2024 GENIE Data Model  
  
RFP: Questions and Answers (7/31/2024)

1. What is the project timeline?
   1. Completion by Q2 2025
2. How many patients are in the source dataset. How many years average of patient data?
   1. There are currently ~185k patients in the registry. Retrospective collection of data elements is typically captured from initial cancer diagnosis, as such, some patient timelines may be greater than 20 years. However, most of our genomic data spans the dates 2013-present.
3. Please describe the registry process today. (1) How is data shared, aggregated and harmonized? (2) What is the user experience? (3) What about this process is working/not working? (4) What tools are being used?
   1. More information and current workflow can be found [here](https://www.aacr.org/wp-content/uploads/2024/07/data_guide.pdf). The current process is well established and defined. The focus of this work is to build and (optionally) facilitate deployment in our infrastructure of a unified data model that facilitates programmatic data collection. Our tech stack includes REDCap as our EDC, Synapse (Sage BioNetworks) to store, harmonize, and provide access/compute, and cBioPortal for visualization.
4. Given the registry is already live, why are you looking to do the clinical abstraction data model now? How do you currently get data?
   1. Past studies deployed different, but related data models and dictionaries. A common base data model closely aligns with our strategic plan and facilitates scaling and better interoperability with other datasets, among other benefits.
5. For the components of the data collection that may need to be completed with the assistance of manual curation, will that be done by clinical staff at each individual cancer center?
   1. Yes
6. Will AACR provide samples of the actual data for the data model construction?
   1. Yes
7. Who will be responsible for training cancer center clinical staff on the manual curation of content into eCRFs? The vendor or AACR GENIE?
   1. AACR GENIE. If the vendor is able to produce curation directives and training materials, please include in the proposal; however, this is not a requirement.
8. Source data is coming from different organizations. Is it normalized to a single format at AACR? If not, how many different formats are there?
   1. The data is harmonized; for more information about the collection and harmonization of Project GENIE data please see the AACR Project GENIE data guide [here](https://www.aacr.org/wp-content/uploads/2024/07/data_guide.pdf).
9. How many new partners/centers will be encouraged in the Genie initiative in the future?
   1. GENIE is always considering new applications to become a consortium member. There is no defined maximum amount of consortium sites allowed in the project.
10. What security and compliance requirements exist?
    1. Our current data security and compliance processes are also outlined in the AACR Project GENIE data guide [here](https://www.aacr.org/wp-content/uploads/2024/07/data_guide.pdf).
11. Is design of data protection and data governance layer in scope for this project?
    1. No, these exist and the proposed solution will be deployed within this infrastructure. We welcome feedback on our existing governance if weaknesses are uncovered during the work.
12. Are users accessing the data on a pay per use model, for example to fund the programs at AACR and/or cover infrastructure costs.
    1. The current business model has no bearing on the scope or implementation of the RFP.
13. Do you have use cases already defined which will drive the scope and granularity of the data model?
    1. Yes, you may review more about how the GENIE data are utilized [here](https://www.aacr.org/professionals/research/aacr-project-genie/aacr-project-genie-publications/).
14. Have the data elements been specified and is the data dictionary built? Does GENIE contain PHI
    1. Yes, a high-level overview of the data elements have now been listed on the RFP webpage. GENIE does not contain PHI - all data are de-identified.
15. For the "other data standards or value sets", will AACR provide the terminology sets to load into the data model?
    1. AACR has defined data standards and value sets. However, we’re open to additional suggestions or recommendations by your team.
16. How often does AACR anticipate this resource being updated? Constantly or once a year? From a scope perspective, are you interested in a model for long term support and further development of the data model or is the scope limited to the initial development only?
    1. The scope of this RFP is focused on the initial development of a model; technologies that facilitate long-term maintenance and continued development by our existing technological partners will be prioritized.
17. Structured vs unstructured data and Manual Curation:
    1. The final work product must support both ingestion of structured as well as unstructured data through both curation and NLP/ML/AI methodologies.
18. If curation is in scope of this project, how much does AACR anticipate manual versus automatic curation? Do they have any idea how standardized the data already is they want to incorporate? If curation is in scope of this project, are there any data processing steps it would be useful to include as part of the curation? For example, automated histopathology results could be generated from image files.
    1. We have not yet determined the relative contributions of manual vs programmatic data collection. Please include any proposed solutions for programmatic data collection for consideration.
19. For the unstructured data, do you envision the use of NLP?
    1. There are parallel efforts underway to collect unstructured data programmatically. Please include any proposed solutions in your response for consideration.
20. Is the data model intended to encompass all 12 cancer types or have 12 separate data model products?
    1. The request is to build a ‘core’ data model with additional cancer-specific modules contained within the same data model solution.
21. Does the model need to be extendable to other cancer types than the ones mentioned in slide 7?
    1. yes, but is not part of this initial scoping. Currently the focus is on a solid tumor cancer model and the 12 initial cancer types.
22. What is current development environment at AACR? Is it cloud based (AWS/Azure/Google) or internal network?
    1. It is a mixture of US-based on-premise and AWS
23. Are there tools/platforms that must be integrated with this solution?
    1. Synapse (via Sage BioNetworks).
24. Are there any existing vendor relationships or technologies that we should consider in our proposal? Is the Synapse infrastructure non-negotiable or can alternative providers be accepted?
    1. Use of Synapse is non-negotiable; however, we will review suggestions for alternative solutions must also be compatible with REDCap.
25. Does the final solution, along with data collection and modeling, for this RFP include a platform for data storage and management, and for users to perform analyses?
    1. No
26. What EHR Systems (types), Registries are utilized by Genie data partners/centers ?
    1. Most use EPIC EHR; however, the sources are varied and pipes to the sources is out of scope for this RFP.
27. Would AACR be able to provide further detail on the desired format of the data model and the final format of the harmonized datasets? Would those datasets ultimately be stored in databases? If yes, would these databases be MongoDB (NoSQL) or relational SQL databases?
    1. The ultimate target for the data is a relational database, currently SQL. Any solution should also be compatible with REDCap.
28. Per RFP, the model should establish a common language and framework and support research by providing a framework for analyzing data. Given this instruction, is the need only for a GENIE model or for a duo of model + data analysis framework?
    1. Currently for this proposal, we’re interested in just a GENIE model (data schema).
29. Do you seek just that data schema model and UML, or do you need a working database with testing and real-world manual curation?
    1. Currently, we’re looking for assistance with developing the data schema model and UML.
30. Do you need a machine-readable data schema, or do you also need the algorithm to use the machine-readable data schema?
    1. We need a machine-readable data schema, preferably in an UML.
31. Is there a preferred format for the "machine-readable data schema"?
    1. No
32. Is data harmonization or automatic data curation expected as a part of the deliverables? Related: what are expectations for the "Modular and editable derived variable code using python to support collected aligning to the AACR GENIE data model framework"?
    1. Data harmonization is out of scope. "Modular and editable" refers to structuring the data model/schema in a way that allows integration of individual modules for analysis while remaining distinct for individual updates. For example, cancer diagnosis data should be kept separate and distinct from treatment data within the code itself such that changing or editing one module’s code will not adversely affect the code of other modules.
33. Does AACR have requirements for data consumption layer? In other words, how does AACR want the data to be presented for downstream analysis or consumption?
    1. We do not. The data will be stored in SQL tables and ingested into analytic layers from there. Providing the model in UML will allow mapping into our analytic layers.
34. OMOP and FHIR are different standards, have you chosen one and if so, what process will be used to decide which model to focus on? Both models are incomplete models for oncology and mapping from OMOP and FHIR is not a 1-1 mapping.
    1. The data schema should be built to the FHIR standard. If you also have a solution for mapping the produced schema to OMOP, please include in your proposal, however, it is not necessary for the data schema to be in the OMOP format.
35. What is the role of HL7/FHIR: is it extraction of HL7/FHIR into OMOP, or creation of an HL7/FHIR model?
    1. Creation of a HL7/FHIR model
36. Will it be required that all source data from participating cancer centers be provided in FHIR or OMOP CDM?
    1. The intention is to collect data programmatically (FHIR based) and map the final dataset to OMOP, but it is not necessary for the data schema to be in the OMOP structure.
37. Could you provide more details on how you envision integrating OMOP CDM, HL7 FHIR, and other data standards (e.g., NAACCR, ICD-10, AJCC) within the clinical data model?
    1. We have already specified the required data elements, value sets, and standard ontologies/standards (e.g., NAACCR, ICD-10, AJCC). The scope of this request is generation of a data schema and model, an optional book of work is to map this model to OMOP; however, it is not a requirement for a successful application.
38. You mention widely accepted standards (e.g., OMOP, CDM, HL7) - do you intend to build an infrastructure only for a restricted set of clinical models (restricting inclusion to GENIE only to those conforming with those models) or is the intent to be able to ingest any type of data models? Also, Regarding OMOP and FIHR, do we expect ingestion of data in FIHR only and delivery of queriable data in OMOP? Or ingestion and query in both?
    1. See answer to #34, standards have been selected and will be iterated as appropriate. If your proposed solution can functionally ETL any model into the model created in this SOW, then please include for consideration; however, it is not a requirement for a successful application. We expect data ingestion via FHIR; mapping to OMOP is optional, please include in submitted proposal if your team has this capability.
39. Regarding Testing and Validation, does AACR want an ongoing validation process that can be reviewed and update the mapping based on results? (For example: a report that identifies how well variables are populated, changes over time, and introduction of new data formats)
    1. AACR will consider inclusions of an ongoing validation processes with the proposal.
40. Do you envision the genomic data to be transformed into OMOP?
    1. While we intend to map the finalized, collected data (including genomic data) to OMOP, it is not necessary for the data schema to be in the OMOP structure.
41. How important is to AACR GENIE team to contribute any model work back to the community (e.g., Standard Development Organizations, terminology SDOs (such as SNOMED international (for SNOMED CT) or Regenstrief Institute (for LOINC), HL7 )? Should the proposal only utilize existing routine healthcare or healthcare research standards and leave the extension of such standard to the larger community) or within reason, try to actively advance?
    1. Proposed solutions should focus on current healthcare/research standards. If an opportunity to advance/expand/iterate a community standard arises, we could explore engagement with the respective organization(s) as appropriate.
42. There is indication in RFP that at least some of the data is in HL7 format. Is it CCDA? Is there FHIR dataset available also?
    1. Currently our data are not in HL7 format.
43. Are vendors expected to actually deploy the developed models into physical databases or FHIR APIs?
    1. Currently not in the scope of this request.
44. Does the Vendor have to provide the specification details for the data model, or is it a collaboration process between the Vendor and AACR experts?
    1. Specification details should be provided, with the ability of AACR staff and associated team to make recommendations or suggestions
45. What, if any, types of molecular and genomic data might be included in or linked to the clinical data model? This is relevant because mutational signature data already known to stratify cancer subtypes (ex HER+, etc), may be more readily incorporated as a clinical feature than something like RNA-seq or scRNA-seq that may be linked to a patient-timepoint. It would be helpful to know more precisely what types of data are expected to advise solutions.
    1. Project GENIE currently collects Next-Generation Sequencing (NGS) data for all patients. You can learn more in the data guide [here](https://www.aacr.org/wp-content/uploads/2024/07/data_guide.pdf). Future efforts will focus on incorporating RNA-seq or single-cell RNA-seq (scRNA-seq) data. A proposal that can accommodate the integration of these types of molecular and genomic data, in addition to NGS, will be highly beneficial. Standard-of-care clinical tests, such as HER2 testing, PSA, etc should be modeled as clinical variables.
46. Given that the OMOP CDM is developed by the OHDSI open-source community, how do you plan to license the clinical data model that will be built through this project?
    1. We plan to make the data model available to the community through creative commons licenses.
47. Are you open to a co-ownership of the code base and IP produced during the project?
    1. Our intention is to make this available to the community through creative commons license.
48. Will the model be open source or closed source?
    1. Open source
49. Additionally, do you intend to share the model and any related tools or resources with the broader community, continuing the spirit of openness and collaboration that defines the GENIE Project?
    1. Yes
50. What communication and project management tools will be used?
    1. AACR Project GENIE Team currently utilizes Box and Freedcamp for communication and project management, respectively, but are open to other tools as needed.
51. For the proposal, are we allowed to include an appendix section with supporting information (e.g., detailed CVs etc)
    1. yes
52. Will AACR GENIE entertain a proposal submitted by two vendors in partnership
    1. Yes
53. If multiple vendors can submit as a single solution, is the response limit 10 pages total or 10 pages for each vendor?
    1. 10 pages total
54. Can you provide us some details on the current stakeholder environment that the chosen vendor's team will have to collaborate with, including AACR staff, volunteers, any other vendors etc?
    1. Other stakeholders beyond the AACR staff will include cBioPortal and Sage BioNetworks supporting partners.
55. How does AACR anticipate collaboration between AACR, its participating centers, and the vendor? Will there be a working group that is established by AACR?
    1. Yes. AACR has already established a working group for this project.
56. Will there be AACR colleagues available to support the core project team? If so, please describe their role(s)
    1. Yes. AACR has a staff of project managers and a clinical data manager that will be supporting the project, in addition to technical individuals through Sage Bionetworks.