotent stem cell-derived organoids. *Gastroenterology*, 160(3), 831–846.e10. <https://doi.org/10.1053/j.gastro.2020.10.002>
2. Soussi, F. E. A., Brusilovsky, M., Buck, E., et al. (2025). Autologous organoid-T cell co-culture platform for modeling of immune-mediated drug-induced liver injury. *Advanced Science*, 12(43), e08584. <https://doi.org/10.1002/adv.202508584>
3. Research Horizons. (2024). *Collaboration to develop liver organoids for drug toxicity screening*. Cincinnati Children's Hospital Medical Center. <https://scienceblog.cincinnatichildrens.org/collaboration-to-develop-liver-organoids-for-drug-toxicity-screening/>
4. Research Horizons. (2025a). *New human liver organoid platform predicts immune-mediated drug toxicity*. Cincinnati Children's Hospital Medical Center. <https://scienceblog.cincinnatichildrens.org/new-human-liver-organoid-platform-predicts-immune-mediated-drug-toxicity/>
5. Research Horizons. (2025b). *A new era for drug safety studies*. Cincinnati Children's Hospital Medical Center. <https://scienceblog.cincinnatichildrens.org/a-new-era-for-drug-safety-studies/>

INDIVIDUAL WINNER

Michele Di Mascio



A former NIH scientist applying two decades of HIV and viral dynamics research experience to advance scientific discovery while advancing rigorous validation and reproducibility in biomedical research.

Michele Di Mascio is a biomedical scientist with over two decades of experience in infectious disease research, with a particular focus on HIV, viral dynamics, and in vivo imaging. Following postdoctoral training in theoretical biology at Los Alamos National Laboratory in the United States, he joined the U.S. National Institutes of Health, where he led parallel programs in AIDS imaging research and mathematical biology at NIAID, overseeing preclinical and translational studies in nonhuman primates and advancing the understanding of viral pathogenesis and treatment strategies. He has authored over 50 peer-reviewed publications, including in leading journals, and received an NIAID award for establishing a pioneering core facility for in vivo imaging of lentiviral infections in nonhuman primates. His work integrates experimental and quantitative approaches to address complex biological questions, with an emphasis on validation and reproducibility. He now works as an independent consultant, advising academic, governmental, and industry partners on biomedical research strategy, data analysis, and scientific communication

1. REPLICATION STRATEGY

1.1 Scope and Objectives

Replication assumes particular importance when foundational biological claims extend beyond advancing scientific knowledge to influencing subsequent mechanistic hypotheses, translational research, and the allocation of substantial public and scientific resources. The present report describes a decade-long replication effort undertaken within the National Institute of Allergy and Infectious Diseases (USA) to determine whether antibody-based nuclear medicine probes can reliably detect viral replication *in vivo* in SIV-infected nonhuman primates. Specifically, the effort examined whether the foundational biological claims advanced in four influential publications could be independently reproduced under rigorous experimental conditions.

Over the past decade, the NIAID research program that I directed prior to my retirement on September 30th, 2025 (AIDS Imaging Research Section, AIRS), conducted a systematic replication effort focused on these publications, which became influential reference points for noninvasive *in vivo* imaging of Simian Immunodeficiency Virus (SIV) using nuclear medicine cameras. This work culminated in our January 2025 publication (Srinivasula et al., 2025), which provides the scientific foundation for the present report.

Scientific Category: This project falls within the preclinical category (NHP imaging), with direct relevance to translational research because of its implications for mechanistic hypotheses and therapeutic development.

The replication effort evaluated a series of publications arising from a sustained collaborative research program between Emory University and the NIAID Laboratory of Immunoregulation. These studies introduced four related claims:

1. Santangelo et al. (2015): feasibility of *in vivo* imaging of SIV replication using radiolabeled anti-gp120 antibodies across multiple tissues, including organs with extremely low viral loads.
2. Santangelo et al., (2018): use of gp120 imaging to detect residual replication in HAART-treated NHPs and to assess the pharmacodynamic effects of an α 4 β 7-based intervention reported to induce SIV functional cure.
3. Samer et al., (2022): visualization of virus reactivation following administration of a latency-reversing agent using a modified anti-gp120 probe in HAART-treated animals.
4. Kim et al., (2024): expanded longitudinal study of the 2022 imaging approach with additional subjects and repeated probe administrations.

Collectively, these publications became influential reference points for translational virology and HIV cure research, motivating subsequent preclinical investigations and early clinical translation.

1.2 Scientific Rationale for Replication

1.2.1 Biological Plausibility Concerns

The gp120 envelope protein exists at extremely low molar concentrations in infected tissues and, based on polymerase chain reaction (PCR) and in situ hybridization (ISH) necropsy data, accumulates primarily in lymph nodes and spleen, while very limited viral replication occurs in kidney, liver, heart, or across the full thickness of the intestinal wall. While viral replication does occur in the thin mucosal layer of the gut, the much larger submucosal compartment contains essentially no virions, rendering whole-organ gp120 visualization in organs other than lymph nodes and the spleen hard to achieve with currently available technologies.

Several features in the published images (Santangelo et al., 2015; Santangelo et al., 2018; Samer et al., 2022; Kim et al., 2024) raise additional methodological concerns. In multiple figures, tracer uptake in the heart -largely reflecting circulating blood activity- appears higher in SIV-infected animals than in uninfected controls, a pattern more readily explained by non-specific alterations in probe biodistribution than by genuine viral signal. Such systemic differences alone could account for the uniformly elevated uptake reported across multiple organs in the earlier studies, including organs known to harbor limited levels of viral replication.

In other cases (Samer et al., 2022), the images suggest extravasation of the radiolabeled antibody during intravenous administration, likely due to suboptimal injection technique. Antibodies inadvertently introduced into subcutaneous or intradermal tissue enter lymphatic drainage and accumulate non-specifically in regional lymph nodes, producing artifactual patterns that can mimic targeted uptake, particularly in the absence of ex-vivo validation.

Additionally, the reported biodistribution suggests unusually rapid whole-body clearance compared with similar radiolabeled antibodies in the literature, as discussed in our publication (Srinivasula et al., 2025). This atypical behavior raises concerns regarding probe degradation or loss of structural integrity during radiosynthesis, further undermining the plausibility that observed signals represent true gp120 detection. Collectively, these biological and methodological observations provided a strong scientific rationale for undertaking an independent replication effort.

1.2.2 Limited ex vivo Validation in the Original Studies

Across the gp120-imaging literature, *ex vivo* validation of tracer binding in infected tissues was limited. With the exception of preliminary efforts reported in the earliest publication (Santangelo et al., 2015), the studies did not include systematic *ex vivo* assays designed to directly confirm tracer specificity in relevant tissues.

In that study, probe specificity was addressed through a single supplementary figure (Fig. S1b), showing tracer uptake in primary cells isolated from infected lymph nodes and spleen. These experiments were conducted on cryopreserved samples and did not include competition assays, which are generally considered an important experimental control to rule out nonspecific effects, such as differential cell viability between infected and uninfected populations that can influence apparent antibody uptake. In addition, the sample size was very limited (for example, the control group for lymph node cells included only a single sample), and no autoradiography data were reported, despite the established role of autoradiography in assessing target engagement and informing the feasibility of ImmunoPET imaging *in vivo*, as discussed previously (Bray et al., 2010; Srinivasula et al., 2025).

Furthermore, none of the published studies reported *in vivo* or *ex vivo* probe-stability assessments using radio-HPLC. Given the differences between the reported biodistribution profiles and those described in the broader literature, as well as those observed with probes generated using the same antibody and radionuclide in our own work (Srinivasula et al., 2025), the contribution of probe instability or degradation cannot be formally excluded.

Conducting comprehensive *ex vivo* validation and probe-stability testing requires the ability to handle radioactive and infected tissues under appropriate BSL-2/3 containment together with specialized radiochemistry and imaging capabilities. Such capabilities were not available within the infrastructure supporting the original studies. By contrast, our program was specifically structured to support these integrated analyses and to address questions related to the feasibility of imaging viral replication and CD4-pool dynamics in SIV/SHIV-infected nonhuman primates. In this context, the unexpected nature of the original findings, together with the limited *ex vivo* corroboration, provided a clear rationale for a systematic and independent replication effort.

1.3 Programmatic Framework and Definition of Replication

1.3.1 Research Program Infrastructure

As described in Section 2, our research program was established nearly two decades ago around an integrated BSL-2/3 imaging infrastructure designed to rigorously develop, validate, and characterize molecular imaging probes intended to visualize immune and viral dynamics noninvasively. This platform enabled:

- radiotracer development and optimization;
- *in vivo* nonhuman-primate imaging across multiple infection states;
- binding assays using primary infected cells and tissues;
- *ex vivo* and *in vivo* probe-stability assessments;
- quantitative and mathematical modeling of imaging and biodistribution data.

1.3.2 Working Definition of Replication

Replication was defined using a hybrid framework combining direct and conceptual replication. Direct replication sought to reproduce the original experimental conditions as closely as possible, preserving identical experimental conditions whenever necessary to isolate and evaluate specific aspects of radiotracer performance, including antibody class, radiolabeling method, infection model, imaging parameters, timing, and quantitative analyses. Conceptual replication incorporated complementary validation procedures - including *ex vivo* binding assays, probe-stability analyses, and alternative antibody scaffolds and radionuclides- to evaluate whether the underlying biological claim remained supported under independent experimental conditions and to distinguish biological signal from biodistribution artifacts.

The unit of replication was the core biological claim: namely, that radiolabeled antibodies with nanomolar-range affinities against gp120 can noninvasively visualize viral replication across multiple tissues *in vivo*, both during chronic untreated SIV infection and under suppressive antiretroviral therapy. Importantly, this formulation reflects the implicit affinity regime of the antibodies used in the original studies, for which binding affinities were not experimentally reported. In our replication work (Srinivasula et al. 2025), we independently measured these affinities of the anti-gp120 probes utilized and found them to be in the nanomolar range, including for the 7D3 antibody used in the original studies. Replication outcomes were evaluated using predefined criteria, including detectable *in vivo* signal above background in lymphoid tissues, reproducible spatial patterns, confirmation of probe–target binding *ex vivo*; and demonstration of probe stability by radio-HPLC.

Using this framework, we implemented both direct and conceptual replication by expanding the original radiolabeling strategies. We radiolabeled the murine anti-gp120 monoclonal antibody 7D3 used in the first two publications (Santangelo et al., 2015; Santangelo et al., 2018) with ⁶⁴Cu (half-life: 12.7 hours), closely mirroring the original methodology to enable direct comparison of biodistribution in uninfected animals. Comparison with the published biodistributions patterns revealed substantial differences, most notably that the apparent whole-body clearance of the tracer reported in the original studies was markedly faster than expected for intact antibodies, whereas the biodistribution observed in our experiments was consistent with established antibody-imaging literature.

As part of the conceptual replication work, we radiolabeled 7D3 with ⁸⁹Zr (half-life: 78.4 hours), enabling image acquisition at later time points following probe administration, after more extensive systemic clearance, thereby increasing the signal-to-noise ratio of the acquired images.

Under these conditions, no differential uptake between infected and uninfected animals was observed. After determining that endogenous anti-gp120 antibodies present in SIV-infected plasma fully abrogated binding of the radiolabeled 7D3 probe, we conducted additional experiments using a fully rhesus ITS anti-gp120 antibody fragmented into F(ab')₂, which did not compete with endogenous anti-gp120 antibodies in most infected subjects and closely matched the molecular weight of the probes used in later publications (Samer et al., 2022; Kim et al., 2024). The ITS-based probe exhibited biodistribution consistent with the broader antibody-imaging literature but not with patterns reported in the original studies, again suggesting unexplained differences in probe preparation.

After confirming that our probes bound gp120 with equal or greater affinity than those of the probes used in their publications (Santangelo et al., 2015; Santangelo et al., 2018; Samer et al., 2022; Kim et al., 2024), we attempted to reproduce the reported *ex vivo* assay demonstrating specific binding of the probe to primary cells obtained from SIV-infected animals relative to uninfected controls (Santangelo et al., 2015). In contrast to the published findings, all experiments yielded negative results. Using competition assays and substantially larger sample sizes, we evaluated cryopreserved and freshly isolated PBMCs, splenocytes, and lymph-node cells and consistently observed no gp120-specific binding.

Together, these experiments operationalized both direct and conceptual replication, enabling independent evaluation of the underlying biological claims as well as the plausibility and stability of the imaging probes. This framework ensured that replication outcomes reflected underlying biology rather than methodological divergence.

1.3.3 Integration into Research Practice

Replication was embedded within the overall research strategy rather than implemented as a downstream verification step. All newly synthesized probes underwent systematic *in vitro* binding and stability testing before animal use or interpretation of imaging studies. Each imaging session paired infected and uninfected subjects matched by weight to control for non-specific biodistribution. Standardized quantitative pipelines ensured comparability across animals, time points, and tracer batches, supported by long-term infrastructure planning, equipment calibration, validated protocols, and standardized training. Within this framework, replication functioned not as a downstream checkpoint but as a guiding principle for experimental design, validation, and interpretation.

1.4 Replication Strategy and Experimental Findings

The original biological claims were evaluated using complementary experimental approaches that included:

- gp120-targeting antibodies radiolabeled under tightly controlled conditions;
- *in vivo* imaging across acute and chronic SIV infection;
- systematic probe-stability assessments;
- *ex-vivo* binding assays on primary infected cells and tissue sections with competition assays;
- an alternative gp120-targeting antibody to exclude interference from endogenous antibodies;
- standardized quantitative analyses across all animals.

As discussed in our publications, autoradiography on tissue sections provides the most informative approaches for assessing the feasibility of molecular imaging. Negative autoradiography findings indicate either insufficient probe affinity, insufficient target abundance, or a combination of both. Autoradiography was not performed in any of the four studies under evaluation (Santangelo et al., 2015; Santangelo et al., 2018; Samer et al., 2022; Kim et al., 2024), nor in subsequent NIH-funded publications inspired by those reports (Obregon-Perko et al., 2021; Beckford-Vera et al., 2022).

Experimental Findings:

Across all complementary experimental approaches, the findings were internally consistent:

- no detectable gp120-specific signal *in vivo* in any tissue, including lymph nodes and spleen;
- no gp120-specific binding *ex vivo* in infected cells or tissues;
- no reproducible spatial imaging patterns;
- biodistribution in uninfected controls markedly different from that reported previously, suggesting probe instability or non-specific behavior.

These findings are described in detail in our 2025 publication (Srinivasula et al., 2025), which systematically compares each of the original biological claims with the corresponding replication outcomes.

1.5 Innovative Aspects of the Replication Strategy

The replication effort incorporated capabilities not integrated into the original experimental framework, including coupled *in vivo* imaging and immediate *ex vivo* validation, probe-stability analysis, use of multiple antibody scaffolds, and a purpose-built BSL-2/3 imaging environment. Collectively, these elements illustrate a replication framework in which methodological validation is integrated throughout the research process rather than introduced only after experimental completion.

1.6 Connection to Related Replication Efforts

Two additional replication projects from our program reinforce the gp120 findings: failure to reproduce the $\alpha 4\beta 7$ -based “functional cure” claim (Santangelo et al., 2018; Byraredy et al., 2016; Di Mascio et al., 2019) and failure to reproduce CD4-pool imaging in the gastrointestinal tract (Byraredy et al., 2016; Di Mascio et al., 2018;

Kim et al., 2023). Independent studies have shown that even highly expressed lymphocyte receptors can be reliably imaged only in lymph nodes and spleen, but not throughout the gastrointestinal tract or other non-lymphoid organs. The gp120-imaging studies (Santangelo et al., 2015; Santangelo et al., 2018; Samer et al., 2022; Kim et al., 2024), which reported strong multi-organ signals despite gp120 being orders of magnitude less abundant, are therefore most parsimoniously explained by biodistribution artifacts or probe instability. Following the same line of reasoning, because the reported gp120 imaging findings were used to further support biological conclusions regarding reductions in viral reservoirs following $\alpha 4\beta 7$ immunotherapy (Santangelo et al., 2018), the independent failure to reproduce the reported immunotherapeutic effect (Di Mascio et al., 2019) naturally strengthened the scientific rationale for reexamining whether the imaging signals themselves truly reflected viral replication rather than probe biodistribution artifacts or probe instability. Together, these replication efforts converge on the conclusion that the underlying biological claims are not reproducible under rigorous experimental conditions.

1.7 Impact and Scientific Importance

Because the original biological claims influenced mechanistic hypotheses, therapeutic strategies, translational research, and the allocation of substantial public and scientific resources, their independent evaluation carries important scientific, ethical, and societal implications. In 2021, an independent Australian human study (McMahon et al., 2021), inspired by the original nonhuman-primate imaging reports, likewise reported negative findings. Nevertheless, additional preclinical and clinical investigations continued despite the absence of *ex vivo* validation, illustrating how influential biological claims can shape downstream translational research even when independent confirmation remains limited. The AIRS 2025 replication study (Srinivasula et al., 2025) clarifies important methodological limitations of antibody-based nuclear medicine imaging in lentiviral infection while providing a practical framework for strengthening methodological standards, independent validation, and the responsible stewardship of public and scientific resources in future translational imaging research.

2. ENVIRONMENT

2.1 Environment Description

The replication activities were conducted within NIAID and were supported by nearly two decades of sustained infrastructure development dedicated to high-containment molecular imaging of SIV/SHIV infection in nonhuman primates. When this research program was initiated in 2003, no NIH laboratory had the capability to integrate nuclear medicine imaging with NHP lentiviral research. The initial imaging capability was established through a contract with Bioqual, a nonhuman-primate research facility in Maryland, where a surplus SPECT camera from the NIH Clinical Center was repurposed to create the first NIAID platform for nuclear imaging of infected NHPs under NIAID oversight.

This platform operated until 2008, when the contract ended and imaging studies were temporarily halted. During this period, institutional support was secured to rebuild an internal NIH capability. In 2011, NIAID re-established a fully integrated BSL-2/3 imaging laboratory at the NIH Animal Center, enabling resumption of CD4-pool imaging and early attempts to visualize viral replication. PET capability was added alongside SPECT; subsequent studies demonstrated that imaging modality alone did not overcome the central biological limitation, as *ex vivo* binding assays in primary cells and tissues remained negative.

During the same period, complementary NIAID-supported imaging studies were conducted at Emory University in collaboration with the NIAID Laboratory of Immunoregulation (Santangelo et al., 2015; Santangelo et al., 2018), using higher-resolution PET systems but at lower biosafety containment levels that limited the ability to perform the integrated *ex vivo* validation studies described in Section 1.

Despite differences in instrumentation, our published findings consistently indicated that the fundamental barrier was biological rather than technological, as no specific gp120-associated signal could be demonstrated in any tissue across platforms. Over subsequent years, and despite delays imposed by the COVID-19 pandemic, we expanded and strengthened our validation capabilities. By that time, the AIRS laboratory had evolved into a uniquely integrated high-containment molecular imaging environment, combining radiotracer synthesis, quantitative PET/SPECT imaging, and immediate *ex vivo* validation of radioactively labeled tissues from infected nonhuman primates within a single coordinated workflow. To our knowledge, at that time no other laboratory worldwide provided this integrated capability within a single BSL-2/3 research environment.

Replication studies were conducted at the NIH Animal Center in Poolesville, MD, within the Integrated Research Facility (IRF), by a multidisciplinary team led by me as Chief of the AIDS Imaging Research Section and the Mathematical Biology Section. The multidisciplinary team included radiochemists responsible for probe design, synthesis, and radio-HPLC quality control; technologists responsible for probe administration and quantitative imaging; veterinarians and animal-care staff responsible for NHP procedures and tissue collection; biologists responsible for *ex vivo* assays; and mathematicians responsible for developing analytical pipelines for data interpretation.

All activities were supported exclusively by federal funds allocated to NIAID. No industry or commercial funding contributed to the program, ensuring full methodological independence in experimental design, execution, data analysis and interpretation. This purpose-built interdisciplinary environment provided the infrastructure required for rigorous evaluation of high-impact biological claims and established a reproducible framework for imaging-based virology studies.

3. IMPACT

3.1 Replication Activity Impact

The replication activities had important scientific, translational, and editorial implications extending beyond the specific biological question under investigation. The initial NIAID-funded NHP studies published in 2015 and 2018 (Santangelo et al., 2015; Santangelo et al., 2018), later expanded in 2022 and 2024 (Samer et al., 2022; Kim et al., 2024), served as the preclinical foundation for two clinical trials published in 2021 (McMahon et al., 2021) and 2022 (Beckford-Vera et al., 2022). These clinical studies reached different conclusions. The independently conducted Australian trial (McMahon et al., 2021) reported negative results, whereas the NIH-funded study (Beckford-Vera et al., 2022) reported partial feasibility while acknowledging important methodological limitations that were further examined in our 2025 replication study (Srinivasula et al., 2025). Additional preclinical and clinical investigations worldwide continue to build upon the four NHP studies (Santangelo et al., 2015; Santangelo et al., 2018; Samer et al., 2022; Kim et al., 2024) discussed in the present report.

These developments illustrate a broader structural challenge in translational biomedical research. Independent replication of complex nonhuman-primate imaging studies requires specialized BSL-2/3 containment, radiochemistry infrastructure, and multidisciplinary expertise available to only a limited number of laboratories worldwide. Because such studies are technically demanding, ethically sensitive, and often cost well in excess of one million dollars, independent replication is rarely performed rapidly, allowing influential biological claims to shape subsequent research for extended periods before independent verification becomes available. The present replication effort provided an independent scientific benchmark against which these foundational biological claims could be evaluated.

Through direct and conceptual replication, we demonstrated that two independent anti-gp120 radiolabeled probes (7D3 and ITS-F(ab')₂) did not detect viral replication *in vivo* nor in infected primary cells or tissues *ex vivo*, and that their biodistribution profiles in uninfected subjects differed markedly from those reported in the four prior publications. These findings clarified important biological limitations of immunoPET in lentiviral infection and provide a foundation for refining the current understanding of the feasibility of antibody-based imaging of viral replication.

The translational implications extend beyond imaging methodology. Both clinical studies (Beckford-Vera et al., 2022; McMahon et al., 2021) involved healthy volunteers exposed to radiation and HIV-infected patients without direct therapeutic benefit, circumstances that demand particularly strong scientific justification. Our findings reinforce the importance of rigorous biological validation before advancing high-risk molecular imaging approaches to clinical investigation.

More broadly, noninvasive imaging of viruses such as SIV or HIV, if biologically feasible, would be transformative, enabling new insights into pharmacodynamics, reservoir biology, and prophylaxis. Since no wild-type virus had been imaged *in vivo* prior to the 2015 claim (Santangelo et al., 2015), the evidentiary bar was high. The present replication effort helped realign expectations with the underlying biological constraints.

Finally, this experience highlights a broader challenge for scientific publishing and translational research: Over more than a decade, publications supporting gp120-based viral imaging became widely cited as foundational, influencing funding priorities, research directions, and translational strategies before independent replication became available. By providing the first comprehensive independent evaluation of these biological claims, our

work illustrates how rigorous replication can strengthen scientific self-correction, improve methodological standards, and support the responsible stewardship of public funding, specialized scientific infrastructure, and ethically sensitive research resources when influential translational hypotheses are evaluated.

4. RESPONSE TO BARRIERS

4.1 Publication Barriers

One of the principal barriers encountered during this replication effort involved dissemination of the findings. Multiple submissions to high-impact scientific journals were declined on the basis of "insufficient interest," without the scientific validity of the replication work being questioned as the stated basis for the editorial decision.

This experience illustrates what appears to be a broader asymmetry within scientific publishing. Novel biological claims that introduce potentially transformative concepts naturally receive substantial editorial visibility because of their capacity to stimulate scientific investigation. By contrast, our experience suggests that independent studies challenging influential biological claims may be regarded as being of insufficient editorial interest, even when they fundamentally reassess the very body of evidence that originally attracted widespread scientific attention and motivated the scientific investments whose validity they now seek to evaluate.

This observation raises a broader question regarding how scientific importance is assessed. Once biological claims have influenced years of basic research, translational studies, clinical investigation, funding priorities, and extensive scientific citation, rigorous independent evaluations that substantially revise the evidentiary landscape may carry scientific importance comparable to that of the original publications. Their scientific value lies not primarily in novelty, but in their potential to refine current understanding, redirect subsequent research, and prevent continued investment in biological hypotheses whose robustness has not yet been independently established.

For the self-correcting process of science to function efficiently, independent replication should ideally reach the same scientific audience that was influenced by the original reports. Otherwise, influential biological claims and the independent evidence evaluating their robustness may propagate through different dissemination pathways, slowing scientific correction and prolonging scientific uncertainty.

The purpose of these observations is not to question individual editorial decisions, which necessarily involve many considerations beyond scientific validity, but rather to highlight a structural feature of the current publication ecosystem that may influence the pace at which scientific knowledge is corrected. This asymmetry may represent an important contributor to what I propose later in this report as *scientific risk*.

4.2 Toward a Framework for Scientific Risk

Biomedical research routinely evaluates multiple dimensions of risk, including biosafety risk, biosecurity risk, and ethical risk. The present case suggests that a fourth complementary dimension may merit explicit consideration in translational biomedical research: scientific risk.

Scientific risk may be defined as the potential for influential biological claims, when not independently evaluated in a timely manner, to shape subsequent basic research, translational development, clinical investigation, funding priorities, and scientific consensus before their robustness has been independently established.

Unlike biosafety or research ethics, scientific risk does not arise from hazards associated with biological agents or experimental procedures themselves. Rather, it arises from the downstream scientific consequences of uncertainty surrounding influential biological observations.

Importantly, scientific risk is independent of whether an original hypothesis ultimately proves correct or incorrect. Instead, it reflects the potential consequences of allowing influential biological claims to guide subsequent scientific activity before independent verification becomes feasible. Even hypotheses that are ultimately confirmed

may initially carry scientific risk if independent validation is delayed and the resulting uncertainty influences major scientific or translational decisions.

Scientific risk is expected to be greatest when independent replication requires substantial public investment, specialized multidisciplinary expertise, rare scientific infrastructure, or ethically sensitive research resources, making timely validation inherently difficult. Under these circumstances, uncertainty surrounding influential findings may persist for many years while continuing to shape scientific priorities and resource allocation.

The present case illustrates how scientific risk may translate into prolonged scientific uncertainty, continued investment of public funding, utilization of specialized scientific infrastructure, use of ethically sensitive research resources, and delayed correction of the scientific record. Explicit consideration of scientific risk may therefore complement existing frameworks for biosafety, biosecurity, research ethics, and scientific stewardship when evaluating high-impact translational hypotheses.

5. TIMELINE & BUDGET

5.1 Program Development and Replication Timeline

The replication effort required an exceptional long-term investment of scientific expertise, specialized infrastructure, and public funding extending over more than two decades. Rather than representing a single experimental study, the work described here reflects the culmination of a sustained research program established to develop and rigorously validate molecular imaging approaches for studying SIV/SHIV infection in nonhuman primates.

Infrastructure development began in 2003, when NIAID recruited me to establish a program integrating nuclear medicine with SIV/SHIV nonhuman-primate research, an area for which no intramural capability existed at that time. Because NIH could not yet support this work on campus, the first imaging laboratory was established externally at Bioqual under an NIAID contract. A surplus SPECT camera from the NIH Clinical Center was repurposed to establish the first government-supported molecular imaging platform dedicated to studying SIV/SHIV-infected nonhuman primates.

In 2008, the NIAID contract supporting the Bioqual imaging platform ended, resulting in an interruption of experimental imaging activities. Although the long-term scientific objectives of the program remained unchanged, experimental work resumed only after a new BSL-2/3 molecular imaging laboratory became operational at the NIH Animal Center in 2011.

During this transition period, NIAID continued pursuing the same overarching scientific mission that had originally motivated establishment of our program through extramural funding, supporting molecular imaging studies at Emory in collaboration with the NIAID Laboratory of Immunoregulation (LIR). As discussed in Section 2, these studies were conducted in a research environment operating at a lower biosafety containment level than the newly established NIAID platform, which did not support the integrated *ex vivo* validation analyses that later became central to our replication strategy. The objective remained the development of nuclear medicine approaches capable of noninvasively visualizing viral replication and immune-cell dynamics in SIV/SHIV-infected nonhuman primates. By the time the new NIAID imaging laboratory became operational in 2011, influential publications reporting the feasibility of noninvasive *in vivo* imaging of SIV replication had begun to emerge, first in untreated infection and subsequently during antiretroviral therapy (Santangelo et al., 2015; Santangelo et al., 2018).

Because the newly established NIAID platform had been specifically designed to address precisely these biological questions through integrated quantitative molecular imaging and immediate *ex vivo* validation under high-containment conditions, independent evaluation of the published findings became not only a scientific opportunity but also a responsibility associated with the stewardship of a unique publicly funded research infrastructure. Rather than simply reproducing the published experiments, our program was therefore redirected toward systematic evaluation of the biological plausibility, reproducibility, and methodological robustness of the reported imaging claims.

Over the following decade, the platform evolved into an integrated high-containment research environment combining radiotracer development, quantitative PET/SPECT imaging, and immediate *ex vivo* validation of radioactive tissues from infected nonhuman primates. To our knowledge, at that time no other laboratory worldwide integrated all of these capabilities within a single BSL-2/3 research environment.

Between 2011 and 2024, the replication experiments included:

- radiolabeling the original 7D3 antibody with ^{64}Cu under rigorously controlled conditions;
- extending studies with ^{89}Zr to assess later post-injection imaging time points;
- implementing extensive *ex vivo* validation workflows (primary-cell assays, competition assays, tissue autoradiography);
- testing a second gp120-targeting antibody (ITS F(ab')₂) to address interference by endogenous antibodies;
- conducting repeated *in vivo* imaging across acute and chronic infection states.

Collectively, these studies constituted one of the most comprehensive independent evaluations of antibody-based immunoPET imaging of viral replication performed in nonhuman primates. Across all experimental approaches, neither probe demonstrated specific detection of viral replication *in vivo* or *ex vivo*.

5.2 Costs and Resources

Preclinical molecular imaging in nonhuman primates represents one of the most resource-intensive areas of biomedical research.

Each major experimental phase required radiochemistry production; BSL-2/3 containment with high-cost NHP husbandry; tracer stability analysis, validation, quantitative imaging pipelines; and specialized multidisciplinary personnel. Individual experimental phases typically exceeded \$1 million reflecting the combined costs of nonhuman-primate resources, radiation-use constraints, biosafety operations, specialized personnel, and long-term infrastructure support. All activities were supported exclusively through NIAID intramural funding.

5.3 Impact of Coordination and Collaboration Barriers on Timeline

The duration of the replication effort reflected not only the intrinsic complexity of the experimental work but also the challenges associated with independently evaluating influential biological claims requiring specialized infrastructure, multidisciplinary expertise, and ethically sensitive research resources. Unlike conventional replication studies, the work required evaluation of multiple biological hypotheses, independent optimization of radiochemistry, development of new validation assays, and integration of complementary *in vivo* and *ex vivo* methodologies capable of distinguishing genuine biological signals from methodological artifacts.

Because no single experimental result could resolve the discrepancies between the published findings and our observations, multiple complementary hypotheses had to be investigated independently. This required evaluation of different radiotracers, antibody formats, probe-stability analyses, *ex vivo* validation strategies, and quantitative imaging approaches before alternative methodological explanations could be systematically excluded. Collectively, these activities substantially extended both the duration and the cost of the replication effort.

5.4 Lesson for Future NIH-Supported Programs

The present case illustrates that the resources required for independent replication may approach, and in some circumstances exceed, those required to generate the original observations. When influential biological claims depend upon specialized infrastructure, rare multidisciplinary expertise, substantial public investment, or ethically sensitive research resources, independent verification inevitably becomes slow, expensive, and feasible only within a limited number of research environments.

This experience suggests that reducing the time and cost associated with future replication efforts may require mechanisms that facilitate earlier independent evaluation of influential biological claims. Earlier access to primary data, structured opportunities for scientific dialogue, transparent data sharing, and dedicated pathways supporting independent validation could substantially accelerate scientific self-correction while improving stewardship of public funding, specialized scientific infrastructure, and ethically sensitive research resources.

More broadly, the present case illustrates how *scientific risk* may translate directly into long-term commitments of time, infrastructure, and public investment. When independent replication depends upon rare capabilities available only within a few specialized laboratories, uncertainty surrounding influential biological claims may persist for many years before robust independent evaluation becomes feasible. Recognizing this dimension of scientific risk

may help inform future funding strategies, editorial practices, and institutional policies aimed at strengthening the efficiency, transparency, and self-correcting nature of translational biomedical research.

6. COLLABORATION

6.1 Key Collaborations

The replication activities illustrate both the value of scientific collaboration and the importance of developing mechanisms capable of rapidly resolving uncertainty surrounding influential biological claims. Imaging viral replication *in vivo* during suppressive antiretroviral therapy represents a transformative scientific goal with broad implications for biology, translational medicine, and clinical trial design. In such settings, scientific independence and collaboration should be viewed as complementary rather than competing components of scientific progress.

In retrospect, stronger coordination between complementary intramural and extramural research programs pursuing the same biological objective might have accelerated resolution of the discrepancies that subsequently emerged following the initial publications (Santangelo et al., 2015; Santangelo et al., 2018). During the period in which the NIAID imaging platform was temporarily unavailable, complementary extramural research programs advanced this area of investigation and generated the first influential reports describing the feasibility of noninvasive *in vivo* imaging of SIV replication. Following re-establishment of the NIAID imaging platform, the two research environments continued addressing the same biological question through independent experimental efforts.

In practice, opportunities for direct scientific comparison remained largely limited to conventional mechanisms such as peer-reviewed publications and scientific conferences. While these mechanisms are indispensable for disseminating scientific knowledge and stimulating new lines of investigation, they are not specifically designed to enable rapid methodological resolution of conflicting experimental findings. Scientific publications necessarily operate on extended editorial timelines, whereas conferences primarily serve to communicate emerging findings to the broader scientific community rather than to facilitate structured experimental reconciliation between independent investigators. As a consequence, important methodological uncertainties may persist for years before they are independently resolved. As commonly occurs in biomedical research, scientific independence, intellectual-property considerations, and healthy scientific competition naturally favor parallel rather than jointly designed studies. While this model appropriately preserves scientific independence, it may also delay resolution of influential biological questions when complementary expertise resides in different research environments.

In retrospect, even a single jointly designed methodological validation experiment, performed early after publication and integrating complementary expertise in radiochemistry, quantitative imaging, and integrated *ex vivo* validation, might have resolved the central methodological uncertainty within days or weeks rather than years. Such an experiment could have rapidly distinguished between methodological differences, alternative biological interpretations, or technical artifacts, substantially reducing subsequent scientific uncertainty while avoiding duplication of resource-intensive nonhuman-primate studies. Importantly, the objective of such experiments is not to determine which investigators are "right," but to identify, as early as possible, the origin of divergent findings so that the scientific community can converge more rapidly toward robust and reproducible knowledge.

Viewed through the framework proposed in Section 4, this represents a missed opportunity for mitigation of scientific risk. When influential biological claims emerge from research requiring substantial public investment, specialized infrastructure, multidisciplinary expertise, and ethically sensitive research resources, the cost of delayed independent evaluation may greatly exceed the cost of performing an early collaborative methodological validation experiment. In many instances, a single carefully designed methodological validation experiment may prevent years of scientific uncertainty, unnecessary duplication of effort, unnecessary expenditure of public resources, and avoidable delays in scientific and translational progress.

More broadly, this experience suggests that, for influential biological claims associated with substantial scientific risk, the research ecosystem may benefit from complementary mechanisms specifically designed to facilitate early methodological reconciliation through jointly designed methodological validation experiments, while fully preserving scientific independence and intellectual property. Such mechanisms would not replace peer-reviewed publications, scientific conferences, or independent replication; rather, they would complement these established pillars of scientific communication by accelerating resolution of critical methodological uncertainties before they propagate through subsequent basic, translational, and clinical research. In doing so, they may simultaneously

reduce scientific risk and strengthen scientific stewardship, particularly for research involving specialized infrastructure, BSL-2/3 containment, radioactive materials, and nonhuman primates.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Current and Future Sustainability Plans

The replication framework developed here provides a simple, scalable, and policy-relevant model for ensuring that future imaging claims are grounded in reproducible biology while strengthening scientific stewardship of public resources. Its core principle is straightforward: no viral-imaging study should advance to costly NHP experiments or human trials without first demonstrating *ex vivo* probe binding in infected tissues. Such verification, via autoradiography or equivalent assays, requires only hours of laboratory work, yet was absent from all four influential publications motivating this replication effort.

Institutionalizing this minimal validation step would immediately improve sustainability, reduce unnecessary use of high-cost resources, mitigate scientific risk at an earlier stage, and help prevent translational detours driven by insufficiently validated biological claims. The approach is readily transferable across pathogens, imaging platforms, and research programs, and could be incorporated into funding strategies, peer-review processes, project planning, investigator training, and other research-governance activities.

Looking forward, this model offers a blueprint for stronger federal stewardship of public resources: rigorous validation first, costly translation only when justified. Adopting these principles would ensure that transformative technologies, such as noninvasive viral imaging, advance on solid scientific footing rather than momentum alone.

More broadly, this case study encourages further discussion of *scientific risk* as a complementary framework alongside biosafety, biosecurity, research ethics, and scientific stewardship. Anticipating and mitigating scientific risk at an early stage will likely require the development of dedicated frameworks and practical mechanisms for timely methodological reconciliation and independent validation. Such approaches may help strengthen the efficiency, robustness, and sustainability of translational biomedical research while maximizing the scientific and societal value of public investment.

8. REFERENCES

1. Beckford-Vera, D. R., Flavell, R. R., Seo, Y., Martinez-Ortiz, E., Aslam, M., Thanh, C., Fehrman, E., Pardons, M., Kumar, S., Deitchman, A. N., Ravanfar, V., Schulte, B., Wu, I. W. K., Pan, T., Reeves, J. D., Nixon, C. C., Iyer, N. S., Torres, L., Munter, S. E., ... Henrich, T. J. (2022). First-in-human immunoPET imaging of HIV-1 infection using ⁸⁹Zr-labeled VRC01 broadly neutralizing antibody. *Nature Communications*, 13, 1219. <https://doi.org/10.1038/s41467-022-28727-5>
2. Bray, M., Di Mascio, M., de Kok-Mercado, F., Mollura, D. J., & Jagoda, E. (2010). Radiolabeled antiviral drugs and antibodies as virus-specific imaging probes. *Antiviral Research*, 88, 129–142. <https://doi.org/10.1016/j.antiviral.2010.08.005>
3. Byraredy, S. N., Arthos, J., Cicala, C., Villinger, F., Ortiz, K. T., Little, D., Sidell, N., Kane, M. A., Yu, J., Jones, J. W., Santangelo, P. J., Zurla, C., McKinnon, L. R., Arnold, K. B., Woody, C. E., Walter, L., Roos, C., Noll, A., Van Ryk, D., ... Ansari, A. A. (2016). Sustained virologic control in SIV+ macaques after antiretroviral and α4β7 antibody therapy. *Science*, 354, 197–202. <https://doi.org/10.1126/science.aag1276>
4. Di Mascio, M., Lifson, J. D., Srinivasula, S., Kim, I., DeGrange, P., Keele, B. F., Belli, A. J., Reimann, K. A., Wang, Y., Proschan, M., Lane, H. C., & Fauci, A. S. (2019). Evaluation of an antibody to α4β7 in the control of SIVmac239-nef-stop infection. *Science*, 365(6457), 1025–1029. <https://doi.org/10.1126/science.aav6695>
5. Di Mascio, M., Srinivasula, S., Kim, I., Duralde, G., St. Claire, A., DeGrange, P., St. Claire, M., Reimann, K. A., Gabriel, E. E., Carrasquillo, J., Reba, R. C., Paik, C., & Lane, H. C. (2018). Total body CD4+ T cell dynamics in treated and untreated SIV infection revealed by in vivo imaging. *JCI Insight*, 3(13), e97880. <https://doi.org/10.1172/jci.insight.97880>
6. Kim, I., Srinivasula, S., DeGrange, P., Long, B., Jang, H., Carrasquillo, J. A., Lane, H. C., & Di Mascio, M. (2023). Quantitative PET imaging of the CD4 pool in nonhuman primates. *European Journal of Nuclear Medicine and Molecular Imaging*, 50, 14–26. <https://doi.org/10.1007/s00259-022-05940-4>

7. Kim, J., Bose, D., Arainga, M., Haque, M. R., Fennessey, C. M., Caddell, R. A., Thomas, Y., Ferrell, D. E., Ali, S., Grody, E., Goyal, Y., Cicala, C., Arthos, J., Keele, B. F., Vaccari, M., Lorenzo-Redondo, R., Hope, T. J., Villinger, F., & Martinelli, E. (2024). TGF- β blockade drives a transitional effector phenotype in T cells, reversing SIV latency and decreasing viral reservoirs in vivo. *Nature Communications*, 15, 1348. <https://doi.org/10.1038/s41467-024-45555-x>
8. McMahon, J. H., Zerbato, J. M., Lau, J. S. Y., Lange, J. L., Roche, M., Tumpach, C., Dantanarayana, A., Rhodes, A., Chang, J., Rasmussen, T. A., Mackenzie, C. A., Alt, K., Hagenauer, M., Roney, J., O'Bryan, J., Carey, A., McIntyre, R., Beech, P., O'Keefe, G. J., ... Lewin, S. R. (2021). A clinical trial of non-invasive imaging with an anti-HIV antibody labelled with copper-64 in people living with HIV and uninfected controls. *EBioMedicine*, 65, 103252. <https://doi.org/10.1016/j.ebiom.2021.103252>
9. Obregon-Perko, V., Bricker, K. M., Mensah, G., Uddin, F., Rotolo, L., Vanover, D., Desai, Y., Santangelo, P. J., Jean, S., Wood, J. S., Connor-Stroud, F. C., Ehnert, S., Berendam, S. J., Liang, S., Vanderford, T. H., Bar, K. J., Shaw, G. M., Silvestri, G., Kumar, A., ... Chahroudi, A. (2021). Dynamics and origin of rebound viremia in SHIV-infected infant macaques following interruption of long-term antiretroviral therapy. *JCI Insight*, 6(23), e152526. <https://doi.org/10.1172/jci.insight.152526>
10. Samer, S., Thomas, Y., Arainga, M., Carter, C., Shirreff, L. M., Arif, M. S., Avita, J. A., Frank, I., McRaven, M. D., Thuruthiyil, C. T., Heybeli, V. B., Anderson, M. R., Owen, B., Gaisin, A., Bose, D., Simons, L. M., Hultquist, J. F., Arthos, J., Cicala, C., ... Martinelli, E. (2022). Blockade of TGF- β signaling reactivates HIV-1/SIV reservoirs and immune responses in vivo. *JCI Insight*, 7, e162290. <https://doi.org/10.1172/jci.insight.162290>
11. Santangelo, P. J., Cicala, C., Byraredy, S. N., Ortiz, K. T., Little, D., Lindsay, K. E., Gumber, S., Hong, J. J., Jelacic, K., Rogers, K. A., Zurla, C., Villinger, F., Ansari, A. A., Fauci, A. S., & Arthos, J. (2018). Early treatment of SIV+ macaques with an $\alpha 4\beta 7$ monoclonal antibody alters virus distribution and preserves CD4+ T cells in later stages of infection. *Mucosal Immunology*, 11, 932–946. <https://doi.org/10.1038/mi.2017.112>
12. Santangelo, P. J., Rogers, K. A., Zurla, C., Blanchard, E. L., Gumber, S., Strait, K., Connor-Stroud, F., Hoxie, J. A., Vidakovic, B., Ansari, A. A., Hunter, E., & Villinger, F. (2015). Whole-body immunoPET reveals active SIV dynamics in viremic and antiretroviral therapy-treated macaques. *Nature Methods*, 12, 427–432. <https://doi.org/10.1038/nmeth.3320>
13. Srinivasula, S., Kim, I., Jang, H., DeGrange, P., Brown, H., Dalton, V., Badralmaa, Y., Natarajan, V., Long, B., Carrasquillo, J. A., & Di Mascio, M. (2025). Whole-body PET imaging of simian immunodeficiency virus using gp120-targeting probes fails to reveal regions of specific uptake in rhesus macaques. *European Journal of Nuclear Medicine and Molecular Imaging*, 52, 2645–2657. <https://doi.org/10.1007/s00259-025-07110-8>

TEAM WINNER

Predictive/UNC Team



A UNC-based lab and startup using AI to extract real-world human toxicity data, building a reproducible alternative to animal testing for regulatory science.

Team Members: Alexander Tropsha · Eugene Muratov · Ricardo Scheufen Tieghi.

Team Predictive/UNC comprises members of the Laboratory for Molecular Modeling (MML) at the UNC Eshelman School of Pharmacy and of the UNC startup Predictive, LLC, which was cofounded in 2019 by Alexander Tropsha, Eugene Muratov, Vinicius Alves (all of UNC) and Kevin Causey as CEO. Our R&D studies have focused on methodology development and the generation of experimentally testable hypotheses. MML has published over 400 peer-reviewed research papers and developed multiple software tools to support biomedical research with the emphasis on data curation, reproducibility, and rigorous model validation. Predictive, LLC was founded to promote these concepts and commercialize computational technologies and software developed at UNC as in silico NAMs. Our joint team is committed to rigor in the analysis of experimental data reproducibility and computational model validation. We are happy that this commitment is now recognized by the NIH Common Fund's Replication Prize.

1. REPLICATION STRATEGY

1.1 Strategy Overview

A central challenge in replicating biomedical computational research is that failures are often attributed to experimental variability or algorithmic instability, when in fact they originate upstream at the level of data representation and provenance. In computational toxicology and related biomedical AI applications, replication frequently fails not because methods are irreproducible, but because the objects being modeled are not equivalent across analyses. It has been shown that on average, there are two errors per each medicinal chemistry publication with an overall error rate for compounds indexed in one of the medchem databases as high as 8% and that chemical structures are encoded inconsistently, biological annotations are duplicated or discordant, and datasets aggregate heterogeneous sources without transparent reconciliation. Under these conditions, independent teams may appear to “replicate” a study while unknowingly operating on substantively different data. Our replication strategy was designed explicitly to address this hidden but pervasive gap. We adopt a working definition of conceptual replication that is intentionally stricter: replication is achieved when independent researchers, starting from the same declared raw data sources and applying the same transparent curation and modeling workflow, can reproduce both model feasibility and predictive performance within expected uncertainty ranges. Under this definition, a model that cannot be rebuilt, audited, or externally validated after re-curation is not considered replicated, regardless of its originally reported accuracy. Our strategy operates primarily in the basic, translational and pre-clinical categories. Machine learning and computational models are increasingly used to influence regulatory decision-making, prioritization, and replacement of animal tests. These domains are particularly vulnerable to irreproducibility because legacy datasets were never generated with computational reuse in mind. By embedding our protocols early in project development, we have consistently shown that we can greatly increase model reliability, accelerate regulatory acceptance, and scale the external applicability of biomedical research.

1.2 Retrospective Efforts for Computational Research Replicability

We have pioneered the data curation, modeling, and gold standards for establishing replicability in Quantitative Structure Activity Relationship (QSAR) modeling [3–5] and chemical informatics. We suggest the broad implementation of these premier standards for data cleaning and chemical structure standardization, as described in our published chemical and biological data curation workflows [3–5]. Briefly, chemical curation involves steps to remove structures that cannot be appropriately handled by conventional cheminformatics platforms, including the triage of inorganic and organometallic compounds, counterions, salts, and mixtures, the validation of chemical correctness by properties of valence, formal charge, and hybridization state, and normalization of ring

aromatization, enol-keto tautomers, and other specific chemotype substructures (Figure 1). Biological curation looks for duplicates with discordant measurements and suspicious activity values. Major curation steps include (i) structure cleaning and standardization, (ii) rigorous assay duplicate analysis for readout concordance, (iii) evaluation of source trustworthiness, measured by the frequency of incorrect structures, outlier readout values, or duplicate discordance, (iv) verification of the dataset MODelability Index (MODI) [6], and (v) consensus model generation to add level of trust to the readout values in the original chemogenomics data. Our workflow ensures the efficient and thorough elimination of non-standardized chemical structures, duplicate data points, records with unreliable data sources or high experimental variability, outliers, unverified activity cliffs, and incorrectly labeled or unidentifiable compounds (Figure 2).

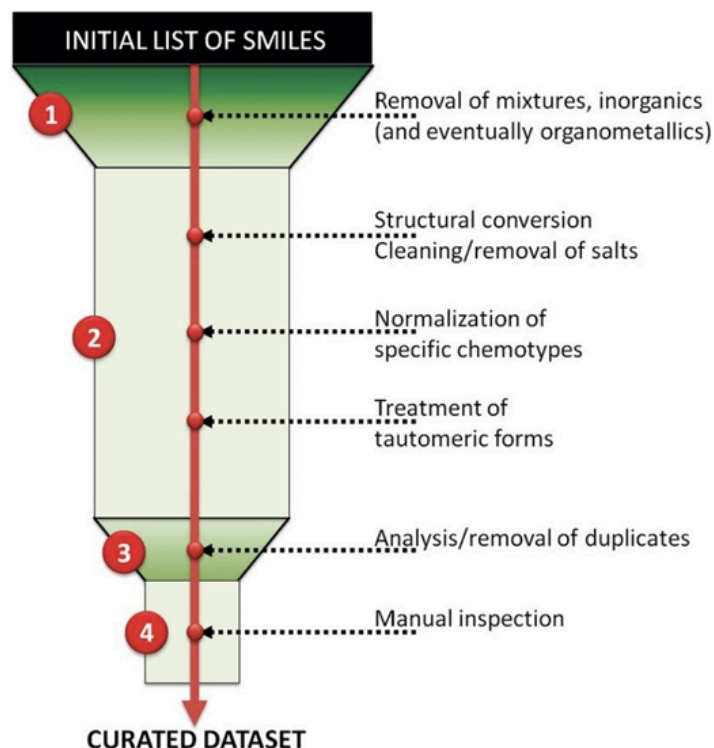


Figure 1. Standard Quantitative Structure-Activity Relationship Modeling Curation Pipeline.

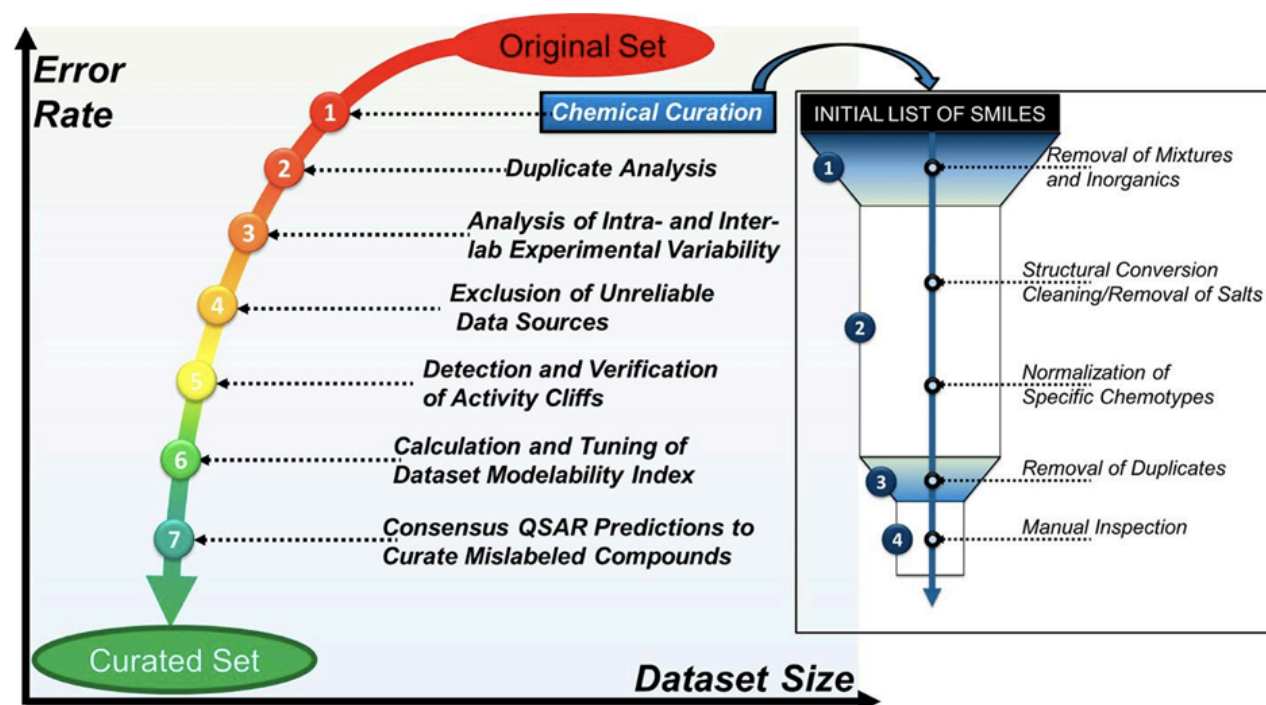


Figure 2. General workflow for comprehensive curation of chemogenomic data sets.

We have benchmarked the best practices of QSAR modeling in a highly cited review that continues to provide commonly acceptable guidance for model development and validation [7]. The key elements of our QSAR modeling workflow are summarized in Figure 3. Each step of the pipeline includes elements of internal or external validation to ensure that the resulting models are rigorously vetted to maximize the external predictivity metrics, such as Correct Classification Rate (CCR) in the case of categorical data or cross-validated R^2 (q^2) for continuous data. Typically, before modeling, a fraction (10-20%) of the curated dataset is withheld from the model-building process as a truly external “Validation” set. The remaining “Modeling” set is further divided into several subsets, subdivided further into “Testing” and “Training” sets (typically also at a 10-20% partition) in which each test set is withheld from modeling and used to test the resulting model. Data is interpreted and used to generate statistical models using modern ML algorithms, such as k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), Random Forest (RF), and other methods including Deep Learning (DL) that we recently adopted for QSAR modeling [8]. As a matter of internal validation, the modeling data set is y-randomized multiple times, to shuffle compound assay readouts. When models built with y-randomized data demonstrate external predictivity, we conclude that the underlying modeling approach leads to overfitting. Models are tested externally using respective withheld subsets and each model is required to pass specific predictivity metrics to be considered acceptable. All validated predictive models can be used (within their Applicability Domain, AD) for assessing novel compounds.

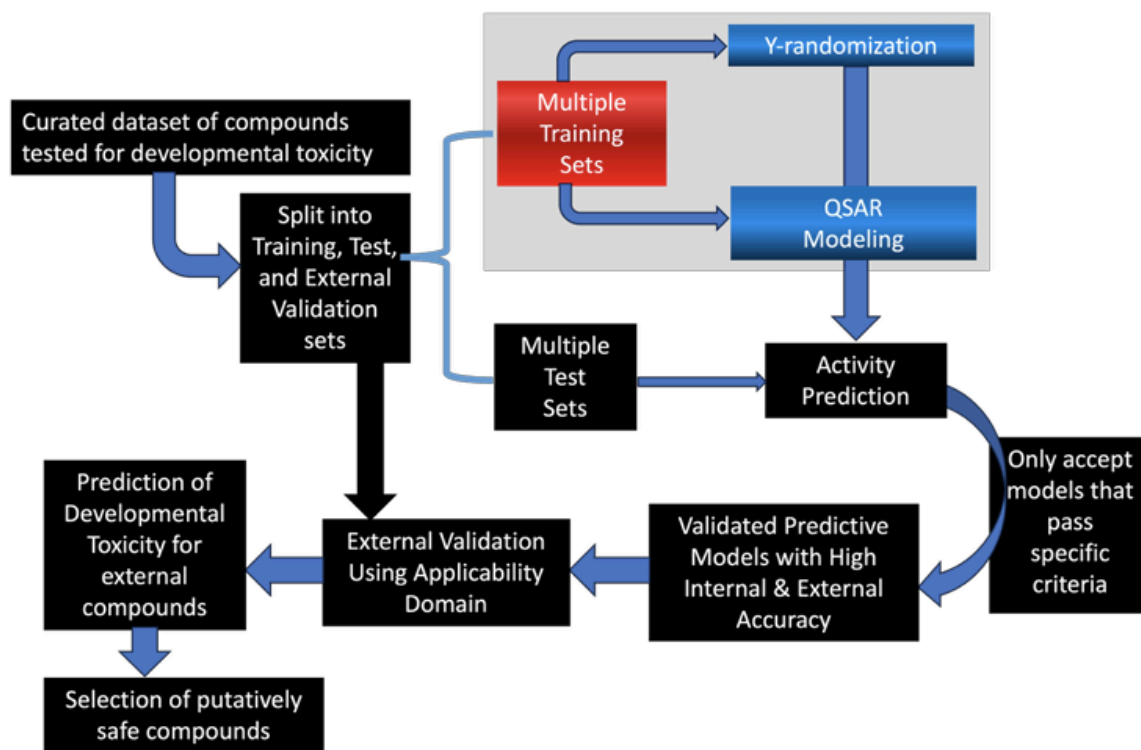


Figure 3. Suggested QSAR Model Development Workflow.

Our foundational approach to data collection, curation, and modeling, has had long-lasting impacts to translational science, becoming the foundational fabric of computational toxicology [9–11], antiviral research [12], material informatics [13], and drug discovery [14,15] projects. We have been adamant in our pursuit of replicable modeling, and have been strong advocates when groups do not develop transparent, reproducible models [16]. We believe our strong dissemination efforts have genuinely inspired the wide-scale adoption of our workflows, which has resulted in strong international collaborations for replicable models, accelerating research reproducibility in cheminformatics. For example, we have been adamant advocates for proper data curation, processing, and modeling, especially in a collaboration between NCATS, NIEHS, and 17 other computational teams to model estrogen receptor activity [17], a standard-setting approach that has resulted in similar efforts, with the same data curation and modeling pipeline given its immense success [18–20].

1.3 Prospective Efforts for Research Replication

As an exemplar study, we have demonstrated the value of data curation and model accuracy replication using the developmental neurotoxicity (DNT) use case. We posit that this study provides guidance to many similar studies focusing on effectiveness and safety of chemical agents across the entire Complement-ARIE Program (<https://commonfund.nih.gov/complementarie>). To comment on the importance of this issue, in high-income countries, approximately 80% of pregnant women take prescription medications, and an even higher proportion take over-the-counter medications [21]. Despite the high usage of medication among pregnant women, drug safety in this population is grossly understudied [22]. For example, clinical trials rarely include pregnant women, leaving knowledge gaps as to the potential toxicity to the fetus and pharmacokinetic differences between the non-pregnant and pregnant populations [23]. Additionally, there remains a lack of sizeable epidemiological cohort studies following children who were exposed to medications in utero [24]. Therefore, pregnant women and their physicians are often faced with the difficult decision of whether to continue medication throughout pregnancy based on limited data. Further investigation into the replicability of assays testing medication safety for pregnant women and the developing fetus is warranted. Therefore, there is a pressing need to find alternative testing strategies and approaches for developmental neurotoxicity testing [25]. Developmental neurotoxicity (DNT) refers

to the impairment of the developing nervous system as a result of early development exposure to environmental hazards.

The complex processes that occur during the early development of the human brain are sensitive to environmental disturbances, resulting in increased risk of irreparable damage due to inhibition or alteration of said processes [26]. There is extensive literature linking motor, cognitive, and behavioral disabilities to exposure to developmental neurotoxicants, including autism spectrum disorder, deficit hyperactivity disorder, and others [27,28]. Over 350,000 chemicals are registered for commercial use worldwide and are commonly present in American households, yet fewer than 1% have ever been tested for developmental neurotoxicity [29,30]. Traditional methods for determining DNT toxicants are time and resource-intensive, using hundreds of rats, rabbits, or mice [31,32], meaning traditional in vivo testing for developmental neurotoxicity is costly and time-consuming [33–35]. In 2018, the US EPA estimated that developmental toxicity and developmental neurotoxicity studies cost between \$128,000 and \$750,000 per compound [33]. Although innovative alternative methods, such as Mouse Ovarian Tumor (MOT), DNT-EST, and HEPM, do not employ animals, the evaluation of a large number of compounds also costs time and money. These methods lack the interaction among different mechanisms, which are crucial to development and are often considered as less relevant for human hazard assessment [36,37]. These animal tests raise ethical concerns, have low reproducibility (~60%), and do not always reliably replicate the human response [31,38]. There is a need to reform the assessment of developmental neurotoxicity, and to improve reproducibility in terms of human translatability and cross-species concordance in DNT assays. These improvements align with recent federal initiatives from the National Institutes of Health (NIH) and the U.S. Food and Drug Administration (FDA), which aim to phase out the use of animals and phase in more ethical, reproducible, and human-relevant models [39,40].

Addressing the challenge of reproducibility in DNT requires learning from published data sources, such as the FDA [41] and TERIS [42], thorough database curation, verification, and computational molecular modeling, which are specific areas of strength for our team [3,4,43]. Such approach provides the means to aid sustainable development of replicable tools to predict the developmental neurotoxicity potential of environmental toxicants, industrial chemicals, and cosmetics [44]. In preliminary studies, we demonstrated that QSAR models of developmental toxicity built with an integrative machine learning approach that employed both human and non-human data were more predictive of the human response than traditional in vivo developmental toxicity assays, and more replicable [44]. Our developmental toxicity approach was the first to develop trimester-specific predictions, based on human data. We plan to leverage our existing replicability and data curation protocols to build a series of new DNT-relevant new approach methods (NAMs), which may serve as candidates for replacement of animal testing. These new series of models will be built to target key events and adverse outcomes in DNT pathways, and will reinforce existing in vitro batteries to assess this complex endpoint [45]. The continued implementation of our data collection, curation, and modeling best practices into future computational models of DNT is poised to significantly accelerate basic and translational research in this space, improving reproducibility in chemical safety assessments [3].

2. ENVIRONMENT

2.1 Environment Description

The standard-setting modeling and data curation approaches developed by our team were conducted within the University of North Carolina (UNC) Eshelman School of Pharmacy. The core team included computational toxicologists, cheminformaticians, and software engineers from the Molecular Modeling Lab and the UNC Eshelman School of Pharmacy. These efforts were a result from over a decade of periodic meetings to develop the optimal standards for the field. We developed the data curation workflows based on years of leadership by Prof. Eugene Muratov and Dr. Alexander Tropsha, specialists in QSAR modeling and its applications to chemical toxicity predictions. Dr. Alexander Tropsha is also CSO of the UNC startup Predictive LLC founded with the goal of commercializing the scientific methods and models developed at UNC. Our Predictive-UNC team is highly collaborative with diverse backgrounds and rich experience in creating in silico tools.

Our paper describing the development of the PreS/MD (Predictor of Sensitization hazard for chemical substances released from Medical Devices), summarizing the successful completion of a Phase 1 SBIR, was regarded by Risk Assessment Specialty Section (RASS) as “Best Published Paper Advancing the Science of Risk Assessment Award” at the 2023 Society of Toxicology (SOT) annual meeting [11]. Additionally, our team has also been recognized for building the first tool using exclusively human data to assess skin sensitization, improving its replicability by creating a public-facing version and enabling comparisons to other animal assays, which received

the Society of Toxicology 2024 Paper of the Year Award in the Alternative Methods category. The Common Fund Replication Prize would enable us to continue to disseminate our curation and modeling workflows, as we have done with materials informatics [46], antiviral research [47], nanomaterials [48], and support expansions into other research spheres, especially in translational science.

2.2 Rigor of Prior Research

Our team has accumulated decades-long experience in developing and implementing the widely accepted best practices for QSAR model development and validation [3–5]. We have used this expertise to build predictive models for various toxicity endpoints, such as skin sensitization [9,49–52], acute systemic toxicity [9,11], carcinogenicity [53–55], and reproductive toxicity [17,18]. Our team has extensive expertise on building user-friendly, web accessible platforms of regulatory relevance for computational modeling and knowledge inference [56–58]. Previously, we developed Chemotext [56], a publicly available online PubMed-mining tool that extracts assertions linking MESH terms, such as drugs, targets, and diseases, from the ~30 million papers annotated in PubMed. As additional evidence, NIH recently expanded the value of the NAMs concept beyond chemical toxicity assessment as a general modern paradigm of biomedical research and created a new Complement-ARIE program to support the development of NAMs. Our team was one of the 20 only winners across the nation of the Complement-ARIE challenge [59]. Given our extensive expertise in developing QSAR models of multiple chemical toxicity endpoints, including several listed in the FDA Guidance document [60], we are confident in our team's ability to continue scaling our curation pipeline for broader application and usability, including for developmental neurotoxicity and its associated mechanisms. Our regulatory grade data curation and modeling practices have found widespread adoption, including from international collaborators in Brazil [50] and Europe [18], but also within U.S. federal agencies such as the National Institute of Environmental Health Sciences and the National Center for Advancing Translational Sciences [61,62].

3. IMPACT

3.1 Replication Activity Impact

Since the original development in 2010s, we have witnessed the large-scale adoption of our QSAR replicability and transparency workflows. The protocols we have developed for proper curation and modeling of chemical bioactivity data have transformed QSAR from a fragile exercise into an approach that can be repeated, audited, and trusted across teams and over time. Post-curation outcomes demonstrated that structured curation is not cosmetic; it changes what can be modeled, how reliably it can be validated, and how model performance can be assessed honestly. In the chemical-structure curation framework, replication impact included measurable gains in model feasibility and predictivity once non-standardized structures and discordant records were removed, and chemotypes were normalized, because these steps prevent descriptor errors, reduce label contradictions, and remove sources of artificial inflation in external performance estimates. Replication outcomes were evaluated using metrics explicitly designed to detect false confidence. These included: (i) feasibility of model construction before and after curation; (ii) stability of external validation metrics across repeated splits; (iii) reduction in sensitivity–specificity imbalance; and (iv) convergence of consensus models built using distinct descriptor spaces and algorithms. In multiple case studies, replication reduced performance variance and eliminated spurious gains driven by duplicate leakage or mislabeled compounds [3–5].

Our structured curation workflows materially increased model quality and replicability, and we advocate for their inclusion as a mandatory prerequisite for rigorous QSAR models that are acceptable to regulators, rather than an optional pre-processing step [5,16]. The post-curation metrics used, demonstrate that structured curation pipelines demonstrably improve reproducibility. In a case study for drug-induced liver injury, we demonstrated how modeling was not possible before curation. In contrast, after curation, a model could be developed with reported 5-fold external cross-validation accuracy in the 62–68% range and external set accuracies reported across 56–73%, anchored by a curated set reduced from 1061 compounds (~10% removed as problematic) [3]. Thus, curation enabled previously “unmodelable” endpoints to become “modellable” within a replicable workflow, allowing independent teams to repeat the pipeline and expect comparable results [3]. Our team has included this standard data curation pipeline in all its published QSAR work since its inception, from endpoint toxicity prediction [9,11,44,63,64] to antiviral research [47], to drug discovery [65,66], and others [67]. By implementing this approach, model performance benchmarks became more balanced: specificity–sensitivity gaps narrowed, and curated models showed closer alignment than non-curated counterparts, reducing the risk of deploying models

that appear strong while systematically failing one class. Thus, the application of standardized, replicable, trustworthy modeling pipelines led to external performance stabilization and reduced bias in class performance, which can otherwise undermine downstream decision-making.

Computational experimental replication is not a private craft, but a transferable standard that all computational modeling groups should apply. We have been consistent advocates for proper model development by developing the protocols for appropriate data collection and curation and by demonstrating the risks of omitting chemical curation and replicability procedures [16]. We continue to publicize our workflows during conference presentations, and have made the data, code, and tools for proper data preprocessing and modeling freely available [51]. With continued dissemination efforts, curation will remain a mandatory prerequisite for trustworthy QSAR, including public calls to improve reproducibility standards in chemogenomic data [3,4,16].

Beyond internal outcomes, the replication strategy materially increased confidence in the team's research outputs. Following the suggested data curation procedures, data preparation steps become explicit and auditable, and lead to improved models development and reception by both academic and regulatory audiences. The same curation-driven workflows have supported models used in regulatory-relevant contexts [68], contributed to international QSAR guidance discussions [69], and informed collaborative efforts evaluated by regulatory agencies [19]. Replication thus functioned not only as quality control, but as a credibility multiplier, enabling broader acceptance of computational evidence in contexts where trust is essential.

3.2 Prospective improvements to research replicability in developmental neurotoxicity

We plan to continue to extend our curation and modeling pipelines to answer complex research questions in developmental neurotoxicity. The Organization for Economic Co-operation and Development (OECD), a consortium of over 38 countries, recently released its initial guidelines for testing DNT with in vitro systems [45]. Our objective is to supplement these initial in vitro assays with more detailed, mechanistic understanding of chemicals affecting brain development. Approaches leveraging machine learning, artificial intelligence, and our detailed curation protocols are poised to significantly accelerate chemical safety assessments and drive a deeper understanding of the key events and adverse outcomes resulting in developmental neurotoxicity. Our suite of machine learning models, dubbed DNT-NAM, will use highly replicable workflows and curated data to predict risks of perturbations to key processes in developmental neurotoxicity, such as cell migration, differentiation, neurulation, and related processes and mechanisms. Overall, using our DNT-NAM approach will ensure that we are accelerating human-relevant biomedical research, using highly effective and scalable technology.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

One of the most persistent scientific barriers encountered during the implementation of our replication strategy was the lack of structural consistency in historical chemical datasets, even among widely cited and ostensibly authoritative sources, such as PubChem and ChEMBL [70,71]. For example, developmental neurotoxicity data is rarely generated under harmonized protocols, and chemical structures are frequently deposited in databases without systematic verification of valence, charge state, tautomeric form, salt composition, or stereochemical correctness. These inconsistencies introduce silent but consequential variability that undermines reproducibility long before any model is trained or evaluated. Early attempts to replicate published QSAR models using nominally identical datasets revealed substantial performance degradation, not due to algorithmic instability, but due to hidden discrepancies in chemical representation and data preprocessing.

At the organizational level, a major barrier to replication was the lack of standardized, shareable data curation workflows, which historically rendered chemical data preparation idiosyncratic and irreproducible across teams. Even when researchers acknowledged the importance of data curation, procedures were often undocumented, manually executed, or embedded in local scripts that could not be independently reconstructed. This barrier limited the ability of external groups to replicate published findings [16], even when datasets and algorithms were publicly available. We addressed this challenge by formalizing chemical and biological curation as a first-class, replicable research output, encapsulated in transparent, tool-agnostic workflows using widely available software. The framework approach shifted curation from an implicit preprocessing step to an explicit methodological component [5].

For upcoming replication work in developmental neurotoxicity, persistent scientific barriers continue to preclude significant advancements in the space. Developing an understanding of developmental neurotoxicity has typically relied on building adverse outcome pathways for developmental toxicity [72]. While heroic and necessary, these efforts typically require dedicated teams of subject matter experts and years of work. In the meantime, we are faced with the frustrating reality that we must test chemicals for developmental safety using animals, primarily rats, rabbits, or mice [73]. Yet, these methods are expensive, time consuming, and raise significant ethical concerns due to the large number of animals tested. Additionally, cross-species analyses have demonstrated that the animal tests have a significant failure of identifying toxic chemicals, with positive predictability rates with human data at only 60% [32]. Thus, we suggest that future efforts leverage our modeling pipelines and knowledge graph approaches [57] to accelerate the discovery of human-relevant DNT mechanisms. Preliminary investigations suggest that by harnessing highly curated data sources of respectable provenance, such as the NIEHS DNT-DIVER initiative [74] and Alliance for Genome Resources, we could significantly accelerate discovery of mechanisms by uncovering previously unlinked chemical-gene-phenotype interactions. Our combined DNT-NAM approach using machine learning and knowledge is a step in the right direction for the continued development of human-relevant, scalable, and replicable technologies at the NIH.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

Data curation and replication activities were conducted progressively between 2010 and 2025. Initial data curation efforts focused on developing and validating chemical, biological, and chemogenomics data curation workflows, demonstrating that “uncurated” datasets fail to support reproducible modeling and that they tend to yield inflated performance estimates. Subsequent phases expanded replication by elucidation of erroneous data points, identification of low quality or irreproducible assays and data sources, etc. Ultimately, replication was no longer treated as a discrete activity but as an intrinsic property of the modeling pipeline, enabling consistent reuse of curated datasets and workflows across endpoints and studies without restarting the process.

5.2 Associated Costs

Costs associated with this research were mostly represented by personnel effort, reflecting the subject matter expertise required for data curation, duplicate resolution, and model verification. A mixed personnel model was employed, combining senior faculty oversight with research professors and postdoctoral fellows. Historically, efforts totaled 15 months in a period between 2010 and 2016 and costs associated with replication activities are estimated at circa \$250,000. This research was funded by the NIH grants (GM66940 and GM96967) obtained by Dr. Tropsha for molecular modeling grants at UNC. The efforts related to conceptual replication proposed in this application, i.e., on developing hazard predictors for binary developmental neurotoxicity risk assessment with the focus on the human outcome, occurred in 2024-2025 and were inspired and partially sponsored by the National Institutes of Health Complement-ARIE challenge prize of \$50,000 obtained by Predictive-UNC Team. The prospective studies described on the development of in silico DNT NAMs focusing on replicable modeling of human outcomes will require circa \$320,000, which are not included in the proposed budget, but will be funded from other sources or conducted in kind.

The overview of both retrospective and prospective timelines can be found in Tables 1 and 2, respectively. The NIH Common Fund replication prize would be used for partial reimbursement (\$50,000) of the expenses required for the development of data curation workflows conducted at Molecular Modeling Lab and sponsored by Dr. Tropsha’s NIH grants, but also for continued dissemination and advocacy efforts for the integration of proper curation and modeling.

Milestones	Q 1	Q 2	Q 3	Q 4
<i>(i) Chemical data curation pipeline</i>	2010	2010		
<i>(ii) Biological data curation pipeline</i>		2015	2015	
<i>(iii) Chemogenomics data curation pipeline</i>				2016
<i>(iv) DeTox NAM</i>	2024	2024	2024	2025

Table 1. Overview of Implementation Timeline (retrospective)

Milestones	Q 1	Q 2	Q 3	Q 4
<i>(i) Data collection, curation, and integration into high-quality curated database.</i>				
<i>(ii) Development and mining of DNT-KG</i>				
<i>(iii) Model development and validation</i>				
<i>(iv) Fully operational DNT NAM</i>				
<i>(v) Interactive user dissemination & feedback</i>				

Table 2. Overview of Implementation Timeline (prospective)

6. COLLABORATION

6.1 Key Collaborations

Successful implementation of the replication activities described in this submission depended on sustained scientific collaboration across institutional boundaries, with a longstanding partnership with the National Institute of Environmental Health Sciences (NIEHS) playing a central role. Rather than functioning as a contractual or transactional collaboration, the NIEHS partnership operated as a scientific alignment around shared concerns: the reproducibility crisis in toxicology, the reliability of computational evidence, and the need for human-relevant alternatives to animal testing that regulators can trust. Over multiple years, this collaboration shaped how data cleaning and standardization was conceptualized, evaluated, and ultimately integrated into routine research practice.

Our group has had over a decade of collaboration with the National Interagency Center for the Evaluation of Alternative Test Methods, at the NIEHS. Together, we have leveraged our chemical curation and modeling protocols to build internationally recognized collaborative modeling projects. These international collaborations have ensured that our modeling approaches are being employed consistently and accurately. The collaboration has resulted in highly replicable models and development pipelines for acute toxicity [19], androgen receptor activity [18], and estrogen receptor activity [17], that are being evaluated by regulatory agencies for its utility and applicability as potential replacement of animal tests [19].

We have also had incredibly fruitful collaborations with the National Center for Advancing Translational Sciences (NCATS) across multiples lies of research from implementing our data curation protocols into drug development workstreams [75], to improving replicability in antiviral [62], to developing novel treatments for bacterial resistance with machine learning models [66]. We plan to continue expanding our collaborations with computational modeling to ensure curation and modeling standards are maintained for decades to come.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Current Sustainability Plans

Unlike experimental replication, which incurs substantial recurring costs and infrastructure demands, computational replication through data curation scales efficiently. Once workflows are established, the marginal cost of replication is substantially decreased, and replication can be performed repeatedly across new datasets with minimal additional effort. This cost structure enables replication at a scale that is impractical for traditional methods, particularly in chemical safety assessment where thousands of substances require evaluation.

FAIR Principles. By virtue of the infrastructure described, we ensure that the data curation pipelines and tools are inherently FAIR, meaning: findable through automated indexing and metadata creation; accessible through multiple interfaces that can be automatically deployed; interoperable through transformation to a common community-driven data model; and reusable through automated provenance building and modular design approach. Transferability was a guiding principle in the design of the replication strategy. Although initially applied to developmental neurotoxicity and related toxicological endpoints, the underlying replication framework is endpoint-agnostic. The same chemical structure standardization, duplicate analysis, and dataset verification procedures can be applied to other areas of toxicology, pharmacology, and chemical safety assessment. This generalizability is particularly important for fields that rely on legacy data generated under heterogeneous conditions, where irreproducibility often stems from inconsistencies in data provenance rather than from the biological complexity of the endpoint itself. By abstracting replication away from any single assay or outcome, the approach enables adoption across sectors, including academia, regulatory agencies, and industry, without requiring specialized infrastructure or proprietary software. All the code and software used to deploy our standard data curation and modeling pipelines are described extensively in our prior publications [3,9,51]. The pipelines leverage a combination of Python packages and open-source tools, with extensive documentation, for all steps, from structural chemical curation to large-scale machine learning modeling.

8. REFERENCES

1. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*. 2018;23:1538–1546.
2. Olah M, Rad R, Ostopovici L, Bora A, Hadaruga N, Hadaruga D, Moldovan R, Fulias A, Mracec M, Oprea T. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In: *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*. New York: Wiley; 2007.
3. Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf Model*. 2010;50:1189–1204. doi: 10.1021/ci100176x.
4. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model*. 2016;56:1243–1252. doi: 10.1021/acs.jcim.6b00129.
5. Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf Model*. 2010;50:1189–1204. doi: 10.1021/ci100176x.
6. Golbraikh A, Muratov E, Fourches D, Tropsha A. Data Set Modelability by QSAR. *J Chem Inf Model*. 2014;54:1–4. doi: 10.1021/ci400572x.
7. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform*. 2010;29:476–488. doi: 10.1002/minf.201000061.
8. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv*. 2018;4:eaap7885. doi: 10.1126/sciadv.aap7885.
9. Borba JVB, Alves VM, Braga RC, Korn DR, Overdahl K, Silva AC, Hall SUS, Overdahl E, Kleinstreuer N, Strickland J. SToxTox: An in Silico Alternative to Animal Testing for Acute Systemic and Topical Toxicity. *Env Health Perspect*. 2022;130:027012. doi: 10.1289/EHP9341.
10. Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminformatics*. 2018;10:10. doi: 10.1186/s13321-018-0263-1.

11. Alves VM, Borba JVB, Braga RC, Korn DR, Kleinstreuer N, Causey K, Tropsha A, Rua D, Muratov EN. PreS/MD: Predictor of Sensitization Hazard for Chemical Substances Released From Medical Devices. *Toxicol Sci.* 2022;189:250. doi: 10.1093/toxsci/kfac078. PMID: 35916740.
12. Tarasova OA. Current Perspectives in Antiviral Research. *Int J Mol Sci.* 2023;24:14555. doi: 10.3390/ijms241914555. PMID: 37833992.
13. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature.* 2018;559:547–555. doi: 10.1038/s41586-018-0337-2. PMID: 30046072.
14. Nittinger E, Clark A, Gaulton A, Zdrazil B. Biomedical data analyses facilitated by open cheminformatics workflows. *J Cheminformatics.* 2023;15:46. doi: 10.1186/s13321-023-00718-8. PMID: 37069670.
15. Waitman KB, Martin H-J, Carlos JAEG, Braga RC, Souza VAM, Melo-Filho CC, Hilscher S, Toledo MFZJ, Tavares MT, Costa-Lotufo LV, et al. Dona Flor and her two husbands: Discovery of novel HDAC6/AKT2 inhibitors for myeloid cancer treatment. *Comput Biol Med.* 2025;197:110990. doi: 10.1016/j.compbimed.2025.110990. PMID: 40882477.
16. Alves V, Borba J, Capuzzi S, Muratov E, Andrade C, Rusyn I, Tropsha A. Oy Vey! A Comment on "Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships Outperforming Animal Test Reproducibility." *Toxicol Sci Off J Soc Toxicol.* 2019;167:3–4. doi: 10.1093/toxsci/kfy286.
17. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ Health Perspect.* 2016;124:1023–1033. doi: 10.1289/ehp.1510267.
18. Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, Andrade CH, Bai F, Balabin I, Ballabio D, et al. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ Health Perspect.* 2020;128:027002. doi: 10.1289/EHP5580. PMID: 32074470.
19. Mansouri K, Karmaus AL, Fitzpatrick J, Patlewicz G, Pradeep P, Alberga D, Alepee N, Allen TEH, Allen D, Alves VM, et al. CATMoS: Collaborative Acute Toxicity Modeling Suite. *Environ Health Perspect.* 2021;129:047013. doi: 10.1289/EHP8495. PMID: 33929906.
20. Sessions Z, Bobrowski T, Martin H-J, Beasley J-MT, Kothari A, Phares T, Li M, Alves VM, Scotti MT, Moorman NJ, et al. Praemonitus praemunitus: can we forecast and prepare for future viral disease outbreaks? *FEMS Microbiol Rev.* 2023;47:fuad048. doi: 10.1093/femsre/fuad048.
21. Daw JR, et al. Prescription drug use during pregnancy in developed countries: a systematic review. *Pharmacoepidemiol Drug Saf.* 2011 [cited 2023 Dec 19]. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/pds.2184>.
22. The safety of asthma medications during pregnancy and lactation: Clinical management and research priorities. ScienceDirect [Internet]. [cited 2024 Jan 8]. Available from: <https://www.sciencedirect.com/science/article/pii/S0091674921003699>.
23. Waitt C, Astill D, Zavala E, Karron RA, Faden RR, Stratton P, Temkin SM, Clayton JA. Clinical trials and pregnancy. *Commun Med.* 2022;2:1–5. doi: 10.1038/s43856-022-00198-1.
24. Risk of neonatal drug withdrawal after intrauterine co-exposure to opioids and psychotropic medications: cohort study. *The BMJ* [Internet]. [cited 2024 Jan 8]. Available from: <https://www.bmj.com/content/358/bmj.j3326.full>.
25. Seiler A, Oelgeschläger M, Liebsch M, Pirow R, Riebeling C, Tralau T, Luch A. Developmental toxicity testing in the 21st century: the sword of Damocles shattered by embryonic stem cell assays? *Arch Toxicol.* 2011;85:1361–1372. doi: 10.1007/s00204-011-0767-1.
26. Grandjean P. Only One Chance: How Environmental Pollution Impairs Brain Development—and How to Protect the Brains of the Next Generation. Oxford University Press; 2015.
27. Rossignol DA, Genuis SJ, Frye RE. Environmental toxicants and autism spectrum disorders: a systematic review. *Transl Psychiatry.* 2014;4:e360. doi: 10.1038/tp.2014.4.
28. Ye BS, Leung AOW, Wong MH. The association of environmental toxicants and autism spectrum disorders in children. *Environ Pollut.* 2017;227:234–242. doi: 10.1016/j.envpol.2017.04.039.
29. Sachana M, Shafer TJ, Terron A. Toward a Better Testing Paradigm for Developmental Neurotoxicity: OECD Efforts and Regulatory Considerations. *Biology* [Internet]. 2021 [cited 2025 Dec 18];10. doi: 10.3390/biology10020086.
30. US EPA. Evaluating Developmental Neurotoxicity Hazard: Better than Before [Internet]. 2017 [cited 2025 Dec 18]. Available from: <https://www.epa.gov/sciencematters/evaluating-developmental-neurotoxicity-hazard-better>.
31. Niethammer M, Burgdorf T, Wistorf E, Schönfelder G, Kleinsorge M. In vitro models of human development and their potential application in developmental toxicity testing. *Development.* 2022;149:dev200933. doi: 10.1242/dev.200933.

32. Bailey J. Developmental Toxicity Testing: Protecting Future Generations? *Altern Lab Anim.* 2008;36:718–721. doi: 10.1177/026119290803600618.
33. US EPA. Cost Estimates of Studies Required for Pesticide Registration [Internet]. 2018 [cited 2023 Dec 30]. Available from: <https://www.epa.gov/pesticide-registration/cost-estimates-studies-required-pesticide-registration>.
34. Van Norman GA. Limitations of animal studies for predicting toxicity in clinical trials: is it time to rethink our current approach? *JACC Basic Transl Sci.* 2019;4:845–854.
35. Dubovický M, Kovačovský P, Ujházy E, Navarová J, Brucknerová I, Mach M. Evaluation of developmental neurotoxicity: some important issues focused on neurobehavioral development. *Interdiscip Toxicol.* 2008;1:206–210.
36. Faustman EM. Short-term tests for teratogens. *Mutat Res.* 1988;205:355–384. doi: 10.1016/0165-1218(88)90027-4.
37. Steele VE, Morrissey RE, Elmore EL, Gurganus-Rocha D, Wilkinson BP, Curren RD, Schmetter BS, Louie AT, Lamb JC, Yang LL. Evaluation of two in vitro assays to screen for potential developmental toxicants. *Fundam Appl Toxicol.* 1988;11:673–684. doi: 10.1016/0272-0590(88)90130-3.
38. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature.* 2023;616:673–685.
39. National Institutes of Health. NIH to prioritize human-based research technologies [Internet]. [cited 2025 June 26]. Available from: <https://www.nih.gov/news-events/news-releases/nih-prioritize-human-based-research-technologies>.
40. U.S. Food and Drug Administration. FDA Announces Plan to Phase Out Animal Testing Requirement for Monoclonal Antibodies and Other Drugs [Internet]. 2025 [cited 2025 June 26]. Available from: <https://www.fda.gov/news-events/press-announcements/fda-announces-plan-phase-out-animal-testing-requirement-monoclonal-antibodies-and-other-drugs>.
41. U.S. Food and Drug Administration [Internet]. 2024 [cited 2024 Jan 8]. Available from: <https://www.fda.gov/>.
42. TERIS [Internet]. [cited 2023 Dec 30]. Available from: <https://deohs.washington.edu/teris/>.
43. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol.* 2015;11:535. doi: 10.1038/nchembio.1881.
44. Tieghi RS, Rath M, Moreira-Filho JT, Wellnitz J, Martin H-J, Gates K, Hogberg HT, Kleinstreuer N, Tropsha A, Muratov EN. DeTox: an In-Silico Alternative to Animal Testing for Predicting Developmental Toxicity Potential. *Environ Health Perspect* [Internet]. [cited 2025 June 9]. doi: 10.1289/EHP15307.
45. OECD. Initial Recommendations on Evaluation of Data from the Developmental Neurotoxicity (DNT) In-Vitro Testing Battery [Internet]. OECD; 2023 [cited 2024 Aug 13]. Available from: https://www.oecd-ilibrary.org/environment/initial-recommendations-on-evaluation-of-data-from-the-developmental-neurotoxicity-dnt-in-vitro-testing-battery_91964ef3-en.
46. Hart M, Idanwekhai K, Alves VM, Miller AJM, Dempsey JL, Cahoon JF, Chen C-H, Winkler DA, Muratov EN, Tropsha A. Trust Not Verify? The Critical Need for Data Curation Standards in Materials Informatics. *Chem Mater.* 2024;36:9046–9055. doi: 10.1021/acs.chemmater.4c00981.
47. Martin H-J, Melo-Filho CC, Zakharov AV, Muratov E, Tropsha A. On the importance of data curation for knowledge mining in antiviral research. *Sci Prog.* 2025 [cited 2025 Dec 18]. Available from: <https://journals.sagepub.com/doi/full/10.1177/00368504241301535>.
48. Tropsha A, Mills KC, Hickey AJ. Reproducibility, sharing and progress in nanomaterial databases. *Nat Nanotechnol.* 2017;12:1111–1114. doi: 10.1038/nnano.2017.233.
49. Alves VM, Capuzzi SJ, Braga RC, Borba JVB, Silva AC, Luechtefeld T, Hartung T, Andrade CH, Muratov EN, Tropsha A. A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustain Chem Eng.* 2018;6:2845–2859. doi: 10.1021/acssuschemeng.7b04220.
50. Borba JVB, Braga RC, Alves VM, Muratov EN, Kleinstreuer N, Tropsha A, Andrade CH. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem Res Toxicol.* 2021;34:258–267. doi: 10.1021/acs.chemrestox.0c00186.
51. Scheufen Tieghi R, Moreira-Filho JT, Martin H-J, Wellnitz J, Otoch MC, Rath M, Tropsha A, Muratov EN, Kleinstreuer N. A Novel Machine Learning Model and a Web Portal for Predicting the Human Skin Sensitization Effects of Chemical Agents. *Toxics.* 2024;12:803. doi: 10.3390/toxics12110803.
52. Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. *Toxicol Appl Pharmacol.* 2015;284:273–280. doi: 10.1016/j.taap.2014.12.013.

53. Zhu H, Rusyn I, Richard A, Tropsha A. Use of Cell Viability Assay Data Improves the Prediction Accuracy of Conventional Quantitative Structure–Activity Relationship Models of Animal Carcinogenicity. *Environ Health Perspect*. 2008;116:506–513. doi: 10.1289/ehp.10573. PMID: 18414635.
54. Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. The use of dose-response qHTS data as biological descriptors improves the prediction accuracy of QSAR models of acute rat toxicity. *Env Health Perspect*. 2011;119:364–370.
55. Predictive Modeling of Chemical Hazard by Integrating Numerical Descriptors of Chemical Structures and Short-term Toxicity Assay Data. *Toxicological Sciences* [Internet]. [cited 2025 Dec 19]. Available from: <https://academic.oup.com/toxsci/article-abstract/127/1/1/1641504>.
56. Capuzzi SJ, Thornton TE, Liu K, Baker N, Lam WI, O'Banion CP, Muratov EN, Pozefsky D, Tropsha A. Chemotext: a publicly available web server for mining drug–target–disease relationships in PubMed. *J Chem Inf Model*. 2018;58:212–218.
57. Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, Morton K, Fecho K, Tropsha A. ROBOKOP KG and KGB: integrated knowledge graphs from federated sources. *J Chem Inf Model*. 2019;59:4968–4973.
58. Beasley J-MT, Korn DR, Tucker NN, Alves ETM, Muratov EN, Bizon C, Tropsha A. ExEmPLAR (Extracting, Exploring, and Embedding Pathways Leading to Actionable Research): a user-friendly interface for knowledge graph mining. *Bioinformatics*. 2024;40:btad779. doi: 10.1093/bioinformatics/btad779.
59. NIH Common Fund. Complement-ARIE Challenge Prize Winner Summaries [Internet]. 2024 [cited 2025 Oct 27]. Available from: <https://commonfund.nih.gov/complementarie/challengewinnersummaries>.
60. U.S. Food and Drug Administration, Center for Devices and Radiological Health. Use of International Standard ISO 10993-1, "Biological evaluation of medical devices - Part 1: Evaluation and testing within a risk management process" [Internet]. FDA; 2023 [cited 2025 Dec 19]. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-international-standard-iso-10993-1-biological-evaluation-medical-devices-part-1-evaluation-and>.
61. Alves VM, Auerbach SS, Kleinstreuer N, Rooney JP, Muratov EN, Rusyn I, Tropsha A, Schmitt C. Curated data in—trustworthy in silico models out: the impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing. *Altern Lab Anim*. 2021;49:73–82.
62. Martin H-J, Melo-Filho CC, Zakharov AV, Muratov E, Tropsha A. On the importance of data curation for knowledge mining in antiviral research. *Sci Prog*. 2025;108:00368504241301535. doi: 10.1177/00368504241301535.
63. HuSSPred: Human Skin Sensitization Predictions [Internet]. 2024 [cited 2024 Nov 25]. Available from: <https://husspred.mml.unc.edu/>.
64. Tieghi RS, Melo-Filho CC, Martin H-J, Moreira-Filho JT, LaPratt T, Allen D, Strickland J, Tropsha A, Kleinstreuer N, Muratov EN. External Validation of SToxTox – Novel Alternative Method (NAM) for Acute Systemic and Topical Toxicity. *Environ Health Perspect* [Internet]. [cited 2025 Nov 26]. doi: 10.1289/EHP16647.
65. Beasley J-MT, Korn DR, Popov KI, Dumpf RL, Sessions ZL, Rath MK, Alves VM, Causey K, Rua D, Muratov EN, et al. Integrated approach to elucidate metal-implant related adverse outcome pathways. *Regul Toxicol Pharmacol*. 2022;136:105277. doi: 10.1016/j.yrtph.2022.105277.
66. Beasley J-M, Dorjsuren D, Jain S, Rath M, Tieghi RS, Tropsha A, Simeonov A, Zakharov AV, Muratov E. Breaking the Phalanx: Overcoming Bacterial Drug Resistance with Quorum Sensing Inhibitors that Enhance Therapeutic Activity of Antibiotics [Internet]. *bioRxiv*; 2025 [cited 2025 Jan 23]. p. 2025.01.17.633658. Available from: <https://www.biorxiv.org/content/10.1101/2025.01.17.633658v1>.
67. Rath M, Wellnitz J, Martin H-J, Melo-Filho C, Hochuli J, Silva GM da, Beasley J-M, Maxfield T, Sessions Z, Popov K, et al. Novel Pharmacokinetics Profiler (PhaKinPro): Model Development, Validation, and Implementation as a Web-Tool for Triaging Compounds with Undesired PK Profiles [Internet]. *ChemRxiv*; 2023 [cited 2023 May 30]. Available from: <https://chemrxiv.org/engage/chemrxiv/article-details/6413d7a4aad2a62ca1e32e07>.
68. Hong H, Chen M, Ng HW, Tong W. QSAR Models at the US FDA/NCTR. *Methods Mol Biol*. 2016;1425:431–459. doi: 10.1007/978-1-4939-3609-0_18. PMID: 27311476.
69. OECD. (Q)SAR Assessment Framework: Guidance for Assessing (Q)SAR Models and Predictions [Internet]. 2023 [cited 2025 Dec 18]. Available from: <https://www.oecd.org/en/events/2023/11/qsar-assessment-framework-guidance-for-assessing-qsar-models-and-predictions.html>.
70. PubChem [Internet]. [cited 2024 Nov 10]. Available from: <https://pubchem.ncbi.nlm.nih.gov/>.
71. ChEMBL Database [Internet]. [cited 2024 Jan 8]. Available from: <https://www.ebi.ac.uk/chembl/>.

72. Edwards SW, Tan Y-M, Villeneuve DL, Meek ME, McQueen CA. Adverse Outcome Pathways—Organizing Toxicological Information to Improve Decision Making. *J Pharmacol Exp Ther*. 2016;356:170–181. doi: 10.1124/jpet.115.228239.
73. OECD. Test No. 426: Developmental Neurotoxicity Study [Internet]. Paris: Organisation for Economic Co-operation and Development; 2007 [cited 2024 Nov 19]. Available from: https://www.oecd-ilibrary.org/environment/test-no-426-developmental-neurotoxicity-study_9789264067394-en.

TEAM WINNER

SOAR (Spatial Omics Analysis Replication)



A University of Washington and Fred Hutch collaboration building standardized lab protocols and open-source software to make spatial proteomics research reproducible across institutions.

Team Members: *University of Washington (UW) Tacoma:* Dr. Ka Yee Yeung-Rhee · Dr. Ling-Hong Hung and Mr. Bryce Fukuda. *School of Engineering and Technology. Fred Hutch Cancer Center (FHCC):* Dr. Cecilia Yeung · Ms. Kimberly Smythe · and Dr. Jocelyn Wright.

The FHCC team consists of Dr. Cecilia Yeung, Ms. Kimberly Smythe, and Dr. Jocelyn Wright. Cecilia Yeung, MD, is a pathologist and medical director of the Clinical Testing Labs at FHCC. Dr. Yeung conceptualized the replication study, and her lab developed the TMAs. Ms. Smythe has a BS in Biochemistry and is the Head of Translational Pathology Lab at FHCC which generated spatial proteomics images. Dr. Wright has a PhD in experimental pathology and extensive experience with spatial proteomic data analysis and worked with engineers to develop the analysis workflow.

The UW team consists of Dr. Ka Yee Yeung-Rhee, Dr. Ling-Hong Hung and Mr. Bryce Fukuda. Dr. Yeung-Rhee has a Ph.D. in Computer Science. She contributed to the conceptualization of the replication study, management of the project, and collaborative activities. Dr. Hung has a Ph.D. in Biochemistry. He contributed to the design and development of containerized analytical workflows that support graphical output. Mr. Fukuda has a Master's degree in computer science and systems. He implemented and tested containerized software tools to reproducibly execute bioinformatics pipelines. He worked closely with the FHCC team in data analysis.

1. REPLICATION STRATEGY

1.1 Strategy Description

Spatial proteomics provides a spatially resolved view of protein expression and localization within cells and tissues by mapping the location and abundance of many proteins simultaneously, and was selected as Method of the Year 2024 by Nature Methods [1]. The clinical impact of spatial proteomics includes the identification of spatial biomarkers associated with disease states that can be subsequently used to guide personalized treatment strategies and patient stratification [2]. Techniques such as co-detection by indexing (CODEX), run on the PhenoCycler-Fusion (PCF) platform, and Lunaphore COMET, are currently among the leading spatial proteomics imaging approaches. Analysis of protein staining must consider not only expression level of a biomarker, but also the location of the staining (nuclear, cytoplasmic, or membranous), the staining pattern (dot like, linear, contiguous, diffuse and homogenous), and integration of cell morphology to allow for correct interpretation. There is a need for a systematic approach with built-in quality assurance/control (QA/QC) to ensure reproducibility of spatial proteomics studies.

Analyzing these multi-modal spatial proteomics data requires an analytical workflow or pipeline consisting of multiple software tools. A quality control (QC) pre-processing step is essential for data validation particularly in proteomic data using antibody staining markers to verify the correct staining pattern in terms of cell compartment and marker co-expression on each image. For cell type identification/annotation, analysis of protein staining must consider not only the expression level of biomarkers but also the integration of cell morphology for correct interpretation. Without these interactive steps, the results will not be valid and can result in erroneous conclusions. Therefore, the transition of this powerful data type into translational research is hampered by the need for systematic quality assessment in laboratory protocols, as well as easy-to-use end-to-end computational tools that support reproducible multi-modal analysis and interactive user input.

In this report, we present our team's efforts in replication studies using spatial proteomics data by: 1) Developing robust laboratory protocols that include instrument verification, assay validation, and quality assurance/control (QA/QC) to ensure reproducibility [3], and 2) Re-analyzing spatial proteomics datasets using alternative computational tools [4].

1.2 Laboratory Protocols with QA/QC

Our team at the Fred Hutchinson Cancer Center (FHCC) developed a multi-cancer tissue microarray (TMA) consisting of tonsil and various tumor types generated for PhenoCycler high-plex immunohistochemistry (IHC). The PhenoCycler-Fusion platform utilizes a self-contained, cyclic workflow to visualize 56 markers on a single 3–4 μm section of tissue or TMA sectioned onto a standard positively charged slide. It employs barcode-conjugated antibodies against protein targets of interest followed by an iterative process of revealing up to three antibodies at a time via fluorescently tagged reporters. The stain fixation and image acquisition steps were performed using Akoya's PhenoCycler-Fusion protocols. The images were acquired with the DAPI, Cy3, Cy5, and Cy7 filters and processed into stitched QPTIFF images. The TMA contained both normal (tonsils) and tumor tissues consisting of common and antigenically diverse cancers including breast adenocarcinoma (BRC), non-small cell lung carcinoma (NSCLC), colorectal cancer (CRC), melanoma (MEL), hepatocellular carcinoma (HCC), and Merkel cell carcinoma (MCC). To study replicability of our laboratory protocols, 2 technical replicates of TMAs (labelled as TMA-A1 and TMA-A2) were generated.

Primary PhenoCycler QA/QC involves confirming there were no equipment failures during the run and performing a precursory visual check of the QPTIFF image to verify that tissues were in focus with minimal staining artifacts and exposures issues. TMA cores and regions of interest (ROI) that passed the preliminary QC step were selected for cellular analysis. A pathologist manually performed a qualitative assessment of segmentation accuracy. Quality metrics for PhenoCycler images include comparison of the cell counts and densities for the duplicate tissues between the two TMAs. We showed that the densities between tissue-type duplicates were consistent regardless of differences between tissue sizes. Next, we verified PhenoCycler markers by comparing each of our 56 markers' observed expression to the expected expression profiles. We showed that 62.5% of the markers performed as per the expected annotations [3].

1.3 Data Analysis

Each step of the analysis was performed using a different computational method and involved user interactions. StarDist [5] for QuPath [6] was used to detect (segment) the individual cells to generate a segmentation image for each tissue and the cell object data for each cell containing x, y coordinates and mean cell fluorescence pixel intensity (MFI) values for each marker. Empty barcodes were included in the staining panel to identify non-specific staining of "sticky" cells and areas. The resulting empty barcode MFIs could be used to exclude false-positive cells. An anatomic pathologist reviewed the segmentation images to visually verify the quality of the cell detections, and cell object data were exported for further analysis. Each marker stain was also manually reviewed to assess quality (positive signal above background) and compared to expected expression. Uniform Manifold Approximation and Projection (UMAP) projections and unsupervised clustering using Leiden were computed. Cell-type clusters were manually labeled by inspecting the pattern of differentially expressed proteins across clusters. Mean cell intensities were batch-corrected by fitting independent two-component Log-Normal mixture models to each marker across tonsil cores and scaling marker intensities by the ratio between the respective mean values of the lower mixture-model components. These adjusted marker intensities were then log-transformed and visualized by projecting into two-dimensions using UMAP.

1.4 Alternative End-to-End Computational Pipeline

While there are many workflow execution engines to run bioinformatics pipelines, most of these are command line-based tools, and thus not designed for graphical and interactive workflows using spatial proteomics data. We implemented a modular and interactive graphical platform that empowers biomedical researchers to reproducibly execute complex spatial proteomic analysis in an automated and end-to-end reproducible workflow that integrates multiple open-source software tools and custom scripts through containerization [4]. **Figure 1** shows a screenshot of the workflow.



Figure 1. A screenshot showing the high-throughput graphical containerized workflow consisting of modular widgets. Steps of the workflow include visualization of the stained image for assessing quality control of individual stains using QuPath; segmentation of cells into cell data units using StarDist in QuPath; exporting these data to perform unsupervised clustering of cell types using custom scripts in a Jupyter notebook; creation of masks to be superimposed onto the image for visual validation; and generation of data summaries such as heatmaps, sortable spreadsheets, and dimension reduction techniques (e.g., UMAP) to facilitate cell-type annotation and distribution between samples. This figure was originally published in GigaByte [4].

Containerization enables reproducibility and portability by bundling a software application, along with all its dependencies and specific configurations, into a single file known as a container image. This allows the software to run consistently across various computing environments. Our spatial proteomics imaging workflow supports graphical output and is packaged in modular version-controlled software containers (such that each step is encapsulated in an independent container), ensuring reproducibility, flexibility and interoperability. Our end-to-end workflow includes visualization of the stained image for assessing quality control of individual stains using QuPath; segmentation of cells into cell data units using StarDist [5] in QuPath; exporting these data to perform cell-level normalization across biomarkers, batch correction, and unsupervised clustering of cell types using custom scripts in a Jupyter notebook; creation of masks to be superimposed onto the image for visual validation; and generation of data summaries such as heatmaps, sortable spreadsheets, and dimension reduction techniques (e.g., UMAP) to facilitate cell-type annotation and distribution between samples (see **Figure 2** for example results). Our workflow is customizable to allow changes in user defined analysis parameters, inclusion of additional analysis modalities, and integration with other software for spatial analysis of cell types.

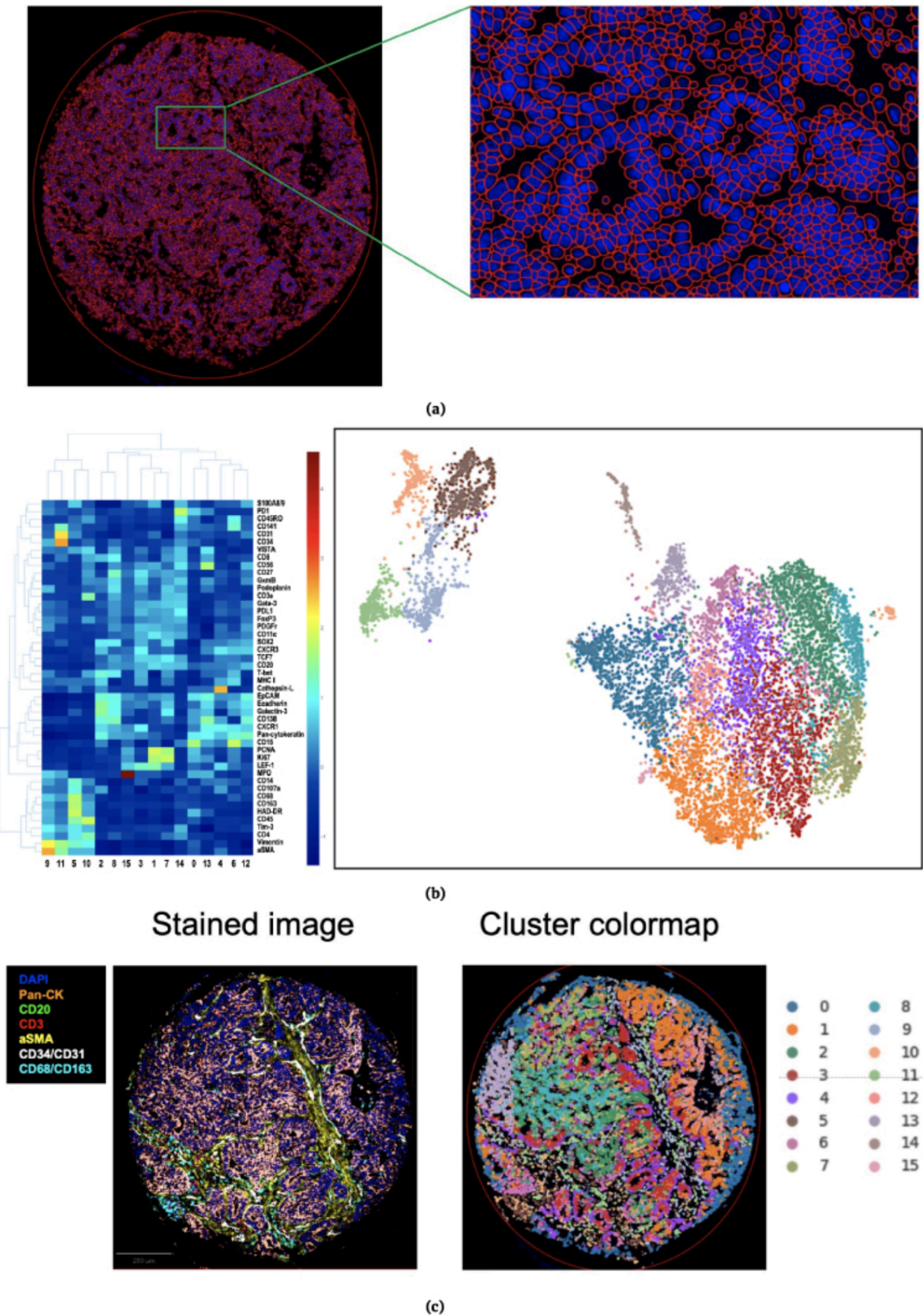


image in QuPath illustrating the output from our containerized workflow on a colorectal cancer region from a stained TMA image. Results of cell segmentation are shown as red cell outlines surrounding DAPI (blue) stained cell nuclei that define the cell data objects. (b) Results of Leiden unsupervised clustering of exported cell data from the segmented image as a heatmap (left) and as a UMAP (right). (c) The results of importing clustering results back into QuPath and displaying it as a colormap over the original colorectal cancer tissue image. This figure was originally published in GigaByte [4].

1.5 Innovation

A key feature of our approach is the flexibility to integrate open-source tools that are widely adopted by the bioimage community, thus reducing barriers of adoption by the research community. This desirable feature is illustrated in the integration of interactive image analysis using QuPath and Fiji [7] in our prototype. Our analytical workflow is customizable to allow changes in user-defined analysis parameters or inclusion of additional analysis modalities for spatial analysis of cell types. This feature enables systematic and automated tuning of input parameters to enhance performance of analytical workflows. Additionally, our end-to-end software platform enables users to interact with graphical visualizations and provide domain knowledge to guide the analyses. Most workflow execution engines focus on automating, orchestrating, and optimizing the execution of complex tasks but often lack native support for graphical outputs, e.g. Nextflow [8] and Snakemake [9]. However, support for interactive graphical analysis is essential to facilitate interpretation of imaging data. Importantly, we support on-prem analysis such that our platform can be deployed locally, on internal computer servers or on the cloud such that data security and compliance can be directly managed by the user. This contrasts with other commercial solutions in which sensitive patient data must be uploaded and analyzed on an external proprietary server for which the user has limited control.

1.6 Working Definition

We define “replicability” as achieving consistent results across repeated laboratory procedures and applying alternative computational tools to analyze the data.

Category Identification: Translational

2. ENVIRONMENT

2.1 Activity Context

This replication project is part of an ongoing collaboration between the Fred Hutchinson Cancer Center (FHCC) and the University of Washington (UW) Tacoma. The replication project originally resulted after publication of a manuscript which compared the 2 replicate TMAs for spatial data produced from 10x Xenium (transcriptomics) to Phenocycler (proteomics) was conducted that included assessment of QA/QC parameters [3]. In this project, we show the replicability of spatial proteomics results when analyzed by a different algorithm and data analysis pipeline.

2.2 Location, Timing, and Team Details

All wet lab experiments were performed at the Fred Hutchinson Cancer Center (FHCC) in Seattle, WA. The containerized end-to-end pipeline that supports graphical output and user interactions was developed at University of Washington (UW) Tacoma, Tacoma, WA.

The FHCC and the UW team are identified at the beginning of this write-up.

2.3 Funding Support

This work was supported by funding from the National Institutes of Health (NIH), National Cancer Institute (NCI) U2C CA271902, NIH, NCI R01 CA264646, NIH, NCI P01 CA225517, the NIH Human Immunology Project Consortium U19AI128914, and the Immunotherapy Integrated Research Center of the Fred Hutchinson Cancer Center.

3. IMPACT

3.1 Replication Activity Impact

Our application addresses current limitations to clinical implementation of spatial proteomic work, namely the lack of pre-analytical assessment for quality tissues, the need for a systematic method to select regions of interest while retaining comparable morphology, and the need for easy-to-use end-to-end computational tools that support reproducibility, trackability, and interactive user input. Our laboratory protocols with QA/QC can be used by other labs to facilitate reproducibility. Our end-to-end containerized pipeline allows users to incorporate software tools that best fit each analytical task and to leverage existing software tools developed for omics data when analyzing multi-modal data. Additionally, our tools are publicly available, can be adopted and customized by other labs.

3.2 Long-Term Impact

Our laboratory protocols with quality assessment can be used to enhance replicability of spatial proteomics data generation. Our end-to-end and interactive imaging workflows for spatial proteomic analysis can empower biomedical researchers to reproducibly customize, execute, and disseminate complex analyses.

3.3 Research Rigor Improvements

Our open-source software enhances transparency and trust in analysis results.

3.4 Dissemination and Adoption

The FHCC team has adopted our platform in their Clinical Laboratory Improvement Amendments (CLIA) certified laboratory and applied this pipeline to analyze spatial proteomics data generated using the PhenoCycler-Fusion system. We are actively expanding the utilities of our pipeline to meet the needs of translational research, and testing our pipeline on data generated by other spatial proteomics technologies, such as Lunaphore COMET. The wet lab work was published in Cancer Biomarkers 2025 [10].

Our graphical end-to-end pipeline was published in GigaByte 2026 [11]. A preprint is available from BioRxiv 2025 [12].

Our end-to-end containerized pipeline is publicly available at <https://github.com/BioDepot/SpatialProteomics>. A pre-configured instance of our pipeline is available at GitHub Codespaces https://github.com/codespaces/new?hide_repo_select=true&ref=main&repo=979017184&skip_quickstart=true. A demonstration video of our pipeline is available at <https://youtu.be/vHjdCAZQhfE>.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

It took some time for the computational team at UW to learn all the intricacies of analyzing spatial proteomics data. This is because multiple software tools (both academic and commercial) were initially used by the FHCC team in their analysis design. Having frequent communications between the UW and FHCC teams greatly facilitated collaboration. Bryce Fukuda served as a liaison between the two teams.

4.2 Organizational Levels

It is often difficult to get replication studies published in a journal, and to acquire funding to support these types of studies. We are very grateful for this opportunity.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

Here is a breakdown of how long it took us when referring to the components of the workflow shown in Figure 1:

1. QuPath is a published open-source tool. Creating a Docker container for QuPath took a couple of days.
2. Writing a script to "export data" took a couple of days.
3. The Jupyter notebook contains code to perform normalization and clustering. This took 1-2 weeks.
4. Writing the script to import images back into QuPath took a couple of days also.

So, the total time for implementation is approximately 3 weeks. But this assumes that the workflow has already been designed. Most of the work is in designing the workflow, not the technical implementation.

To run our workflows, the user needs to install Docker. If the user is technically proficient, it takes a couple of hours to get started.

5.2 Associated Costs

There are no associated costs with running our open-source computational pipeline. The Yeung lab has multiple sets of spatial proteomics data sets with breast cancer, colon cancer, lymphoma, Merkel cell carcinoma, bladder cancer, and lung cancer that have been previously analyzed with other spatial analysis platforms.

5.3 Required Resources

The only required resources associated with running our computational pipeline are computational and expert personnel FTE. Our pipelines can be deployed on desktops, local, or cloud servers.

6. COLLABORATION

6.1 Key Collaborations

This replication project is a collaboration between the Fred Hutchinson Cancer Center (FHCC) and University of Washington (UW) Tacoma.

The FHCC performed all the wet lab experiments and designed the analysis. The UW team implemented containerized computational pipelines and performed data analysis. Our teams had an established track record of collaboration, previously primarily on nanopore sequencing data (R21CA280520 and 3R21CA280520-01S1).

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Sustainability, Transferability, and Generalizability

We believe that our laboratory protocols and computational analytical pipeline can be adopted by other laboratories. All our software tools are open source, with training materials and pre-configured instances available. In particular, a pre-configured instance of our pipeline with test data is available at GitHub Codespaces https://github.com/codespaces/new?hide_repo_select=true&ref=main&repo=979017184&skip_quickstart=true. Codespaces include a quota of 120 hours and 180 hours of free compute time per month for GitHub personal and pro accounts respectively. Users from any laboratory can log in directly and test drive our pipeline without installing any software on their own computers. A demonstration video of our pipeline is also available at <https://youtu.be/vHjdCAZQhfE>. Our training portal provides specific instructions on how to run our workflow using user's own data (https://biodepot.github.io/training/executing_workflows/).

7.2 Current Sustainability Plans

We will actively maintain and update our GitHub repository and documentation for our software.

7.3 Scaling and Application

Our open-source computational pipeline is scalable in the sense that it can be deployed on the cloud or on-prem using internal servers. Internal servers allow for additional security for large hospitals and academic clinical entities to analyze their spatial data. To accommodate increased usage, cloud based deployment can be scaled by upgrading to larger virtual machine instances with more compute/memory or through scaling management services provided by commercial cloud. On internal services, workload managers can be employed to schedule jobs efficiently.

7.4 Future Integration Plans

Currently, our computational pipeline has been tested using PhenoCycler and Lunaphore data. We plan to extend support and testing of our computational pipeline to other spatial proteomics technologies. We are actively soliciting and addressing feedback from users at the FHCC. We will also continue to improve and expand our training materials to facilitate the adoption of our protocols by other labs.

8. REFERENCES

1. Method of the Year 2024: Spatial proteomics. (2024). *Nature Methods*, 21(12), 2195–2196. <https://doi.org/10.1038/s41592-024-02565-3>
2. Horvath, P., & Coscia, F. (2025). Spatial proteomics in translational and clinical research. *Molecular Systems Biology*, 21(6), 526–530. <https://doi.org/10.1038/s44320-025-00101-9>
3. Yeung, C. C., Jones, D. C., Woolston, D. W., Seaton, B., Donato, E. L., Lin, M., & authors. (2025). Spatial proteomics and transcriptomics characterization of tissue and multiple cancer types including decalcified marrow. *Cancer Biomarkers*, 42(1), 18758592241308757. (Note: If you have the full list of authors to replace "et al.", you should list them all up to 20).
4. Singh, P., Wright, JH., Smythe, K. S., Fukuda, B., Hung, L.-H., Yeung, C. C., & authors. (2026). Graphical and interactive spatial proteomics image analysis workflow. *Gigabyte*, 2026. <https://doi.org/10.46471/gigabyte.186>
5. Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (Eds.). (2018). Cell detection with star-convex polygons. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, 11, 21–29. Springer. (Note: Page numbers for this specific conference chapter were added for completeness).
6. Bankhead, P., Loughrey, M. B., Fernandez, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., & authors. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7(1), Article 16878. <https://doi.org/10.1038/s41598-017-17204-5>
7. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., & authors. (2012). Fiji: An open-source platform for biological-image analysis. *Nature Methods*, 9(7), 676–682. <https://doi.org/10.1038/nmeth.2019>
8. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319. <https://doi.org/10.1038/nbt.3820>
9. Molder, F., Jablonski, K. P., Letcher, B., Hall, B. B., Tomkins-Tinch, C. H., Sochat, V., & authors. (2021). Sustainable data analysis with Snakemake. *F1000Research*, 10, Article 33. <https://doi.org/10.12688/f1000research.29032.2>

10. T. Yeung et al. (2025), "Spatial proteomics and transcriptomics characterization of tissue and multiple cancer types including decalcified marrow," *Cancer Biomarkers*.
<https://doi.org/10.1177/18758592241308757>
11. A. Singh et al (2026)., "Graphical and Interactive Spatial Proteomics Image Analysis Workflow," *GigaByte*.
<https://doi.org/10.46471/gigabyte.186>
12. A. Singh et al. (2025), "Graphical and Interactive Spatial Proteomics Image Analysis Workflow," *bioRxiv*.
<https://doi.org/10.1101/2025.05.23.655879>

TEAM WINNER

Team PGRM (Phenotype-Genotype Reference Map)



Built and validated the Phenotype-Genotype Reference Map (PGRM), a tool that lets biobanks assess and improve the quality of their phenotyping data.

Team Members: Lisa Bastarache · Sarah Delozier · Anita Pandit · Jing He · Adam Lewis · Aubrey C. Annis · Jonathon LeFaive · Joshua C. Denny · Robert J. Carroll · Russ B. Altman · Jacob J. Hughey · Matthew Zawistowski · Josh F. Peterson

We are a team of researchers based in Vanderbilt University Medical Center (VUMC) and the University of Michigan (UMich). Our work centers on leveraging electronic health records (EHRs) and biobanks to address clinically meaningful questions and develop tools to improve healthcare outcomes. Drawing on our collective expertise in computer science, informatics, and biomedical research, we focus on devising methods to better represent phenotypes using EHR data to improve quality, transparency and trust in biobank research in the age of the so-called “replication crisis.” Through our work with biobanks and EHR datasets, we have developed a deep understanding of the challenges of using EHR data for research and an understanding of how untested assumptions can lead to consequential errors. EHR data is inherently complex and messy, characterized by ascertainment biases, missing information, and all manner of ambiguities. These challenges frequently result in discrepancies between the populations we aim to study (e.g., individuals with type 2 diabetes) and the operational definitions derived from EHR-based algorithms. Ensuring that EHR-based accurately truly represent the intended populations requires constant vigilance and skepticism. Our team’s collective decades of experience wrangling EHR data have led to a profound awareness that the phenomena we study may not always align well with their representation in the data.

Our longstanding interest in replication stems from this awareness of the gap between representation and reality in EHR-based research. In this work, we not only present thousands of individual replications of phenotype/genotype associations but also introduce a tool designed to evaluate the validity of phenotype definitions and methodological assumptions in biobank-based research. We believe that our approach of “replication as quality control (QC),” which we detail further below, is a powerful strategy to enhance the rigor and reliability of EHR-based research.

As a team that values replication as a cornerstone of scientific progress, we are thrilled to participate in this challenge sponsored by NIH. It is a great opportunity to further promote our work and to cultivate a culture that values the importance of replication in our field.

1. REPLICATION STRATEGY

1.1 Strategy Description

Our replication activity is the development of the phenotype-genotype reference map (PGRM) – a tool to conduct genotype/phenotype replication experiments in population biobanks (Bastarache et al., 2023). In recent years, the use of biobanks has grown enormously, boosted by the creation of large-scale and accessible datasets such as All of Us and UK Biobank. This has led to a profound shift in the way genetic studies are conducted. Early genetic studies, including the first genome-wide association studies (GWAS), were based on recruited cohorts that were carefully phenotyped. Today, the majority of genetic research is conducted on multi-purpose biobanks, and phenotyping algorithms have replaced physical exams and manual chart review.

Despite this seismic shift in the way phenotypes are ascertained for genetic studies, there has been a notable lack of inquiry into how this shift has impacted research data quality. This is true despite the well documented challenges in accurately capturing phenotypes with EHR data including, for example, semantic shift of coded data, the effect of clinical context on measurements, secular trends in diagnoses, and cross-institutional heterogeneity in medical documentation (Bastarache et al., 2022). These problems are well known among researchers; however detecting phenotyping problems in analytic data sets based on biobank data is challenging. Manual validation requires access to the full clinical notes which are not available in most accessible cohorts like All of Us and UK Biobank. Thus, while there are well established quality control (QC) procedures and metrics that

are routinely applied to genomic data (Laurie et al., 2011), there are no equivalent tools to ensure the quality of phenotype data.

We developed the PGRM to assess the quality of biobank data and facilitate replication experiments across and within biobanks. The PGRM is a curated set of 5,879 phenotype-genotype associations derived from 523 GWAS publications from the NHGRI-EBI GWAS Catalog (Sollis et al., 2023). It includes 149 disease phenotypes encoded as phecodes, which are widely used ICD-based phenotypes that are applicable across biobanks (Bastarache, 2021). The PGRM associations were selected through a rigorous curation process to ensure compatibility with replication of genetic associations within general-purpose biobanks, as specified by recommendations from the NCI-NHGRI working group (Chanock et al., 2007). Annotations include alignment of phenotype definitions, effect sizes to support power calculations, and risk allele alignment to facilitate variant matching. As a publicly available R package, the PGRM makes it easy for researchers to attempt replication of previously reported genotype-phenotype associations, lowering the barrier for replication-based studies and addressing an unmet need for scalable tools to assess phenotype quality.

For the implementation of the PGRM strategy, we defined several replication related measures. The simplest of these is the overall replication rate (RR_{All}), which is the total number of associations replicated in the study cohort divided by the total number of replications attempted. Additional measures like the powered replication rate (RR_{Power}) and the actual:expected ratio (A:E) take statistical power into account based on the reported effect size from the discovery cohort and cohort size of the study cohort. For a single association, replication is defined as obtaining statistically significant result ($p < 0.05$) with the same direction of effect in the replication cohort.,

The PGRM rests on the premise that the quality of biobank data can be inferred by its ability to support robust replication of known genotype/phenotype associations. In the publication introducing the PGRM, we demonstrate this "replication as QC" premise by showing that replication rate is sensitive to data corruption. In one experiment, we show that the introduction of random noise to a phenotype definition significantly decreased the replication rate. Through a series of experiments comparing different analytical parameters, we also demonstrate how replication measures can be used to empirically assess different case/control definitions, evaluate the effect of cohort composition, or gauge the effect of missing data.

1.2 Outcome Evaluation Methods

The PGRM was validated through replication studies across four independent biobanks—BioVU, Michigan Genomics Initiative (MGI), UK Biobank (UKB), and BioBank Japan (BBJ). We found that each of these biobanks demonstrated robust replication of known phenotype/genotype association, achieving replication rates at or near what was expected based on power. In this massive cross-biobank replication experiment that included thousands of associations for over 100 unique phenotypes, we were also able to study the replicability of associations, identifying factors that were predictive of replicability. Unsurprisingly, associations with larger effect sizes and lower p-values reported in the discovery cohort were more likely to replicate. We also found that associations related to mental/behavioral disorder phenotypes were less likely to replicate, consistent with prior research demonstrating the challenge of accurately capturing these phenotypes.

1.3 Innovative Features

To our knowledge, the PGRM is the first tool introduced to address global phenotype quality in biobanks. While the PGRM is not sufficient to fully address issues of data quality, it is a powerful tool that can be integrated into essentially any biobank study. We believe that the field would be improved if researchers adopted replication measures as a standard component of an analytical pipeline, just as genotype QC metrics are reported as a standard in GWAS papers. This publicly available tool could promote a culture in which the reporting of phenotype quality metrics is viewed as a standard.

2. ENVIRONMENT

2.1 Environment Description

The origins of the PGRM can be traced back to a 2013 paper that systematically compared phecode-based association results with the GWAS catalog (Denny et al., 2013). That study tested the hypothesis that

associations reported in the NHGRI-EBI GWAS catalog could be robustly replicated using phenome-wide association study (PheWAS) methodology. The study was conceptualized by Josh Denny (JD), Lisa Bastarache (LB), and Dan Roden (DR). LB was the programmer lead on the project and thus was responsible for curating and structuring GWAS catalog association to facilitate replication studies across biobanks.

Following the publication of this study, LB continued to use replication rates as a means of evaluating data quality for new datasets genotyped in BioVU. In two separate cases, low-replication rate indicated data quality problems which were resolved before the datasets were used by researchers. Seeing its value for QC, LB began disseminating the map to external collaborators so they could perform their own checks and teaching trainees to use the tool as an early check prior to conducting research studies.

2.2 Location, Timing, and Team Details

By 2022, a team at VUMC and University of Michigan decided to formalize this tool that had organically generated interest among researchers concerned with data quality. This collaboration took a useful but basic GWAS catalog-based map to a mature and feature rich tool which included validated measures and cross-cohort benchmarks. LB and Josh Peterson (JP) led the conceptual development of the tool. LB and Sarah Delozier (SD) undertook the curation work, which entailed aggregating results from the GWAS catalog and annotating the associations which, in most cases, required reading the original publications (and supplementary data files). Jake Hughey (JH) assisted with the definition of replication metrics. Matt Zawistowski (MZ), Anita Pandit (AP), Aubrey Annis (AA), and Jonathon LeFaive (JL) conducted the replication experiments in MGI and UKB and worked to ensure the cross-biobank compatibility of the PGRM. Robert Carroll (RC) and Russ Altman (RA) advised on the design of the PGRM to support biobanks beyond those included in the flagship paper.

2.3 Funding Support

This work was supported by the National Library of Medicine grant R01-LM010685 (first granted in 2012), the National Center for Advancing Translational Sciences grant UL1TR002243 (2017), the National Human Genome Research Institute grant R01-HG012657 (2022) and the University of Michigan Advanced Genomics Core.

3. IMPACT

3.1 Replication Activity Impact

Since its release in 2023, the PGRM has been used by numerous groups. Several publications report replication rates and related measures like AER as a global assessment of biobank data quality, including All of Us (The All of Us Research Program Investigators, 2024), the Taiwan Precision Medicine Initiative (Chen et al., 2025), and the Alliance for Genomic Discovery (Avery et al., 2025). Newbury et al. (2025) used the PGRM to compare the performance of different phenotyping algorithms in the UK Biobank. The PGRM has also been used to systematically evaluate how ancestry impacts replication, such as this study using the BioMe cohort (Cullina et al., 2024). Finally, Pimplaskar et al. (2025) used the PGRM to quantify the effect of inclusion bias in biobanks and show that weighting samples to account for bias improves replication rates .

The publicly available PGRM R package includes code to facilitate experiments and the replication results of over 10,000 genotype/phenotype associations across four biobanks. Over 4,000 additional replication results have been reported across five new biobanks, with this count reflecting only results published in the scientific literature. These studies provide detailed statistics on replication attempts, including those that yielded non-significant results, thereby promoting greater transparency in the field.

The PGRM and R code to compute replication measures are available for free on GitHub: <https://github.com/PheWAS/pgmr>.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

We overcame numerous barriers to complete the PGRM. First, curating the PGRM required substantial manual effort, more than we anticipated prior to undertaking the study. While we are enthusiastic supporters and contributors to the NHGRI-EBI GWAS catalog, we found that its annotations were insufficient for supporting rigorous replication experiments. Important details regarding the composition of the discovery and phenotype definitions were not provided in the catalog and had to be curated from the reference research publications and embedded in the PGRM mappings.

To validate the PGRM and assess the generalizability of the approach, we needed to test the tool across numerous biobanks. This challenge was met by leveraging longstanding collaborative relationships with biobank researchers. For this effort, we collaborated with researchers at the University of Michigan, who worked seamlessly with the VUMC team to provide replication metrics for two additional biobanks.

A final challenge relates to the establishment of replication measures as a standard QC metric to be reported in biobank research papers. We have made some inroads along these lines, most importantly by reporting PGRM measures in the flagship AoU paper. Changing research practices is challenging and will require continuing efforts on our part. Applying for the NIH replication prize is part of our promotional efforts.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

The implementation of the PGRM tool unfolded over a multi-year period, beginning with its conceptualization and culminating in its publication in 2023. Below is the detailed timeline:

2022: The project was formalized following growing interest in the tool among researchers concerned with data quality. Early development began to transition a basic GWAS catalog-based map to a mature, feature-rich tool.-Conceptual development was led by LB and Josh Peterson (JP), establishing the framework for the tool's design, functionality, and goals.-Curatorial efforts were initiated by LB and Sarah Delozier (SD), involving the aggregation of results from the GWAS catalog and thorough annotation of associations. This labor-intensive process required reading original publications and supplementary data files to ensure accuracy.

2023: Development of replication metrics was carried out with assistance from Jake Hughey (JH). Guidance on tool design to facilitate widespread applicability of the PGRM was provided by Robert Carroll (RC) and Russ Altman (RA) -Validation studies within the MGI and UKB biobanks were conducted by the team at University of Michigan, including Matt Zawistowski (MZ), Anita Pandit (AP), Aubrey Annis (AA), and Jonathon LeFaive (JL). Their work ensured cross-biobank compatibility and robustness of the PGRM. Validation studies within the Biobank Japan were conducted by Jing He (JE) and Adam Lewis (AL). LB developed the R package that operationalized the PGRM, making it accessible to the broader research community. The work was finalized and published, marking the completion of the project.

5.2 Associated Costs and Required Resources

The successful implementation of the PGRM tool required funding to cover person costs, including salaries for team members (LB, JP, SD, JH) who contributed to the conceptual, developmental, curatorial, experimental, and software engineering aspects of the project. Additional costs included server maintenance for the VUMC server used to store, process, and analyze BioVU data, and costs associated with preparing and publishing the findings in a peer-reviewed journal.

6. COLLABORATION

6.1 Key Collaborations

The development of the PGRM grew out of longstanding collaborative relationships among VUMC researchers. LB, JP, and JH have all worked together extensively on biobank and phenotype methods papers in the past. SD was a trainee in the department of bioinformatics.

To identify validation partners for the PGRM, we looked to our longtime collaborators at the University of Michigan. LB has worked extensively with this team, led by MZ. This includes a prior publication assessing the cross-biobank replicability of genetic associations with laboratory measures (Goldstein et al., 2020).

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Current Sustainability Plans

To ensure the long-term sustainability of the PGRM (Phenotype-Genotype Replication Map), we have outlined the following steps:

1. **Maintaining Access and Visibility:** The PGRM package and source map will remain accessible to the research community via GitHub and our website (phewascatalog.org). Regular updates and monitoring will ensure continued usability and relevance.
2. **Updating with Newer GWAS Studies:** We plan to update the PGRM by incorporating findings from newer genome-wide association studies (GWAS), including studies involving diverse genetic ancestries. The current PGRM is primarily based on associations derived from European ancestry cohorts, which limits its generalizability across other genetic populations. Expanding the PGRM to include cross-ancestry associations will strengthen its utility and foster replication studies assessing genotype-phenotype relationships across diverse populations. While this update requires significant investment of time and resources, we believe it is critical for advancing research equity and scientific rigor. To support this effort, we plan to pursue funding through NIH awards.
3. **Scaling and Application:** Efforts are underway to integrate phecodeX, an advanced version of phecodes designed to capture more granular phenotypic classifications than those defined in phecode V1.2, into the PGRM. This integration is supported by a grant from the National Eye Institute (NEI) awarded in 2024 (R01EY036473), which focuses on enhancing the quality of eye phenotypes in biobank datasets. By integrating phecodeX, we anticipate that PGRM will become even more valuable for researchers conducting replication and phenotype-genotype association studies in diverse contexts.

To ensure the long-term growth and impact of the PGRM, we are committed to the following initiatives:

1. **Outreach and Dissemination:** We will actively promote the PGRM through lectures, seminars, and collaborative efforts. For instance, LB delivers an annual phenotyping lecture at UCLA's Computational Genomics Summer Institute (CGSI), where researchers express enthusiasm for the PGRM as a time-saving resource for curated GWAS associations. Additionally, the use of PGRM in several published papers this year will help increase awareness and adoption within the scientific community.
2. **Educational Initiatives:** We are considering applying for an R25 grant to develop and teach a course on EHR phenotyping with a focus on the "replication as quality control (QC)" theme. This course will aim to educate researchers on best practices for leveraging the PGRM and related tools in their studies.
3. **Collaboration and Networking:** We will continue fostering collaborations with colleagues to expand the reach and applicability of the PGRM. Networking with researchers across diverse fields will ensure that the resource evolves to meet emerging scientific needs.

Through these activities, we aim to position the PGRM as a foundational tool in biobank research, ensuring its scalability, transferability, and sustainability for years to come.

8. REFERENCES

1. Bastarache, L., Delozier, S., Pandit, A., He, J., Lewis, A., Annis, A.C., LeFaive, J., Denny, J.C., Carroll, R.J., Altman, R.B., et al. (2023). The phenotype-genotype reference map: Improving biobank data science through replication. *Am J Hum Genet* 110, 1522–1533. <https://doi.org/10.1016/j.ajhg.2023.07.012>.
2. Bastarache, L., Brown, J.S., Cimino, J.J., Dorr, D.A., Embi, P.J., Payne, P.R.O., Wilcox, A.B., and Weiner, M.G. (2022). Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. *Learn Health Syst* 6, e10293. <https://doi.org/10.1002/lrh2.10293>.
Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., Painter, I., ... GENEVA Investigators (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic epidemiology*, 34(6), 591–602. <https://doi.org/10.1002/gepi.20516>
3. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., ... Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic acids research*, 51(D1), D977–D985. <https://doi.org/10.1093/nar/gkac1010>
4. Bastarache, L. (2021). Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* 4, 1–19. <https://doi.org/10.1146/annurev-biodatasci-122320-112352>.
5. Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., et al. (2007). Replicating genotype–phenotype associations. *Nature* 447, 655–660. <https://doi.org/10.1038/447655a>.
6. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* 31, 1102–1111. <https://doi.org/10.1038/nbt.2749>.
7. All of Us Research Program Genomics Investigators (2024). Genomic data in the All of Us Research Program. *Nature*, 627(8003), 340–346. <https://doi.org/10.1038/s41586-023-06957-x>
Chen, H.-H., Chen, C.-H., Hou, M.-C., Fu, Y.-C., Li, L.-H., Chou, C.-Y., Yeh, E.-C., Tsai, M.-F., Chen, C., Yang, H.-C., et al. (2025). Population-specific polygenic risk scores for people of Han Chinese ancestry. *Nature*, 1–10. <https://doi.org/10.1038/s41586-025-09350-y>.
8. Avery, C., Babanejad, M., Baker, J., Bledsoe, X., Blostein, F., Corty, R.W., Ellis, K., Hung, A.M., Lake, A., Shelley, J., et al. (2025). Genome sequencing of 35,024 predominantly African ancestry persons addresses gaps in genomics and healthcare. Preprint at medRxiv, <https://doi.org/10.1101/2025.10.30.25338549>.
9. Newbury, A., Elhussein, A., and Gürsoy, G. (2025). Multi-domain rule-based phenotyping algorithms enable improved GWAS signal. *npj Digit. Med.* 8, 499. <https://doi.org/10.1038/s41746-025-01815-8>.
10. Cullina, S., Shemirani, R., Asgari, S., and Kenny, E.E. (2024). Systematic comparison of phenome-wide admixture mapping and genome-wide association in a diverse biobank. Preprint at medRxiv, <https://doi.org/10.1101/2024.11.18.24317494>.
11. Pimplaskar A, Qiu J, Lapinska S, Tozzo V, Chiang JN, Pasaniuc B, Olde Loohuis LM. Inclusion bias affects common variant discovery and replication in a health-system linked biobank. *Am J Hum Genet.* 2026 Apr 2;113(4):702-714. doi: 10.1016/j.ajhg.2026.02.011. Epub 2026 Mar 10. PMID: 41812654; PMCID: PMC13087405.Pimplaskar, A., Qiu, J., Lapinska, S., Tozzo, V., Chiang, J.N., Pasaniuc, B., and Olde Loohuis, L.M. (2025). Inclusion bias affects common variant discovery and replication in a health-system linked biobank. *medRxiv*, 2025.04.04.25325131. <https://doi.org/10.1101/2025.04.04.25325131>.
12. Goldstein, J.A., Weinstock, J.S., Bastarache, L.A., Larach, D.B., Fritsche, L.G., Schmidt, E.M., Brummett, C.M., Kheterpal, S., Abecasis, G.R., Denny, J.C., et al. (2020). LabWAS: Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *PLOS Genetics* 16, e1009077. <https://doi.org/10.1371/journal.pgen.1009077>.

TEAM WINNER

The RESI Team



Four biostatisticians applying the robust effect size index to improve how effect sizes are reported and replicated across biomedical imaging and psychiatry research.

Team Members: Megan Jones · Kaidi Kang · Xinyu Zhang · Simon Vandekar

The RESI Team is four biostatisticians who have done statistical methodology research using the robust effect size index with the goal of improving effect size reporting and replicability for scientific research studies. Our scientific interests primarily lie in biomedical imaging and psychiatry. Megan Jones recently earned her PhD in Biostatistics at Vanderbilt University and is beginning a postdoc at Queen’s University. Kaidi Kang is an Assistant Professor of Biostatistics and Data Science at Wake Forest University School of Medicine. Xinyu Zhang is a fourth-year PhD student in Biostatistics at Vanderbilt University. Simon Vandekar is Associate Professor of Biostatistics at Vanderbilt Health with 15 years of experience in biomedical imaging statistics.

1. REPLICATION STRATEGY

1.1 The challenge: The replicability crisis in neuroimaging

Human neuroimaging studies (e.g., studies using magnetic resonance imaging) are a critical translational tool for understanding brain-behavior associations in neuropathological conditions in humans that may inform new treatment strategies. Neuroimaging is a translational field, in the early stages of establishing biomarkers and brain-behavior associations for neuropsychiatric and neurological conditions. However, the field is currently facing a well-documented replication crisis. Recent studies have revealed that standard neuroimaging findings from even large-scale neuroimaging studies are often unreliable and fail to be replicated, raising serious concerns about the utility of neuroimaging studies (Marek et al. 2022; Kang et al. 2024; Spisak, Bingel, and Wager 2023; Tervo-Clemmens et al. 2023; Botvinik-Nezer and Wager 2023; Owens et al. 2021). Low replicability in the field has been identified to be associated with several factors, including the use of inefficient or even inappropriate statistical methods, an over-emphasis on p-value hypothesis testing, and underpowered study designs (e.g., small sample sizes, and small effect sizes) (Marek et al. 2022; Kang et al. 2024). Without a fundamental shift in how we design studies, analyze neuroimaging data, and report scientific findings, we risk wasting \$1.8 billion invested annually by the NIH on neuroimaging science (National Institutes of Health n.d.).

To address the replicability crisis in neuroimaging studies, our statistical research group has proposed novel study design and analysis methods centered on standardized effect size reporting. In neuroimaging, replicability is defined as the probability of agreement in hypothesis tests between independent studies. Standardized effect sizes – unitless quantities summarizing the strength of a statistical finding – directly influence replicability (Kang et al. 2024; Botvinik-Nezer and Wager 2023). Reporting effect sizes improves replicability by enabling scientists to evaluate the strength of an association irrespective of its statistical significance (Schäfer and Schwarz 2019; Vandekar et al. 2020) and facilitates meta-analyses to identify consistent associations across many studies by compiling their observed effect sizes (Lakens 2013; Weissgerber et al. 2018; Schubert et al. 2025). Furthermore, reporting standardized effect sizes also improves our ability to perform accurate power analyses, which is essential for designing high-replicability studies (Botvinik-Nezer and Wager 2023; Lakens 2013; Kang et al. 2024). Our strategy comprises three key activities: demonstrating the critical role effect sizes play in increasing replicability, removing barriers to facilitate effect size reporting, and proposing novel, pragmatic study designs that improve replicability.

Replication Activity 1: Demonstrating improved replicability by effect size reporting

To illustrate the impact of effect sizes on improving replicability of neuroimaging studies, we performed reproducible analyses studying the association between age and brain functioning during rest in the Autism Brain Imaging Data Exchange (ABIDE) dataset (Vandekar and Stephens 2021). We demonstrated that analyzing statistical images of brain associations using effect sizes yields more consistent results across various studies with different sample sizes compared to p-value-based analysis (Figure 1). These analyses highlight the advantage of using effect sizes over p-values in neuroimaging research, and reinforce the growing consensus

across scientific disciplines regarding the benefits of effect size reporting for replicable research (Weissgerber et al. 2018; Lakens 2013; Schäfer and Schwarz 2019; Grimes 2024; Monsarrat and Vergnes 2017; Methratta 2025).

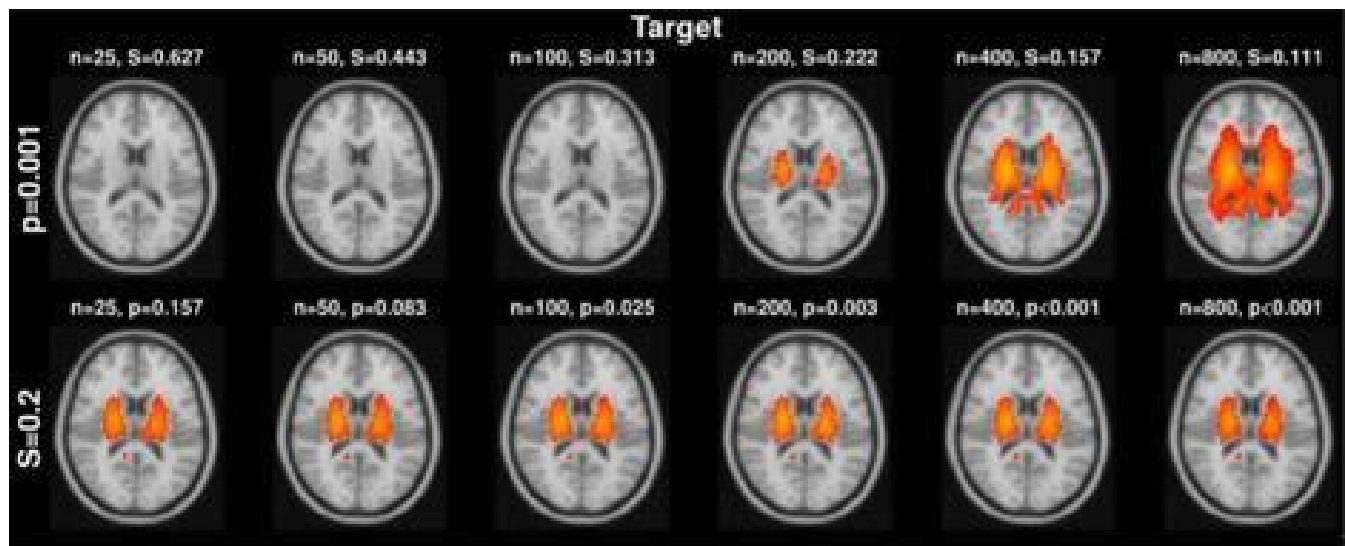


Figure 1. The warm-colored overlays on the brain images indicate regions where a study will, on average, detect age-related effects in resting state brain activity (data from ABIDE). The standard p-value thresholding approach for neuroimaging data (top row) has inconsistent results across sample sizes (n) because larger studies are better at detecting weaker signals, resulting in broader brain activation patterns than smaller studies. In contrast, thresholding based on the robust effect size index (denoted by S, bottom row) yields a consistent target; regardless of sample size, studies target similar effects, yielding more consistent and replicable findings on average. Adapted from (Vandekar and Stephens 2021).

Replication Activity 2: Removing barriers to reporting effect sizes

Scientific progress is often hindered by the difficulty of reporting effect sizes with the current effect size reporting framework: metrics are inconsistent across models and study designs, confidence interval construction methods are either missing or non-robust, and accessible software for easy effect size reporting are scarce (Kang et al. 2023; Fritz et al. 2012; Amaral and Line 2021). To address these limitations, we developed a robust effect size index (RESI) (Vandekar et al. 2020). Traditional effect sizes (e.g. Cohen's d/f or Pearson's r) are model-specific, difficult to compare across metrics, and often sensitive to assumption violations. RESI overcomes these challenges by providing a unified, theoretically grounded measure applicable across a wide range of statistical models and robust to model misspecification. The RESI has been widely applied by many research studies in various fields (including psychology, psychiatry, and nutrition) in addition to neuroimaging, significantly facilitating scientific communication. In our subsequent work (Kang et al. 2023), we demonstrated that standard confidence interval procedures frequently mis-estimate effect sizes. We proposed and validated bootstrap-based procedures that produce accurate, robust confidence intervals even under model misspecification. These tools enable researchers to report effect sizes consistently and reliably, improving transparency and comparability across studies. To facilitate adoption, we developed an open-source R package (<https://statimagcoll.github.io/RESI/>) that computes the RESI and its confidence intervals for multiple statistical models (M. Jones, Kang, and Vandekar 2025). Figure 2 illustrates the RESI software in a tutorial (M. Jones, Kang, and Vandekar 2025) applied to open-access health insurance data, highlighting how RESI identifies the strongest predictors beyond statistical significance. Finally, we also developed tools and software to address the challenge of reporting effect sizes in high-dimensional, spatially correlated neuroimaging data settings. Historically, approaches focused heavily on hypothesis testing and have been criticized for having low replicability (Yeung 2018; Eklund et al. 2016; Greve and Fischl 2018; Flandin and Friston 2016). We developed a RESI-based spatial confidence set approach, implemented in our open-source R package (Zhang et al. 2025), which allows investigators to easily visualize and report effect size with robust uncertainty quantification across a broad range of brain-behavior associations (Figure 3).

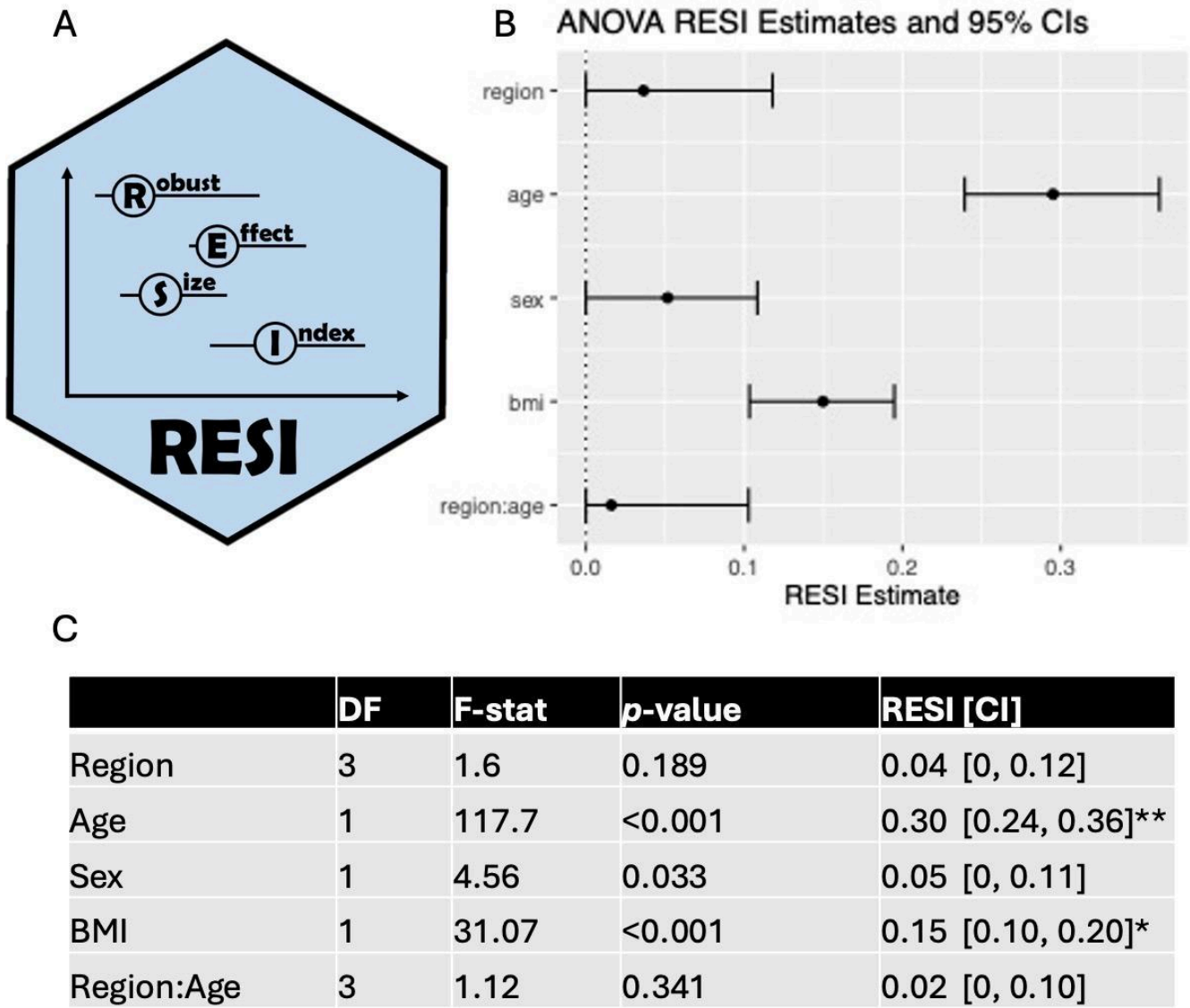


Figure 2. (A) Logo for the RESI software package. (B) An example of visualization of RESI estimates studying the association between demographic covariates and health insurance charges; (C) age, sex, and BMI all have significant effects using p-values, however, age and BMI have the strongest associations with insurance charges indicated by the larger RESI estimates. Prior to our work, there was no approach to robustly summarize effect strength across all types of statistical models. *small effect, **medium effect according to Cohen's recommendations (1988). Adapted from (M. Jones, Kang, and Vandekar 2025).

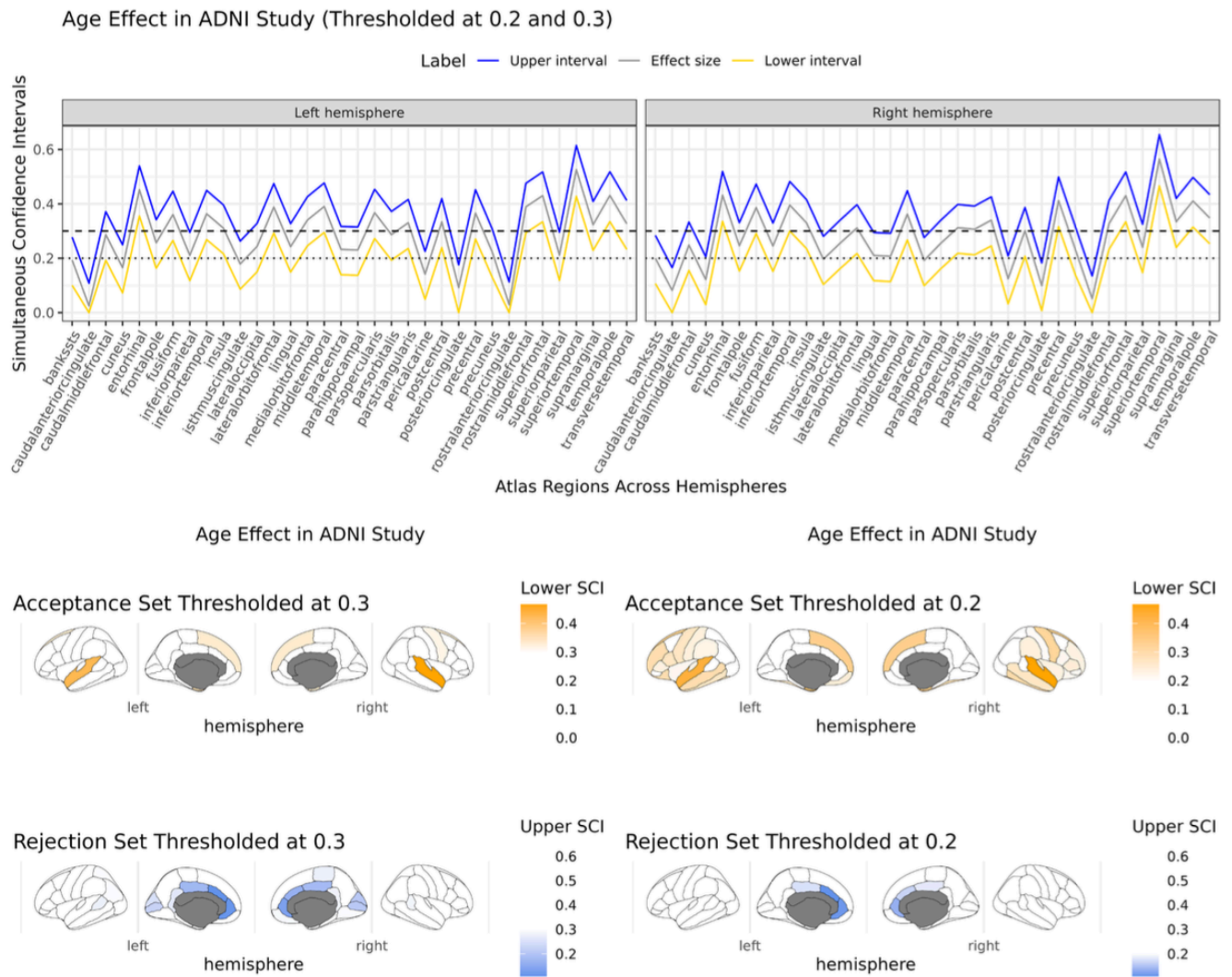


Figure 3. Simultaneous confidence intervals (top) are used to construct confidence sets (bottom) for RESIs for the age-cortical thickness associations across all brain atlas regions. The lower confidence limit (yellow) is used to construct an “acceptance set” using the threshold $S = 0.3$ (dashed line) and $S = 0.2$ (dotted line), which identifies regions of the brain where the true effect size is likely to be larger than that value. The upper confidence limit (blue) is used to construct the “rejection set” using the same threshold, identifying regions where the effect size is likely to be smaller than that value.

Replication Activity 3: Proposing practical study design recommendations to increase replicability

In our recent work published in *Nature*, we proposed and validated practical study design strategies that can be immediately deployed to boost replicability in both cross-sectional and longitudinal neuroimaging studies (Kang et al. 2024). These methods could be implemented in the collection of National Institutes of Health (NIH) funded studies, such as the Adolescent Brain Cognitive Development (ABCD®), to significantly reduce costs and improve replicability. In particular, we recommended accessible two-phase sampling methods to increase the predictor variability that led to 1.8 factor reduction in sample size needed for 80% replicability for most brain-behavior associations. Furthermore, in longitudinal designs, we demonstrated that the largest immediate gain in replicability is driven by the first follow-up records after baseline; collecting additional time points beyond the second measurement yields diminishing additional returns. Validated on one of the largest neuroimaging datasets ever assembled (Lifespan Brain Chart Consortium, over 75,000 participants from 62 primary studies), these cost-effective strategies offer a straightforward path to more reliable neuroimaging research.

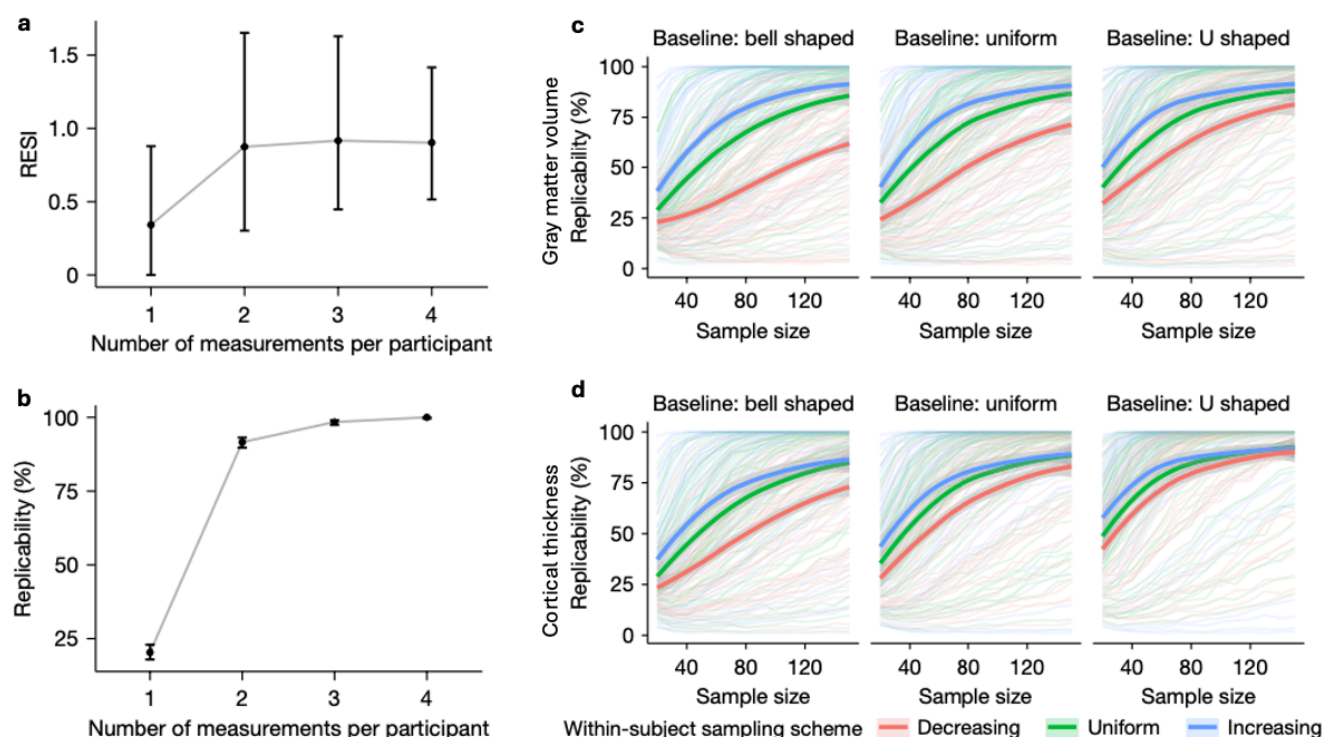


Figure 4. Adapted from (Kang et al. 2024). Evaluation of study design features to improve replicability. (a,b) Increasing the number of measurements from one to two per participant provides the most benefit on standardized (a) effect size and (b) replicability for the total GMV–age association when using uniform between-subject and within-subject sampling schemes and $n = 30$. The points represent the mean standardized effect sizes or estimated replicability, and the whiskers are the 95% confidence intervals. (c, d) Increased replicability for the associations of age with regional GMV (c) and regional cortical thickness (d) across all brain regions under different sampling schemes. The bold curves are the average estimated replicability across all regions with significant uncorrected effects using the full ADNI data. The shaded areas are the corresponding 95% confidence bands. Increasing the between-subject and within-subject variability of age by implementing different sampling schemes can increase the standardized effect sizes of age, and the associated replicability, on regional GMV and regional cortical thickness.

2. ENVIRONMENT

2.1 Environment Description

The research to develop methods to improve replicability was performed at Vanderbilt University Medical Center and Vanderbilt University from August 1st, 2020 to the present and was supported, in part, by National Institute of Mental Health grant R01MH123563. To ensure widespread adoption and immediate help on research replicability, we have implemented these methods in two open-source R packages: RESI (<https://statimagcoll.github.io/RESI/>) and pbj (<https://github.com/statimagcoll/pbj>) (M. Jones et al. 2025). These tools are designed to be computationally efficient and platform independent, running seamlessly on standard computing environments including Windows, macOS, and Linux.

3. IMPACT

3.1 Impact of Replication Activity 1

This activity demonstrated that standard p-value-based approaches are a primary driver of non-replicability in neuroimaging research. By establishing that effect size reporting yields consistent results across studies of different sample sizes (as shown in our ABIDE analysis), we provided a concrete, evidence-based roadmap for the field to shift away from binary hypothesis testing toward more robust quantitative practices.

3.2 Impact of Replication Activity 2

This activity democratized access to robust statistical reporting. By developing the RESI framework and releasing the open-source R packages, we have eliminated technical barriers that previously prevented researchers from reporting accurate effect sizes for complex, longitudinal, and spatial models. This equips the community with the necessary tools to immediately improve the transparency and comparability of their findings.

3.3 Impact of Replication Activity 3

This research delivers immediate, actionable strategies to optimize federal research investment, not only in the neuroimaging sector but across scientific disciplines facing high data acquisition costs. By identifying simple design changes that can reduce required sample sizes by, on average, nearly 50%, we offer a scalable solution to maximize the statistical power and replicability of large-scale NIH-funded initiatives like the ABCD study without increasing budgets. Crucially, these design principles are transferable to other fields, ensuring broader impact on scientific efficiency.

4. RESPONSE TO BARRIERS

4.1 Barrier Identification and Solutions

Our research has overcome technical and practical barriers to robust effect size reporting that have historically hindered the replication of neuroimaging studies.

Barrier 1: Incompatibility with complex neuroimaging analysis settings

The Challenge: Prior neuroimaging confidence-set methods worked only for univariate effect sizes (e.g., Cohen's d) and cross-sectional data. They could not support multi-group contrasts, nonlinear age effects, or repeated measures, exactly the types of analyses common in neuroimaging.

The Resolution: We built a new framework that combines RESI, a general robust effect size index that handles multi-parameter effects, and Generalized Estimating Equations (GEE) for repeated-measures neuroimaging. This allowed effect size inference for longitudinal, nonlinear, and multigroup models.

Barrier 2: Heterogeneity in effect size reporting

The Challenge: Conventional effect sizes are model specific and there is no unified software for reporting effect sizes with confidence intervals across different model types used in scientific research.

The Resolution: We defined a widely applicable effect size (RESI) and wrote open-source software to support its use across scientific disciplines.

Barrier 3: Inaccessible statistical methods limit practical implementation

The Challenge: Advanced statistical methods often fail to improve replicability because they remain inaccessible to applied researchers. A lack of user-friendly software and clear documentation created a barrier to adoption.

The Resolution: We bridged this gap by translating complex theory into accessible tools and actionable guidelines. In our Nature publication (Kang et al., 2024), we distilled sophisticated design principles into practical recommendations (e.g., two-time point longitudinal data and two-phase sampling strategies) that researchers can immediately deploy. To support analysis, we developed and maintained two open-source R packages, RESI and pbj, accompanied by comprehensive tutorials and reproducible code, ensuring that the neuroimaging community can immediately apply these rigorous methods to their own data.

5. TIMELINE & BUDGET

5.1 Implementation Timeline

The methods and software were developed over the span of 4-5 years, and the software is continually maintained and updated.

5.2 Associated Costs

Costs include computing resources, workspace, and salary/stipend/tuition support for the team members. Scientists who want to implement the described methods can use the open-source software immediately free-of-charge, but require funding to support the resources required for their research, and computational resources to use the methods. The computation resources are not beyond those accessible in most academic research institutions.

6. COLLABORATION

6.1 Key Collaborators

The research was driven by an extensive international collaboration among leading research institutions. Our primary research team at VUMC was driven by Megan Jones (MS), Kaidi Kang (PhD), and Xinyu Zhang (MS), and principal investigator, Simon Vandekar (PhD). Extensive evaluation of the proposed methods required a transdisciplinary team of scientists across national and international institutions including principal investigators Aaron Alexander-Bloch (MD; Children's Hospital of Philadelphia), Damien Fair (MD; University of Minnesota), Stephan Heckers (MD; VUMC), Theodore Satterthwaite (MD; University of Pennsylvania), and Neil Woodward (PhD; VUMC). Several of these collaborations were made to support the replication research activities, particularly with data access, processing, and interpretation.

7. SUSTAINABILITY / TRANSFERABILITY

7.1 Current Sustainability Plans

The effect size methods we use can be easily transferred to other scientific fields. The neuroimaging-specific methods we developed can be applied to other spatial datasets. Sustaining and advancing the methods and software requires continuing financial support for our research team, but can be implemented freely by other scientific research groups. Software updates are implemented by our research team approximately biannually. Dissemination and communication strategies include utilizing the effect size reporting software in scientific research and presenting the results and interpretation at conferences.

7.2 Future Outlook

Current research is developing methods to improve computation time for confidence interval methods and developing new methods to improve replicability of machine learning and imaging artificial intelligence models (M. T. Jones et al. 2025).

8. REFERENCES

1. Amaral, Emilyane de Oliveira Santana, and Sergio Roberto Peres Line. 2021. "Current Use of Effect Size or Confidence Interval Analyses in Clinical and Biomedical Research." *Scientometrics* 126 (11): 9133–45. <https://doi.org/10.1007/s11192-021-04150-3>.

2. Botvinik-Nezer, Rotem, and Tor D. Wager. 2023. "Reproducibility in Neuroimaging Analysis: Challenges and Solutions." *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, Reliability of Neurocognitive Measures for Mental Health, vol. 8 (8): 780–88. <https://doi.org/10.1016/j.bpsc.2022.12.006>.
3. Eklund, Anders, Thomas E. Nichols, and Hans Knutsson. 2016. "Cluster Failure: Why fMRI Inferences for Spatial Extent Have Inflated False-Positive Rates." *Proceedings of the National Academy of Sciences*.
4. Flandin, Guillaume, and Karl J. Friston. 2016. "Analysis of Family-Wise Error Rates in Statistical Parametric Mapping Using Random Field Theory." *Human Brain Mapping*.
5. Fritz, Catherine O., Peter E. Morris, and Jennifer J. Richler. 2012. "Effect Size Estimates: Current Use, Calculations, and Interpretation." *Journal of Experimental Psychology: General* 141 (1): 1. <https://doi.org/10.1037/a0024338>.
- Greve, Douglas N., and Bruce Fischl. 2018. "False Positive Rates in Surface-Based Anatomical Analysis." *NeuroImage* 171: 6–14.
6. Grimes, David Robert. 2024. "Towards Replicability and Sustainability in Cancer Research." *BJC Reports* 2 (1): 65. <https://doi.org/10.1038/s44276-024-00090-6>.
- Jones, Megan, Kaidi Kang, and Simon Vandekar. 2025. "RESI: An R Package for Robust Effect Sizes." *Journal of Statistical Software* 112 (March): 1–27. <https://doi.org/10.18637/jss.v112.i03>.
7. Jones, Megan, Kaidi Kang, Simon Vandekar, Gina Yu, and Xinyu Zhang. 2025. *RESI: Robust Effect Size Index (RESI) Estimation*. V. 1.3.2. Released July 29. <https://cran.r-project.org/web/packages/RESI/index.html>.
8. Jones, Megan T., Ishaan Gadiyar, Xinyu Zhang, et al. 2025. "Improved Estimation of Correlation Accuracy for Machine Learning Brain-Phenotype Associations." Preprint, bioRxiv, December 1. <https://doi.org/10.1101/2025.11.26.690778>.
9. Kang, Kaidi, Megan T. Jones, Kristan Armstrong, et al. 2023. "Accurate Confidence and Bayesian Interval Estimation for Non-Centrality Parameters and Effect Size Indices." *Psychometrika*, ahead of print. <https://doi.org/10.1007/s11336-022-09899-x>.
10. Kang, Kaidi, Jakob Seidlitz, Richard A. I. Bethlehem, et al. 2024. "Study Design Features Increase Replicability in Brain-Wide Association Studies." *Nature*, November 27, 1–9. <https://doi.org/10.1038/s41586-024-08260-9>.
11. Lakens, Daniël. 2013. "Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs." *Frontiers in Psychology* 4 (November): 863. <https://doi.org/10.3389/fpsyg.2013.00863>.
12. Marek, Scott, Brenden Tervo-Clemmens, Finnegan J. Calabro, et al. 2022. "Reproducible Brain-Wide Association Studies Require Thousands of Individuals." *Nature* 603 (7902): 654–60. <https://doi.org/10.1038/s41586-022-04492-9>.
13. Methratta, Elizabeth T. 2025. "Effect Size as a Measure of Biological Relevance for Offshore Wind Impact Studies." *ICES Journal of Marine Science* 82 (3): fsaf022. <https://doi.org/10.1093/icesjms/fsaf022>.
14. Monsarrat, Paul, and Jean-Noel Vergnes. 2017. "The Intriguing Evolution of Effect Sizes in Biomedical Research over Time: Smaller but More Often Statistically Significant." *GigaScience* 7 (1): gix121. <https://doi.org/10.1093/gigascience/gix121>.
15. National Institutes of Health (RePORTER: Search Results for Funded Projects.). n.d. "RePORTER: Search Results for Funded Projects." https://reporter.nih.gov/search/_dNnH1VaiEku_vZLZ7L2xw/projects/charts.
16. Owens, Max M., Alexandra Potter, Courtland S. Hyatt, et al. 2021. "Recalibrating Expectations about Effect Size: A Multi-Method Survey of Effect Sizes in the ABCD Study." *PLOS ONE* 16 (9): e0257535.
17. Schäfer, Thomas, and Marcus A. Schwarz. 2019. "The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases." *Frontiers in Psychology* 10 (April). <https://doi.org/10.3389/fpsyg.2019.00813>.
18. Schubert, Anna-Lena, Meike Steinhilber, Heemin Kang, and Daniel S. Quintana. 2025. "Improving Statistical Reporting in Psychology." *Communications Psychology* 3 (1): 156. <https://doi.org/10.1038/s44271-025-00356-w>.
19. Spisak, Tamas, Ulrike Bingel, and Tor D. Wager. 2023. "Multivariate BWAS Can Be Replicable with Moderate Sample Sizes." *Nature* 615 (7951): E4–7. <https://doi.org/10.1038/s41586-023-05745-x>.
20. Tervo-Clemmens, Brenden, Scott Marek, Roselyne J. Chauvin, et al. 2023. "Reply to: Multivariate BWAS Can Be Replicable with Moderate Sample Sizes." *Nature* 615 (7951): 7951. <https://doi.org/10.1038/s41586-023-05746-w>.

21. Vandekar, Simon N., and Jeremy Stephens. 2021. "Improving the Replicability of Neuroimaging Findings by Thresholding Effect Sizes Instead of P-Values." *Human Brain Mapping* 42 (8): 2393–98. <https://doi.org/10.1002/hbm.25374>.
22. Vandekar, Simon, Ran Tao, and Jeffrey Blume. 2020. "A Robust Effect Size Index." *Psychometrika* 85 (1): 232–46. <https://doi.org/10.1007/s11336-020-09698-2>.
23. Weissgerber, Tracey L., Oscar Garcia-Valencia, Vesna D. Garovic, Natasa M. Milic, and Stacey J. Winham. 2018. "Why We Need to Report More than 'Data Were Analyzed by t-Tests or ANOVA.'" *eLife* 7 (December): e36163. <https://doi.org/10.7554/eLife.36163>.
24. Yeung, Andy W. K. 2018. "An Updated Survey on Statistical Thresholding and Sample Size of fMRI Studies." *Frontiers in Human Neuroscience* 12. <https://doi.org/10.3389/fnhum.2018.00016>.

Appendix A: “How did you define replication?”

Submission	Winner Type	Response to "How did you define replication?"
ATCC Federal Solutions	Team	<i>We define replication as the repetition of a study or analysis to confirm that its findings are reliable, valid, and reproducible. Thorough documentation and methodological rigor help ensure that any reproduced findings accurately reflect the reliability of the original results.</i>
Antibody Team	Team	<i>Replication is made possible by a description that allows an experiment to be repeated identically, including a precise definition of all reagents used. It is here that the reproducibility of antibody experiments is often lacking, since antibodies derived from animal serum, ascites fluid, and hybridoma-derived monoclonal antibodies are typically not precisely defined and cannot be reproduced indefinitely.</i>
C20D3 – Vitamin A Replicate	Team	<i>Replication was defined as independent, multimodal verification of the central hypothesis “slowing vitamin A dimerization prevents retinal pigment epithelium damage in mice and halts progression of retinal atrophy in humans.” This encompassed: (1) independent validation of molecular mechanism and efficacy in pigmented and albino animal models using quantitative biomarkers; and (2) confirmation of clinical benefit across patient populations with Stargardt disease and age related macular degeneration</i>
Christina Kackos Brunelle	Individual	<i>For the purpose of these studies, the influenza research community defined replication as consistency in experimental results across participating laboratories. The specific metrics and evaluations used to define and assess consistency for each study are described below.</i>
ConductScience	Individual	<i>We defined replication as the ability of independent laboratories to re-execute the same method — with sufficient specification of the apparatus, procedure, and analysis — so that results can be directly compared across sites and yield consistent measurements when addressing the same experimental question. In practice, this means preserving complete, citable method components (hardware specifications, procedural logic, and analytic definitions) rather than narrative descriptions alone, so that differences in results reflect biology rather than uncontrolled variation in how a study was built and run.</i>
Jagadeesh Ramasamy	Individual	<i>Replication was defined as the independent, remeasurement of mitochondrial retention in red blood cells across multiple cohorts and study sites using harmonized assays, standardized procedures, and orthogonal validation methods. Additionally, we replicated not only the cellular phenotype but also the underlying mechanisms and clinical associations linking mitochondrial retention to oxidative stress, hemolysis, and disease severity.</i>

John Worthley	Individual	<i>Replication in our work is defined as demonstrating that ACE inhibitors, originally developed at sea level, maintain their efficacy and predicted pharmacological effects across the varying environmental conditions found at high altitudes, such as the hypoxia-induced stabilization of Hypoxia-Inducible Factors (HIFs). We assess whether the core drug action remains robust despite physiological and environmental variability by testing in high-elevation natural laboratories like the Donner Lake Basin</i>
Krissandra Johnson-Nealey	Individual	<i>I defined replication as the independent verification of clinical trial outcomes, protocol adherence, and safety reporting through parallel measurement, documentation, and audit-based validation across time, staff, and overlapping patient populations. In practice, replication meant embedding dual verification, cross-protocol outcome comparisons, and continuous quality checks into routine clinical research operations to ensure study findings were consistent, reproducible, and reflective of real-world patient outcomes.</i>
LATCH (LLM-Assisted Testing of Clinical Hypotheses)	Team	<i>In this work, replication is defined as direct computational replication, in which previously published clinical studies are independently re-executed using the methods as documented in the publication.</i>
Magdalena Kasendra	Individual	<i>In this work, we defined replication as a translational de-risking strategy: not only reproducing the core liver organoid biology and compound-specific DILI response phenotypes reported in the original Shinozawa study, but transforming them into a standardized, automation-ready platform with the reproducibility, scalability, and transferability needed for real-world drug safety applications.</i>
Michele Di Mascio	Individual	<i>Replication was defined using a hybrid framework combining direct and conceptual replication. Critical experimental conditions were preserved wherever necessary, while scientifically justified modifications optimized sensitivity and provided an independent test of the original biological hypothesis.</i>
PredictiveUNC Team	Team	<i>We define replication as the ability of independent researchers, using the same raw data sources and applying the same transparent curation and modeling workflow as in the published work, to reproduce both model feasibility and predictive performance within expected uncertainty ranges. We find it critical for any published model to include explicitly referenced sources of data and data curation protocols</i>
SOAR (Spatial Omics Analysis Replication)	Team	<i>We define “replicability” as achieving consistent results across repeated laboratory procedures and applying alternative computational tools to analyze the data.</i>
Team PGRM (Phenotype-Genotype Reference Map)	Team	<i>In this work, we adopted a definition of replication offered by an NCI-NHGRI working group formed in the early days of the GWAS era to establish standards for verifying genetic associations (NCI-NHGRI Working Group on Replication in Association Studies, 2007) [1]. In their 2007 paper, the working group offers several “suggested criteria for establishing positive replication,” including that replication studies should be conducted on independent datasets using a “similar phenotype,” “similar population,” and with “sufficient sample size.” Thus they define a positive replication as a study that meets these criteria and produces a result that is statistically significant and has the same direction of effect as the initial study.</i>

The RESI Team	Team	<i>Replicability was defined as the probability that two independent studies with the same design have significant results in the same direction.</i>

References:

[1] NCI-NHGRI Working Group on Replication in Association Studies; Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype-phenotype associations. *Nature*. 2007;447(7145):655-60. doi: 10.1038/447655a. PubMed PMID: 17554299. <https://pubmed.ncbi.nlm.nih.gov/17554299/>

Corresponding Authors

The contact information below is provided for readers who wish to connect directly with the contributors featured in this collection. Authors have consented to inclusion of their contact details in this publication. Where no contact is listed, the contributor opted not to include their information in this publication.

WINNER	CORRESPONDING AUTHOR CONTACT
ATCC Federal Solutions	
Antibody Team	Katherine Groff Email: katherineg@peta.org 
C20D3-Vitamin A Replicate	 
Christina Kackos Brunelle	Christina Kackos Brunelle Email: cbrunelle@deloitte.com 
ConductScience	Shuhan He, MD Email: shuhan@conductscience.com  
Jagadeesh Ramasamy	Jagadeesh Ramasamy, BSMS, MSc, PhD Email: jagadees@uic.edu  
John Worthley	 
Krissandra Johnson-Nealey	Krissandra Johnson-Nealey Email: krissandrau@gmail.com  
LATCH	Nayoon Gim Email: nygim@uw.edu ; Aaron. Y. Lee Email: leeay@wustl.edu 
Magdalena Kasendra	Magdalena Kasendra, Ph.D. Email: magdalena.kasendra@cchmc.org  
Michele Di Mascio	Michele Di Mascio Email: michele@micheledimascio.com  
Predictive/UNC Team	Alexander Tropsha Email: alex_tropsha@unc.edu  

SOAR (Spatial Omics Analysis Replication)	Ka Yee Yeung-Rhee Email: kayee@uw.edu  
Team PGRM (Phenotype-Genotype Reference Map)	Lisa Bastarache Email: lisa.bastarache@vumc.org  
The RESI Team	Simon Vandekar Email: simon.vandekar@vumc.org  

NIH Replication Prize