Check for updates

METHOD ARTICLE

Efforts to enhance reproducibility in a human performance research project [version 1; peer review: 1 approved with reservations]

Jeffrey A. Drocco[1], Kyle Halliday[2], Benjamin J. Stewart[1], Sarah H. Sandholtz[1], Michael D. Morrison[1], James B. Thissen[1], Nicholas A. Be[1], Christopher E. Zwilling[3], Ramsey R. Wilcox[3], Steven A. Culpepper[4], Aron K. Barbey[3], Crystal J. Jaing[1]

[1]Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, California, 94550, USA
[2]Computing Directorate, Lawrence Livermore National Laboratory, Livermore, California, 94550, USA
[3]Beckman Institute for Advanced Science and Technology and Department of Psychology, University of Illinois Urbana-Champaign, Urbana, Illinois, 61801, USA
[4]Department of Statistics, University of Illinois Urbana-Champaign, Champaign, Illinois, 61820, USA

v1 **First published:** 01 Nov 2023, **12**:1430
https://doi.org/10.12688/f1000research.140735.1
Latest published: 01 Nov 2023, **12**:1430
https://doi.org/10.12688/f1000research.140735.1

Abstract

Background: Ensuring the validity of results from funded programs is a critical concern for agencies that sponsor biological research. In recent years, the open science movement has sought to promote reproducibility by encouraging sharing not only of finished manuscripts but also of data and code supporting their findings. While these innovations have lent support to third-party efforts to replicate calculations underlying key results in the scientific literature, fields of inquiry where privacy considerations or other sensitivities preclude the broad distribution of raw data or analysis may require a more targeted approach to promote the quality of research output.

Methods: We describe efforts oriented toward this goal that were implemented in one human performance research program, Measuring Biological Aptitude, organized by the Defense Advanced Research Project Agency's Biological Technologies Office. Our team implemented a four-pronged independent verification and validation (IV&V) strategy including 1) a centralized data storage and exchange platform, 2) quality assurance and quality control (QA/QC) of data collection, 3) test and evaluation of performer models, and 4) an archival software and data repository.

Results: Our IV&V plan was carried out with assistance from both the funding agency and participating teams of researchers. QA/QC of data

Open Peer Review

Approval Status ?

	1
version 1 01 Nov 2023	? view

1. **Michael Fitzsimons**, The University of Chicago, Chicago, USA

Any reports and responses or comments on the article can be found at the end of the article.

acquisition aided in process improvement and the flagging of experimental errors. Holdout validation set tests provided an independent gauge of model performance.

Conclusions: In circumstances that do not support a fully open approach to scientific criticism, standing up independent teams to cross-check and validate the results generated by primary investigators can be an important tool to promote reproducibility of results.

Keywords
reproducibility of results, validation studies, evaluation methodology, data quality

This article is included in the Reproducible Research Data and Software collection.

Corresponding author: Jeffrey A. Drocco (drocco1@llnl.gov)

Author roles: Drocco JA: Conceptualization, Formal Analysis, Methodology, Project Administration, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Halliday K**: Conceptualization, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation; **Stewart BJ**: Data Curation, Formal Analysis, Methodology, Software, Writing – Review & Editing; **Sandholtz SH**: Data Curation, Formal Analysis, Software, Validation, Writing – Review & Editing; **Morrison MD**: Data Curation, Formal Analysis, Methodology, Validation; **Thissen JB**: Data Curation, Formal Analysis, Methodology, Validation; **Be NA**: Formal Analysis, Methodology, Writing – Review & Editing; **Zwilling CE**: Formal Analysis, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation; **Wilcox RR**: Data Curation, Formal Analysis, Software; **Culpepper SA**: Conceptualization, Formal Analysis, Methodology; **Barbey AK**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision; **Jaing CJ**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. We acknowledge the support of the Defense Advanced Research Projects Agency under contract HR001119S0021-MBA-FP-004.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Drocco JA, Halliday K, Stewart BJ *et al.* **Efforts to enhance reproducibility in a human performance research project [version 1; peer review: 1 approved with reservations]** F1000Research 2023, **12**:1430 https://doi.org/10.12688/f1000research.140735.1

First published: 01 Nov 2023, **12**:1430 https://doi.org/10.12688/f1000research.140735.1

Introduction

Reproducibility of findings is a fundamental requirement of any scientific research endeavor. Nevertheless, for a variety of reasons, reproducibility remains a challenge in many areas of the life sciences.[1] Batch effects, hidden variables, and low signal-to-noise ratios may interfere with researchers' ability to draw broad conclusions based on small quantities of data.[2] Similarly, data snooping may, even unintentionally, lead to the selection of models that have poor generalizability outside an original dataset.[3] These difficulties can be exacerbated in exploratory studies that are by design not limited to the consideration of only one or a small number of pre-specified hypotheses, but rather constructed for the purpose of testing a very large number of possible explanatory variables using a statistical approach.

Transparency in data collection and analysis has been suggested as a potential means of bringing to light methodological or other flaws that may impair the reproducibility of results in various areas of biomedical research. For example, authors who distribute notebooks integrating data, code, and text directly enable others to replicate some or all of the analysis supporting their stated conclusions.[4] While such measures do not exclude all possible errors that might call into question the reliability of published findings, scientists adhering to these practices considerably reduce the ambiguity associated with the steps in their workflow subsequent to data acquisition.[5,6]

Organizations that fund research must take into account these considerations and others in planning new research and development (R&D) programs. The time and money available to obtain answers to the scientific questions of stakeholder interest are generally limited. Reachback tasking that would allow re-analysis of data or models following an original period of performance is not always possible, as studies frequently rely on teams assembled in an *ad hoc* manner to respond to the requirements of a specific project call. Moreover, publication of results is not a guarantee that the artifacts of a research program will be fully preserved. While many journals have adopted standards for sharing of data and code, compliance with these policies is imperfect.[7,8] Indeed, selective publication practices have themselves been implicated as potential sources of bias in the scientific literature.[9,10] Finally, the aspirations of open science may conflict with project constraints when supporting data are not suitable for release into the public domain.

Here we describe the efforts of one research program, Measuring Biological Aptitude (MBA), a four-year effort sponsored by the Biological Technologies Office of the Defense Advanced Research Projects Agency (DARPA), to improve the reproducibility of studies performed with the goal of optimizing human performance in a variety of athletic and cognitive military skills tests. As the MBA program involved data encumbered by restrictions related to personal privacy, medical confidentiality, and national defense, a fully open approach to promoting reproducibility was not practical. Instead, the program sponsored the authors of this manuscript to conduct a comprehensive independent verification and validation (IV&V) program to test and evaluate the results generated by the primary program contractors and modeling teams.

Independent verification and validation

DARPA defines IV&V as "the verification and validation of a system or software product by an organization that is technically, managerially, and financially independent from the organization responsible for developing the product" (DARPA Instruction 70). In recent years, IV&V has become a key component of various DARPA research programs both inside and outside the life sciences domain.[11] In contrast to the standard in open science, which generally relies on the free and voluntary participation of members of the scientific community to verify the results of third party studies, DARPA's policy suggests that independent efforts to support the integrity of scientific results are of sufficient importance to merit direct funding, using teams selected for their expertise in the relevant technical areas.

According to the MBA Broad Agency Announcement (BAA), primary performers were charged to "identify, understand, and measure the expression circuits (e.g., genetic, epigenetic, metabolomic, etc.) that shape a warfighter's cognitive, behavioral, and physical traits, or phenotypes, related to performance across a set of career specializations." The IV&V team, by contrast, was directed to "verify and validate whether the expression circuits, as measured by the molecular targets identified, directly correlate to dynamic changes in performance traits in the individual and independently confirm…that those circuits correlate to selection success or failure." From this followed a corresponding but distinct schedule of tasks for each group (Figure 1).

In 2019, DARPA selected Lawrence Livermore National Laboratory (LLNL) and the University of Illinois Urbana-Champaign (UIUC) to lead the IV&V component of the MBA program. According to the IV&V plan developed by LLNL, the effort comprised four core focus areas: 1) a centralized secure data storage and exchange platform, 2) quality assurance and quality control checklists applied to data acquisition, 3) test and evaluation of performer modeling products, and 4) an archival software repository and data store.

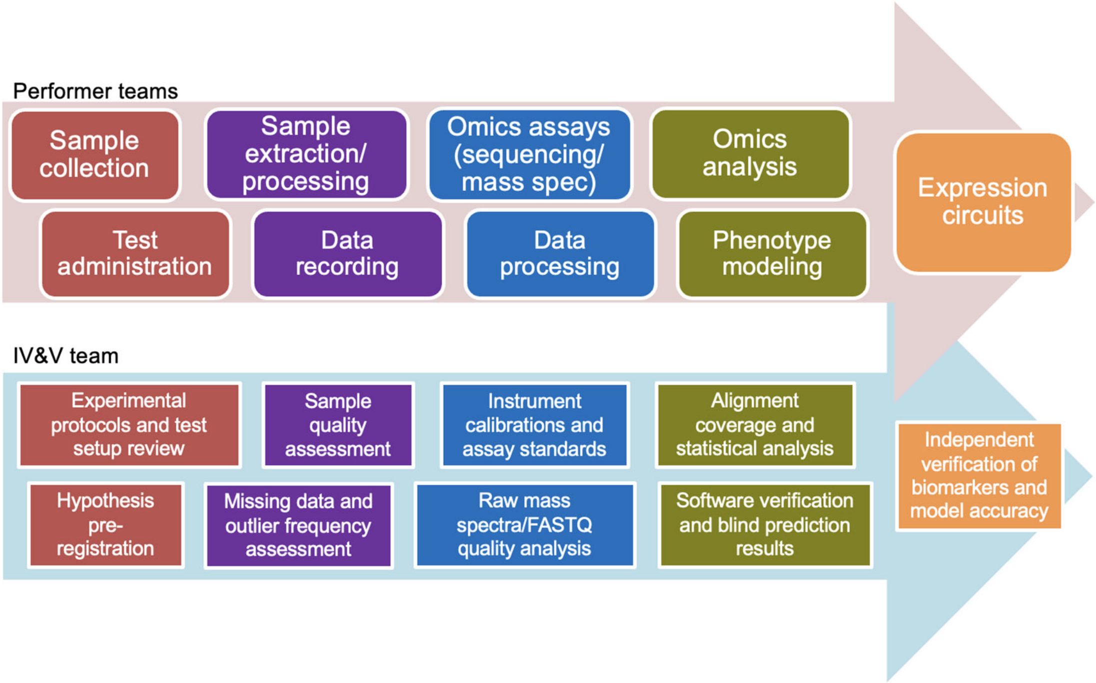
Performer teams
Sample collection
Sample extraction/ processing
Omics assays (sequencing/ mass spec)
Omics analysis
Test administration
Data recording
Data processing
Phenotype modeling
Expression circuits
IV&V team
Experimental protocols and test setup review
Sample quality assessment
Instrument calibrations and assay standards
Alignment coverage and statistical analysis
Hypothesis pre-registration
Missing data and outlier frequency assessment
Raw mass spectra/FASTQ quality analysis
Software verification and blind prediction results
Independent verification of biomarkers and model accuracy

Figure 1. Diagram of the MBA program workflow, denoting the activities of the primary performer teams (top) and corresponding responsibilities of the IV&V team (bottom).

Methods

Secure data storage and exchange platform

While the unit costs of bioinformatic data collection have declined in recent years, the acquisition and processing of large-scale omics data remain both financially and computationally expensive.[12] For reasons of reliability and cost savings, research sponsors may desire a centralized user facility for data storage and pre-processing, even in projects that involve multiple competing investigators and modeling teams. Moreover, if program managers wish to obtain a comparison of performance across several predictive models, centralized data services may help to maximize the time that data scientists and statisticians are able to devote to model selection while minimizing the risk that ambiguities in outcome labels or other metadata may lead different groups to substantially varying interpretations of the same modeling problem.[13,14]

A separate consideration in research involving human subjects involves data security and privacy. In the U.S., federal regulations require institutional review boards (IRBs) to evaluate each proposed project's provisions for protecting the privacy and confidentiality of human subjects information, regardless of whether a study explicitly plans to include data that is covered by other medical privacy laws.[15] In addition, considerations such as the possible re-identification of putatively de-identified health data may warrant additional data protection precautions even when not required by statute or regulation.[16]

To ensure both data security and data consistency for all teams working on the project, LLNL built a computing enclave for storing and analyzing MBA program data (Figure 2). Following best practices employed by other centralized biomedical data repositories, the enclave implemented cyber security controls at the FISMA Moderate policy level with enhanced controls from the NIST 800-53 and 800-66 information security guidelines for privacy and HIPAA compliance.[17,18,19] All accredited users of the enclave were required to complete cyber security training and a human subjects protection course prior to receiving computing accounts.[20] Access to the enclave was established through multifactor authentication.

To minimize risks of intentional and/or accidental duplication of human subjects data, the enclave featured a Virtual Network Computing (VNC) portal through which external collaborators could interact with program data. As the enclave excluded other networking protocols for ordinary users, the visual interface allowed modelers to perform analyses in a standard Linux computing environment while imposing a soft barrier against the bulk download of sensitive data. Modelers were allowed to upload new data or software dependencies to the enclave via the data transfer node. However,

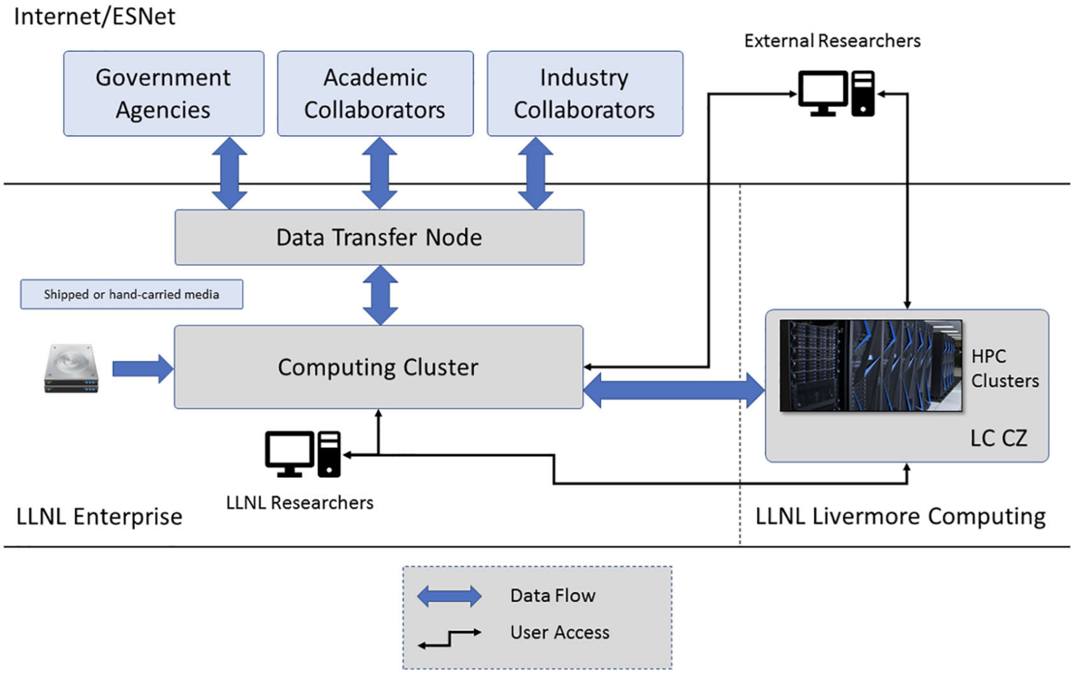

Figure 2. Schematic of the MBA data infrastructure with both a physical enclave (left), including primary compute and data transfer nodes, and an extension to HPC infrastructure (right).

while outbound transfers of finished analysis were supported via the same pathway, these required the additional step of administrator review and approval.

To permit utilization of high-performance computing (HPC) systems, the LLNL secure enclave was extended to include the Livermore Computing Collaboration Zone (CZ; https://hpc.llnl.gov/hardware/zones-k-enclave). This enabled analysis of multiple omics datasets using leadership-class compute platforms such as Mammoth, a 8,800 core cluster acquired via the National Nuclear Security Administration's Advanced Simulation and Computing (ASC) Program.

Quality assurance and quality control of data acquisition

In recent years, various scientific disciplines and consortia have developed minimum standards for the inclusion of data in both centralized repositories and published meta-analyses.[21] These guidelines have encompassed a range of data acquisition formats, including genetic, proteomic, and other biochemical data.[22,23] For example, the Human Proteome Organization's Proteomics Standards Initiative developed the Minimum Information About a Proteomics Experiment (MIAPE) standard for mass spectrometry experiments involving protein and peptide identification.[24] Similarly, in the human subjects field, the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement set minimum metadata standards for collection, archiving, and reporting of epidemiological research data.[25]

Other standards-setting organizations go beyond simple metadata reporting requirements and seek to detail comprehensive processes and systems that can help ensure data quality (quality assurance, or QA) as well as specific tests and benchmarks that can flag errors during and after data collection (quality control, or QC).[26] These types of procedural checks have been adopted, for example, by the Metabolomics Quality Assurance and Quality Control Consortium (mQACC) and the Encyclopedia of DNA Elements (ENCODE) Consortium.[27,28] The MBA program used this type of standard as a model for its own QA/QC efforts.

LLNL experimentalists generated a scoring rubric for each of the molecular and omics data collection modalities employed in the various human trials throughout MBA. Example rubrics are shown in Table 1 and Extended Data Tables 1-3. A pass/fail checklist was used to determine whether each dataset met the minimum quality standards for use by the modeling teams. Some criteria involved best practices in sample handling and study design, while others were specific to the instrumentation used and, in general, followed manufacturer recommendations. For some types of data collection, including genome sequencing data, open source tools such as multiQC were utilized as components of the scoring framework.[29] Following the IV&V team's evaluation of each dataset against the rubrics, a scorecard was

Table 1. Example QA/QC scoring rubric for immunophenotyping using the CytoFlex-S flow cytometer. Reference ranges are derived from 'CytoFLEX Platform Instructions for Use,' Beckman Coulter, rev. 12/11/2019. In addition to the pass/fail ranges, some rubrics included a 'warn' range for borderline data.

Assay Phase	Specific Metric	Pass Range	Fail Range
Sample Collection	Was sample collected properly (tube type, anticoagulant, etc., per SOP)?	Yes	No
Sample Collection	Was sample properly recorded?	Yes	No
Sample Collection	Was sample stored according to proper procedure for the sample type?	Yes	No
Sample Collection	Is an unbroken chain of custody documented, including dates, times, and locations of all custodian changes?	Yes	No
Sample Collection	Were appropriate transportation conditions used, and were temperatures monitored and recorded?	Yes	No
Sample Prep	Is sufficient sample volume present for processing and analysis?	>10 uL each sample	No
Sample Prep	Were samples thoroughly mixed before loading?	Yes	No
Study Design	Number of Technical Replicates	>1	≤1
Calibration	Most recent quality control and standardization of CytoFlex system (with CytExpert QC)	≤1 day	>1 day
Calibration	Were optical filters verified to match the detector configuration?	Yes	No
Calibration	Were QC fluorospheres adequately mixed?	Yes	No
Calibration	Age of QC fluorosphere preparation	≤5 days	>5 days
Calibration	Was gain set in accordance with manufacturer instructions?	Yes	No
Calibration	Was daily cleaning performed in accordance with instructions?	Yes	No
Experimental	Was threshold adequate to exclude sample debris?	Yes	No
Experimental	Was the detection rate appropriate throughout sampling?	< 10,000 events/second	Other
Experimental	Were positive sample concentrations within acceptable range?	$2 \times 10^4 - 2 \times 10^7$ units/mL	Other

transmitted to the performer team or subcontractor responsible for the data collection, and the program office was consulted for a final determination on the inclusion of the dataset in the modeling corpus.

Additionally, UIUC statisticians developed separate sets of metrics for the phenotypic and behavioral data collected during the course of the program. These rubrics were crafted to flag outliers and diagnose other potential data quality issues. The analyses encompassed five general domains: cognition, demographics, human performance, personality, and wearable sensors (see Table 2). Some metrics applied to only a single domain, whereas others (e.g., missing data) were relevant for multiple domains. When there is a quality assurance failure, the Potential Issue column of Table 2 provides plausible mechanisms that may underlie the faulty data collection process. The team also developed customized R software scripts that read in the data and automatically generated tables, figures, and reports.

Test and evaluation of performer modeling products

The primary deliverable from the MBA IV&V effort was the test and evaluation of performer expression circuit models used to predict achievement on military skills tests. While performers trained statistical models to predict pass/fail outcomes for different candidates on a battery of human performance and cognitive tests, the IV&V team was responsible for certifying to the military cadre that the selected molecular observables were in fact predictive of the chosen outcomes.

Several factors complicated the evaluation of performer models according to these criteria, among them: 1) small sample sizes for program cohorts, 2) the potential for subjective evaluation criteria in certain skills tests, and 3) incomplete

Table 2. **Example QA/QC scoring criteria for phenotypic data collection.** Domains Assessed: Cog=Cognition; Demo=Demographics; HP=Human Performance; Per=Personality; WS=Wearable Sensors.

Domain(s) Assessed	Specific Metric	Response Indicative of Failure	Potential Issue
Cog; Demo; HP; Per; WS	Was there missing data for the participant?	Yes	Data missing randomly or systematically.
Cog; Demo; HP; Per; WS	Were there missing data for > 25% of participants?	Yes	Failure in test administration.
Cog; HP; Per	Was a score an outlier, defined by 1.5 times the interquartile range?	Yes	Visually examine violin plot and magnitude of outlier.
Cog; Per	Were any scores outside the allowed range for the test?	Yes	Data recording error.
HP	Did everyone meet the military-defined threshold?	No	Individuals who failed the threshold are excluded from data analysis.
HP	Was an achieved performance metric beyond human limits?	Yes	Measurement error, data recording, or data transcription issue.
HP	For tests with repeated measures under identical conditions, is the Pearson correlation > 0.75?	No	Measurement error, data recording error, or data transcription issue.
WS	Was data collected?	No	Sensor failure or sensor was not used.
WS	Was data collected for expected duration?	No	Sensor was not always turned on or was not used.
WS	Was data in the expected range?	No	Sensor failure or error in data transcription or processing.
Demo	Is there an entry for each field?	No	Data not entered or deleted.
Demo	Is the field entry an allowed value?	No	Data for fields with unallowed values is unusable.

coverage in the acquisition of omics data relative to phenotypic data, for which several years of historical data collection were already available.

To mitigate these complications, we implemented a two-pronged model evaluation strategy consisting of both a qualitative component, based on pre-registration of the key mechanistic hypotheses each performer planned to investigate, and a quantitative component, based on an evaluation of the predictions of each performer model against a held-out validation set of true outcome labels.

Hypothesis pre-registration is a technique used in some disciplines to avoid using the same set of data for both hypothesis generation and hypothesis testing.[30] Hypotheses proposed by MBA performers at the outset of the modeling effort included a variety of potential biological mechanisms underlying task performance, such as sleep quality, metabolism, muscle tone recovery, and several proposed cognitive/psychological mechanisms. These pre-registration documents were retained by the IV&V team for later determination if the identified predictive biomarkers might plausibly correspond to the pre-specified categories.

For quantitative validation, the IV&V team held back 20-30% of the candidate outcome labels from the primary modeling teams during each year of the MBA program. The outcomes for this validation set were kept fully blinded from performer team members to prevent data snooping.[31] Modelers were given all other data from each annual cohort and then asked to submit predictions of the outcomes of the blinded candidates for scoring by the IV&V team. Results were announced at each program review meeting.

Software repository and data archive

Preserving the ability to apply predictive models to new cohorts of individuals following the conclusion of MBA was a key goal of the program. Given the small sizes of individual cohorts, prospective model testing on future data collection was considered a significant component of the overall validation strategy. Furthermore, the IV&V team desired to ensure that, to the extent possible, the models would be independent of the choice of laboratory for omics data processing to avoid vendor lock-in.

To facilitate a single storage location for program data, LLNL data scientists generated a MariaDB database schema to contain all multi-omic, phenotypic, and outcome data collected over the course of the MBA program. As some omics data was too large to practically store within the database itself, the database contained links to the original and processed data files stored in a master data archive. It also contained metadata to track QA/QC results associated with different datasets, as well as to reconcile individual research subjects with their anonymized identifiers.

To support continued usefulness of the predictive models, the IV&V team requested that performers package their analysis and models for future use as research compendia, according to the method of Marwick et al.[32] We chose this format as the R language was the preferred coding environment of the majority of modeling teams. To support long-term portability of the modeling pipelines, containerization of the computing environment using Docker or Singularity was also recommended for each team.

Results

The primary investigator-led teams funded to perform work for MBA offered a high level of cooperation with our IV&V efforts. We were aided by support for our IV&V plan from the research sponsor, particularly when elements of the plan necessitated extra effort by the performer teams, such as in the case of hypothesis pre-registration or periodic data holdbacks.

Data QA/QC added a modest amount of time between the return of results from experimenters and the availability of processed data for use by modelers. However, on several occasions involving both sequencing and mass spectrometry experiments, issues flagged during the QA/QC process spurred additional consultation with the data collection teams and led to process improvement that was incorporated into subsequent data re-analysis.

Regular holdout validation set tests of performer predictive models provided an unbiased, apples-to-apples comparison of model performance that assisted the sponsor in measuring progress against program goals. Unfortunately, program constraints made it difficult to test counterfactual predictions made by the modelers, i.e., predictions that certain individuals would have progressed further in the selection process than they actually did. As a result, measuring improvement over state-of-the-art in quantities such as recall, as envisioned at the outset of MBA, was not possible. Instead, the IV&V team defaulted to the use of prediction accuracy and F-score as the primary endpoints for model evaluation.[33]

Discussion and lessons learned

In recent years, studies have demonstrated that diverse types of omics data are predictive of biological phenotypes supporting human performance characteristics.[34] Nevertheless, this field of research comes with a unique set of challenges that separate it from the much larger pool of clinical research seeking to drive progress in the medical domain. “Success” in the human performance context may be a more multifactorial entity than in the medical context, where it may simply entail the cessation of an identified disease process. Additionally, healthy and, in particular, athletically adept individuals may be more reluctant to participate in invasive specimen collection procedures than individuals already engaged with the medical system.

In working with cohorts that significantly depart from broader population baselines, reference data from publicly available databases may turn out to be of lesser value than modelers initially hope. For example, studies of the metabolic impact of various dietary regimens in aging or pre-diabetic populations may not have high transfer value in the warfighter population. To the extent possible, omics data collection for single individuals over long periods of time may mitigate this issue and limit the need for transfer learning from weakly representative populations.

Alternatively, research sponsors may wish to consider funding short-term but larger multiomic studies that are composed of participants more closely representative of the target population. Phenotypic outcomes could be collected passively and unobtrusively using wearables technology. Cadre members could be polled to determine surrogate endpoints, measurable in this more high-throughput context, that they believe most likely related to their more holistic judgments in the selection process of interest. Additionally, if the surrogate endpoint markers are continuously valued, this type of outcome variable may allow for superior statistical power than dichotomized pass/fail outcome labels.[35]

To participate in blinded prediction contests, modelers may be reluctant to give up scarce training data samples as a validation holdout when the total number of observations in the dataset is small. Statistical techniques that require checking certain prerequisite assumptions, such as the normality of predictor distributions, may become tedious to implement when small amounts of data are released sequentially. The modeler experience might be subjectively improved if there is enough data to constitute multiple test sets, even if some of those are only partially blinded. For example, Kaggle, the competitive data science website, frequently splits datasets into training, "public leaderboard," and "private leaderboard" components, with the first category being fully accessible to modelers, the second providing a basis for competitors to obtain a preliminary score during the competition, and the final category remaining fully blinded until all models have been submitted.[36] Even though the "public leaderboard" data is not truly blinded, since competitors can iteratively query it throughout the model building process, modelers may nevertheless elect to use it judiciously to gauge their performance and to debug basic generalization errors.

Finally, program managers wanting to drive improvements over state-of-the-art outcome prediction, particularly for quantities such as recall, should engage early with program stakeholders to develop means of testing counterfactual predictions made by data scientists. For example, modelers may predict that certain candidates "would have passed" later rounds of a tournament selection process had they been given the opportunity to compete at the higher level. While this information may be of high value from the standpoint of program goals, these predictions are impervious to validation if those individuals are lost to follow-up.

Conclusions

Community-based efforts to promote reproducibility through open sharing of data and code have played an important role in advancing the methodological rigor of many scientific disciplines. We have demonstrated a paradigm for adapting several aspects of this approach to achieve independent verification and validation of results in the context of a research program where unlimited data exchange is not feasible.

Using holdout prediction tests and hypothesis pre-registration, our team was able to certify that predictive modeling benchmarks were achieved in the absence of data snooping. Additionally, a centralized data infrastructure and integrated QA/QC system promoted data integrity and helped to facilitate the preservation of data and algorithms generated in the course of the project for follow-on research efforts.

While our IV&V strategy was developed for projects at the intersection of human performance and defense, we anticipate that similar protocols may prove useful in other research contexts involving multiomic data analysis and sensitive human subjects data. Though the data and trained models from this project are encumbered by distribution restrictions, other artifacts from the study, such as QA/QC rubrics, have been made available to support future work in this area.

Data availability

Underlying data

No data associated with this article.

Extended data

Figshare: Measuring Biological Aptitude Omics QA/QC Rubrics. https://doi.org/10.6084/m9.figshare.23802606.v1.[37]
This project contains the following extended data:

- ExtendedDataTables.pdf (Sequencing, Proteomics, and Metabolomics QA/QC Scoring Rubrics)

Data is available under the terms of the CC-BY 4.0 license.

References

1. Begley CG, Ioannidis JPA: **Reproducibility in science improving the standard for basic and preclinical research.** *Circ. Res.* 2015; **116**(1): 116–126.
Publisher Full Text

2. Robson B: **The dragon on the gold: Myths and realities for data mining in biomedicine and biotechnology using digital and molecular libraries.** *J. Proteome Res.* 2004; **3**(6): 1113–1119.
PubMed Abstract | Publisher Full Text

3. Russo D, Zou J: **How much does your data exploration overfit? controlling bias via information usage.** *IEEE Trans. Inf. Theory.* 2020; **66**(1): 302–323.
Publisher Full Text

4. Gentleman R, Lang DT: **Statistical analyses and reproducible research.** *J. Comput. Graph. Stat.* 2007; **16**(1): 1–23.
Publisher Full Text

5. Morin A, Urban J, Adams PD, *et al.*: **Shining light into black boxes.** *Science.* 2012; **336**(6078): 159–160.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Laurinavichyute A, Yadav H, Vasishth S: **Share the code, not just the data: A case study of the reproducibility of articles published in the journal of memory and language under the open data policy.** *J. Mem. Lang.* 2022; **125**: 104332.
Publisher Full Text

7. Federer LM, Belter CW, Joubert DJ, *et al.*: **Data sharing in plos one: An analysis of data availability statements.** *PLos One.* 2018; **13**(5):

e0194768.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Sholler D, Ram K, Boettiger C, *et al.*: **Enforcing public data archiving policies in academic publishing: A study of ecology journals.** *Big Data Soc.* 2019; **6**(1): 205395171983625.
Publisher Full Text

9. Easterbrook PJ, Berlin JA, Gopalan R, *et al.*: **Publication bias in clinical research.** *Lancet.* 1991; **337**(8746): 867–872.
Publisher Full Text

10. Turner EH, Matthews AM, Linardatos E, *et al.*: **Selective publication of antidepressant trials and its influence on apparent efficacy.** *N. Engl. J. Med.* 2008; **358**(3): 252–260.
PubMed Abstract | Publisher Full Text

11. Raphael MP, Sheehan PE, Vora GJ: **A controlled trial for reproducibility.** *Nature.* 2020; **579**(7798): 190–192.
Publisher Full Text

12. Berger B, Peng J, Singh M: **Computational solutions for omics data.** *Nat. Rev. Genet.* 2013; **14**(5): 333–346.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Edwards PN, Mayernik MS, Batcheller AL, *et al.*: **Science friction: Data, metadata, and collaboration.** *Soc. Stud. Sci.* 2011; **41**(5): 667–690.
PubMed Abstract | Publisher Full Text

14. Levin N, Leonelli S, Weckowska D, *et al.*: **How do scientists define openness? exploring the relationship between open science policies and research practice.** *Bull. Sci. Technol. Soc.* 2016; **36**(2): 128–141.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Boehnen C, Bolme D, Flynn P: **Biometrics irb best practices and data protection.** *Conference on Biometric and Surveillance Technology for Human and Activity Identification XII, volume 9457 of Proceedings of SPIE, BELLINGHAM, 2015. Spie-Int Soc Optical Engineering.* 978-1-62841-573-5.
Publisher Full Text

16. El Emam K, Jonker E, Arbuckle L, *et al.*: **A systematic review of re-identification attacks on health data.** *PLos One.* 2011; **6**(12): 12.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Do N, Grossman R, Feldman T, *et al.*: **The veterans precision oncology data commons: Transforming va data into a national resource for research in precision oncology.** *Semin. Oncol.* 2019; **46**(4-5): 314–320.
PubMed Abstract | Publisher Full Text

18. Navale V, Ji M, Vovk O, *et al.*: **Development of an informatics system for accelerating biomedical research.** *F1000Res.* 2019; **8**: 1430.
Publisher Full Text

19. Barnes C, Bajracharya B, Cannalte M, *et al.*: **The biomedical research hub: a federated platform for patient research data.** *J. Am. Med. Inform. Assoc.* 2022; **29**(4): 619–625.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Braunschweiger P, Goodman KW: **The citi program: An international online resource for education in human subjects protection and the responsible conduct of research.** *Acad. Med.* 2007; **82**(9): 861–864.
PubMed Abstract | Publisher Full Text

21. Liberati A, Altman DG, Tetzlaff J, *et al.*: **The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration.** *Ann. Intern. Med.* 2009; **151**(4): W65–W94.
PubMed Abstract | Publisher Full Text

22. Sumner LW, Amberg A, Barrett D, *et al.*: **Proposed minimum reporting standards for chemical analysis.** *Metabolomics.* 2007; **3**(3): 211–221.
Publisher Full Text

23. Fostel JM: **Towards standards for data exchange and integration and their impact on a public database such as cebs (chemical effects in biological systems).** *Toxicol. Appl. Pharmacol.* 2008; **233**(1): 54–62.
PubMed Abstract | Publisher Full Text

24. Taylor CF, Paton NW, Lilley KS, *et al.*: **The minimum information about a proteomics experiment (miape).** *Nat. Biotechnol.* 2007; **25**(8): 887–893.
PubMed Abstract | Publisher Full Text

25. von Elm E, Altman DG, Egger M, *et al.*: **The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies.** *Lancet.* 2007; **370**(9596): 1453–1457.
Publisher Full Text

26. Groboth G: **Quality assurance in testing laboratories.** *J. Therm. Anal. Calorim.* 1999; **56**(3): 1405–1412.
Publisher Full Text

27. Beger RD, Dunn WB, Bandukwala A, *et al.*: **Towards quality assurance and quality control in untargeted metabolomics studies.** *Metabolomics.* 2019; **15**(1): 4.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Dunham I, Kundaje A, Aldred SF, *et al.*: **An integrated encyclopedia of dna elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Ewels P, Magnusson M, Lundin S, *et al.*: **Multiqc: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Van't Veer AE, Giner-Sorolla R: **Pre-registration in social psychology-a discussion and suggested template.** *J. Exp. Soc. Psychol.* 2016; **67**: 2–12.
Publisher Full Text

31. Roelofs R, Fridovich-Keil S, Miller J, *et al.*: **A meta-analysis of overfitting in machine learning.** *Advances in Neural Information Processing Systems 32 (Nips 2019).* 2019; **32**: 11.

32. Marwick B, Boettiger C, Mullen L: **Packaging data analytical work reproducibly using r (and friends).** *Am. Stat.* 2018; **72**(1): 80–88.
Publisher Full Text

33. Zhang E, Zhang Y: *F-Measure.* Boston, MA: Springer US; 2009; 1147.
Publisher Full Text

34. Kim DS, Wheeler MT, Ashley EA: **The genetics of human performance.** *Nat. Rev. Genet.* 2022; **23**(1): 40–54.
Publisher Full Text

35. Royston P, Altman DG, Sauerbrei W: **Dichotomizing continuous predictors in multiple regression: a bad idea.** *Stat. Med.* 2006; **25**(1): 127–141.
PubMed Abstract | Publisher Full Text

36. Bojer CS, Meldgaard JP: **Kaggle forecasting competitions: An overlooked learning opportunity.** *Int. J. Forecast.* 2021; **37**(2): 587–603.
Publisher Full Text

37. Benjamin J, Morrison Michael D, Thissen James B, *et al.*: **Measuring biological aptitude omics qa/qc rubrics.**
Publisher Full Text

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 22 November 2023

https://doi.org/10.5256/f1000research.154122.r220464

?

Michael Fitzsimons
The University of Chicago, Chicago, Illinois, USA

The authors share their approach to performing an Independent Verification and Validation (IV&V) process for the Measuring Biological Aptitude (MBA) project, which is funded by the Biological Technologies Office of the Defense Advanced Research Projects Agency (DARPA). They discuss IV&V as a general use case for supporting research reproducibility in a field where data cannot be directly shared. They also provide the specifics of their implementation including QA/QC and qualitative and quantitive evaluation of model performance. I found the article interesting and informative, but I think it is lacking some context that would help readers to understand more about IV&V. In particular, I would like to see considerably more detail about how their approach differs from or improves up other IV&V examples. A cursory review of the literature shows many papers on the topic and it would be helpful to understand how the authors' approach differs and compares.

I would also love to see any more details about their actual model evaluation techniques. I presume the sponsors constrains what they are allowed to share, but at present it is somewhat difficult to assess the rigor of their verification and validation process.

Additional small and targeted comments are included below:

Introduction

Data snooping - I don't think this term is understood by all. It might be helpful to add another sentence or phrase about data snooping.

Reachback tasking is not a standard term as far as I know and should be modified or explained.

Independent verification and validation (IV&V) - I would like to see a lot more discussion of what this typically looks like and how your implementation compares. A casual review of the literature shows lots of citations about "independent verification and validation (IV&V)" - how does your paper add to this literature?. I cannot tell if your implementation is good, bad, or typical. Given we cannot see any actual modeling results (I assume this is a restriction by the sponsor), I think it is important to include a broader comparison to the existing IV&V literature.

Methods

Standards discussion is good.

Could you really ascertain the answer to some of the QA/QC questions such as "Were collection tubes labeled with unique, traceable codes?" and "Were samples thoroughly mixed before loading?"

Can you include any more details about the actual models or their evaluation? This is for the quantitative evaluation. Or maybe at least say explicitly what the limitations of what you are allowed to provide.

Results

Are there any results, or at least the format for results, that you can share?

Other suggestions

Mention privacy preserving federated learning as another option for incorporating data that cannot be shared directly.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, data sharing solutions, software

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.