

Observational research with Muse models – What models accelerate and where expertise still matters

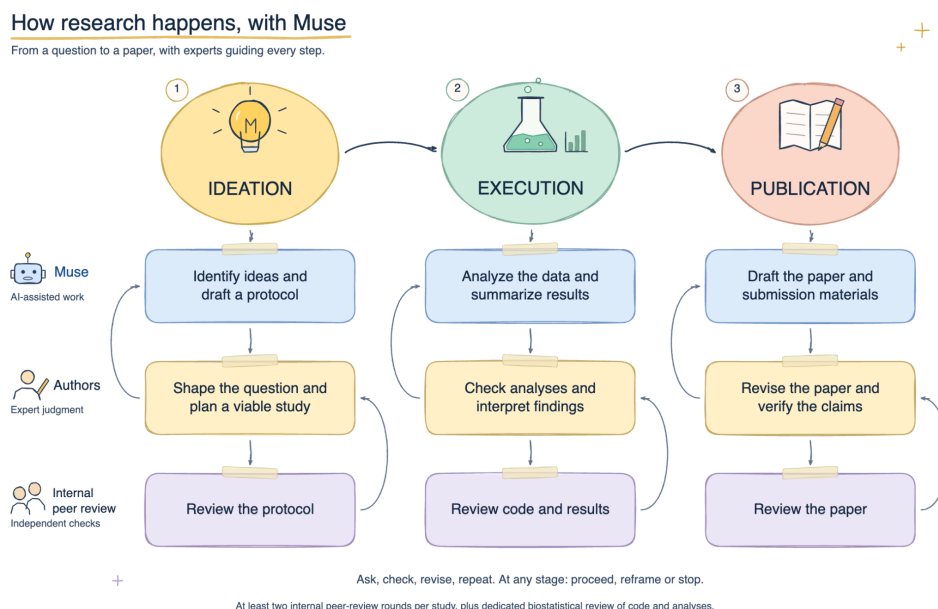
S. Michelle Ogunwole, MD, PhD; Daniel Seung Kim, MD, PhD; MPH, Kritika Singh, PhD; Sam Ashoo, MD; Felipe Lobelo, MD, PhD; Arturo Loaiza-Bonilla, MD, MSED; Jared Durieux, PhD, MS; Christos Colovos, MD, PhD

Observational research on real world data such as electronic health records has become a principal engine of clinical evidence, yet the process that turns such data into a published finding remains slow and largely manual. A single study routinely takes months to develop, execute, analyze, and write up. This process passes through many hands – a principal investigator, a fellow, a data scientist, a biostatistician – each contributing judgment that is rarely documented in a form another team could reproduce.

AI is a powerful tool that can accelerate the research lifecycle by automating parts of research that are most time-intensive and laborious for humans – literature searches, protocol drafting, coding, and report generation. To maximize this efficiency, however, investigators must clearly map the new division of labor, identifying where the model can operate autonomously and where human expertise remains the critical driver of quality.

A research sprint with Muse models

Over several weeks, our team of physician-scientists, biostatisticians, and AI developers used Muse models to take research ideas from hypothesis to manuscript. The team used a coding agent powered by Muse Spark 1.3 and Muse Wonder to carry a project through each stage of research, from ideation through execution to manuscript writing (Figure 1). Each study underwent at least two rounds of internal peer review, plus a dedicated biostatistical and computer scientist review of all code and analyses.



[Figure 1.](#) **Human-AI collaboration from research question to publication.** High-level overview of the complementary roles of Muse, investigators, and peer reviewers during ideation, execution, and publication. The process is iterative: at each stage, a study may proceed, be reframed, or be stopped.

Ten observational studies emerged from the sprint, and many additional ideas are in progress. The studies included in the final sprint drew on two kinds of data. Some used de-identified, longitudinal US Electronic Health Record and claims data with 4.9 million patients. Others used publicly available sources, including National Electronic Injury Surveillance System emergency department data combined with public attention signals from Google Trends, Wikipedia pageviews, and GDELT (Global Database of Events, Language, and Ton) news volume.

One practical benefit of the process was being able to assess ideas before investing in full analyses. Muse models quickly flagged questions already addressed in the published literature, exposed mismatches between research questions and the available data, and proposed alternative study designs that could produce more novel contributions. This allowed investigators to redirect or stop some projects early. Given how much research time is routinely lost to studies that never reach publication, the sprint's clearest gain was not faster drafting but faster termination. By rapidly identifying flaws or a question already answered, the model acted as a critical time-saver, and the hours that would have been spent on them were reallocated to the surviving studies.

What we learned along the way

Muse models are fast, capable, and often help make the work more rigorous and thorough. What they cannot supply on their own is judgment: the clinical, epidemiologic, and statistical wisdom to choose the right question, spot erroneous definitions, and decide whether a result should be trusted. That is where human expertise mattered most. Rather than slowing the work down, human expertise pointed the work in the right direction and identified weakness early, while problems could still be fixed, leading to stronger, higher-quality work. Appendix Table 1 summarizes this division of labor across the research process, with examples of where Muse helped, where human expertise was essential, and how collaborative checks improved the work.

AI readily maps the structure of a dataset, but it cannot judge the clinical value of an analysis nor always identify inherent data limitations. Observational research fundamentally struggles to guarantee if a finding is both empirically true and medically useful. Without human guidance, a model will simply execute commands and produce technically accurate but practically useless inferences. The investigator's domain expertise acts as the essential compass, steering the AI away from sterile data exercises and toward clinically relevant insights.

We learned that a useful AI framework needed two components to work across studies. The first was the workflow itself: the steps of a study, and how each step should be carried out. The second, which was easier to miss, was a data contract: a single shared definition between structured and unstructured data – detailing its contents, how records link, and how studies must access it – with checks that enforce those rules. Lacking this centralized framework, individual studies would be forced to independently decipher and reconstruct the underlying data architecture, resulting in highly redundant and inefficient workflows. Implementing this shared schema allowed Muse models to utilize standardized analyses across studies, ensuring teams did not reinvent the wheel for every new project.

The system behaved as a high-throughput, unevenly reliable research assistant: strong at drafting, extraction, and code generation - including analytic methods beyond a given investigator's expertise - and unreliable at judging whether a source existed, whether a recorded field measured the intended construct, and whether its own output was correct. It was, on documented occasions, assertively incorrect, with no uncertainty signal a reviewer could use. More than once, the model told us something wasn't available in the dataset when it actually was; a clinician who knew it should be there, or had used it in another project, identified the error.

We did not let the models' choices stand unexamined. When Muse models picked a cutoff, a variable definition, an exclusion, or a statistical method, we asked why and what evidence supported the choice. Often the first output was something the investigator did not yet understand or accept. The work improved through the exchange that followed: probing, supplying clinical knowledge, disagreeing, and at times asking for the model's opinion, until the investigator could be confident in the methods. While Muse models produced fluent first drafts, investigators revised heavily for clarity, accuracy, and argument. That back-and-forth is where the investigator's contribution most clearly shaped the quality and impact of the work.

Above all, accountability always remained with the investigator. Our process required the investigator to verify each consequential step and ultimately ensure that the speed gained from working with the model did not result in merely high-volume, low-quality work. Instead, investigators assumed final responsibility – ensuring that the power of Muse models served to root out weaknesses, making the resulting work more rigorous, valid and acceptable to the scientific community.

Where this approach sits among AI research systems

Systems for AI-augmented research now sit on a continuum defined by how much of the research process the model is allowed to decide. At one end are fully autonomous "AI scientist" systems that ultimately aim to fully remove the human from the loop and, in doing so, risk unverifiable citations, opaque provenance, and the displacement of scientific judgment. At the other end are tools that automate a single stage, such as evidence synthesis or review screening, and leave the arc from question to manuscript ungoverned as a whole. We deliberately placed ourselves between the two. Muse models accelerated every stage, and an investigator governed each one with prespecified gates and checks at critical points in the process.

We offer our experience as a candidate best practice rather than a finished standard. It still requires researchers who know how to move a study from question to dataset to analytic approach to interpretation. Observational evidence, however rigorous, approximates rather than replaces randomized control trials. It has important value in clinical/medical research and especially in population health and real-world evidence exploration where randomized controlled trials are not feasible. What this approach changes is time. Work that once took months of data cleaning, or weeks to answer a reviewer, can substantially advance in days. Further, this approach allows for the democratization of the scientific process by providing researchers a model to act as expert consultants that smaller, less well-funded, research groups struggle to access. Ultimately, this lets more investigators test the full reach of an idea before committing to it, and spend more of their effort on the questions most worth asking.

Acknowledgements

Support to use the Meta Muse AI platform for this study was provided by the Meta AI engineering team, including Ashish Shenoy, Benoit Corda, DJ Petersen, Hongru Zhu, Jason Wei, Kay Ji, and Xilun Chen.

Appendix

Table 1. Complementary Roles of Muse and Expert Judgment Across the Research Workflow

Stage	Where Muse helped	Where human expertise was essential	Example of Collaborative Checks and Balances	Next Iteration
Ideation and novelty	Searched and synthesized the literature then proposed a short list of testable hypotheses that had passed initial feasibility checks. Scored each idea on novelty, whether the exact question had already been answered, and actionability, whether the available data could answer it Flagged when a teammate was pursuing a similar hypothesis	Judged whether a question was meaningful and clinically important, not only novel and actionable Recognized questions arising from patient care and new developments in the field Read beyond what the model could access, because it could read open-access articles in full but only summaries of subscription articles, and those summaries can leave out limits and unanswered questions	Two investigators were exploring GLP-1 medicines for diabetes and weight management in relation to fertility Muse flagged that they were working on similar questions They compared the questions and found that the aims were distinct, so both proceeded without duplication	Full-text access to subscription journal articles through publisher agreements, rather than summaries alone
Can the data answer the question?	Tested whether the exposure, outcome, and covariates could be measured before the long work of analysis began	Decided whether "measurable" meant "valid" for the specific question Checked whether a recorded field truly measured the condition or treatment the study meant to study Checked whether information was missing more often in one group than another, which would make the comparison unfair	Muse recognized that dates in Protégé had been shifted to protect patient privacy, so the true calendar date and season were not known It flagged that a flu vaccine question of interest to the researcher could not be answered in this dataset, because flu vaccine outcomes depend on season. That recognition saved time	Given a question, recommend which of many available datasets can best answer it, accounting for each dataset's limitations
Data quality and limitations	Ran direct queries quickly to quantify missingness, completeness, and collection artifacts Counted missing information and surfaced unusual patterns in the data	Asked the right question, because a clinician who knows the expected pattern can spot an implausible number. Directed the next query to determine whether the pattern reflected missing data, incomplete capture, or how the variable was collected. Recognized that missing information is not the same as absence of disease	Muse calculated 1% mortality rate for patients with <i>Staphylococcus aureus</i> bacteremia, a serious bloodstream infection An infectious disease physician knew that published studies report mortality closer to 30%, so the team stopped and audited how death was recorded which revealed other implausible patterns	Collaboratively assess the quality of large datasets in a rapid, systematic manner Develop best practices for a data quality check skill that can be applied broadly

Stage	Where Muse helped	Where human expertise was essential	Example of Collaborative Checks and Balances	Next Iteration
Defining variables	Proposed definitions, time windows, and cutoffs, and explained the trade-offs among them	Matched the definition to the clinical construct of interest Chose definitions that kept the comparison fair, using the same kind of information to define each group Made sure eligibility at the start of the study used only information available at that time	Heparin can be given as a tiny flush to keep an intravenous line open or as a full dose to treat a blood clot Both records name the same drug, but they are different exposures. The Model counted them as the same exposure The investigator defined exposure by dose and route, so a flush was not counted as treatment and improved the accuracy of the exposure definition. This would have been a critical error if left unchecked. Imaging results were often pasted into progress notes instead of filed as formal reports, so counting only formal reports would have missed real cases.	Addition of a discussion skill, inviting the model to ask questions and review data fields with investigators conversationally, to prevent assumptions and improve analysis.
Advanced methods and adversarial review	Explained specialized specialty designs, such as target trial emulation, and helped draft the analysis code while keeping track of the assumptions each method required Produced structured critiques of sources, methods, interpretation, and citations before the manuscript was drafted	Chose a method suited to the question, the data, and the audience Checked that the effect the study meant to estimate matched the calculation the code actually performed, and that the reported uncertainty matched both. Checked that the protocol, the code actually run, and the claims in the manuscript described the same analysis	A method section states one target effect, while the implemented weights target another. The reviewer traces the code and outputs, identifies the source of the disagreement, and determines whether the remedy concerns code, reporting, or the intended comparison	A structured review assistant that follows an experienced reviewer's checklist and links each recommendation to the relevant code or result, while the investigator makes the final decision
Clinical meaningfulness	Offered a reasonable first interpretation of the findings	Judged whether the result was large, precise, and credible enough to matter to patients or change care, rather than deciding from statistical significance alone Considered whether bias or the patients studied	Practicing clinicians disagree about a cancer question, so investigators used their knowledge of current treatment choices to judge whether the benefit was large enough to matter in practice	Clinicians who do not conduct research could surface questions from practice for research teams to vet and study

Stage	Where Muse helped	Where human expertise was essential	Example of Collaborative Checks and Balances	Next Iteration
		limited how broadly the result could be applied	They did not decide from statistical significance alone	
Publication	Assessed journal fit candidly and built submission materials. Turned reviewer-requested sensitivity analyses around in a day rather than weeks	Verified every citation against its source. Checked that the Methods described the analysis actually run, and that the results and claims matched. Supplied the interpretation that shaped the Discussion	AI engineers built a model to predict the odds of acceptance into a top tier and/or specialty journal. This tool helped experts gauge if the potential impact of their science and if the idea was worth improving or pivoting from	A research assistant/coordinator that finalizes all the publication-ready documents and goes through submission quality control checks