TransMed Evaluation Plan

This evaluation plan is based on some of the components of the RFI response created by the Monarch and Orange teams available here. It also includes the use case/competency question documents for the Fanconi Anemia and Undiagnosed Disease demonstrators, which are available within linked documents.

TRANSLATOR DEMONSTRATORS

The Translator must be benchmarked against the demonstrators to demonstrate overall data content and connectivity progress in the data integration processes. As a reminder, each demonstrator is described briefly below. Specific documents related to each demonstrator are linked, and more general Translator evaluation criteria reference the demonstrators further below where relevant.

Fanconi Anemia (FA)

FA is a rare, complex disease that exhibits phenotypes seen in both rare and common diseases with mutations in at least 21 genes. Phenotypic features are highly variable but typically include bone marrow failure (BMF), a range of developmental abnormalities, accelerated ageing, and cancer. The FA genes include two of the major breast cancer susceptibility genes (BRCA1, BRCA2) and several other newly-recognized cancer susceptibility genes. Knowledge of FA is limited in that the majority of FA genes have no known function and few recognizable functional protein motifs. Recently, it was determined that the FA pathway is required to handle the damage to DNA from reactive aldehydes from endogenous and exogenous sources (e.g., formaldehyde and acetaldehyde). As a result, FA patients who also have genetic variants that reduce their ability to detoxify reactive aldehydes have greatly increased risk for bone marrow failure and certain cancers. Eight percent of the world's population has a variant in the alcohol dehydrogensase gene, ALDH2*2, with reduced ability to detoxify acetaldehyde, which is an endogenous metabolic product as well as a metabolite of alcohol. The ALDH2*2 phenotype has been recognized as the "Asian Flushing syndrome". Increased alcohol consumption by individuals with ALDH2*2 is associated with increased cancer risk. What is not known is if FA gene variants that subtly reduce function of FA proteins would predispose to cancer, developmental abnormalities, and bone marrow dysfunction. FA variants that reduce the function of the FA proteins might also coexist in individuals with variants in ALDH2 and other enzymes (e.g., ADH5) that process reactive aldehydes. A key feature of this use case is that evidence for interaction of ALDH2*2 variants in Japanese FA patients has been shown and will serve as a positive control.

To assess the ability to shed light on the mechanism of action of FA genes/pathways and on the clinical correlations between genotype, phenotype, and environmental variables, we have created a full complement of competency questions. These competency questions (CQs) will provide the capability to benchmark the Translator's data content (e.g. do we have the sources we need to answer the CQs) and data connectivity (e.g. are the data integrated in such a way so as to support answering the CQs). See below for additional and more specific aspects of the use the FA demonstrator in the evaluation.

Rare and undiagnosed dysmorphologies and inborn errors of metabolism.

The diagnosis of very rare diseases is extremely difficult for a number of reasons, such as lack of knowledge relating to the diversity in clinical manifestations or of candidate human variants. There is simply a lack of data available to inform variant prioritization. Many commercial and research tools exist to help prioritize variants from trios (parents and child) of Whole Exome or Whole Genome Sequence files (WES or WGS), but most of these tools focus on variant frequency, inheritance patterns, and prediction of variant pathogenicity. We and others have demonstrated success in diagnosing rare and undiagnosed diseases using integrated genomic and phenomic data. For the Translator, we hypothesize that integration of additional features, such as metabolomics data, external influences including treatments, drugs, diets, and exposures, and the intersection (symptom-based and mechanistic) with common diseases to the algorithms and tools will improve diagnostic capabilities and the potential for treatment identification for rare diseases.

We will evaluate two cohorts of patients from the NIH Undiagnosed Disease Network, OHSU, and JHU: a) Patients with a known rare disease diagnosis, and b) Patients with no diagnosis. We will create phenotype profiles based on HPO terms, obtain clinical analysis data, and where possible, metabolomics profiles and treatment data. These will serve as the entry point into the Translator, allowing the detection and matching of patterns to existing patients in the Translator. For example, a metabolome profile may match the predicted effect of perturbing a biochemical pathway, and orthologs of genes in that pathway may present with matching phenotypes in other organisms (e.g. an undiagnosed Shprintzen-Goldberg patient). We will investigate mechanisms, prognoses, and treatments for candidate variants (from WES, WGS, or inferred) for both the undiagnosed and diagnosed cohorts. We will document a diversity of patients genotypic, phenotypic, and environmental characteristics for benchmarking purposes. The first patient examples are documented here. Essentially, each patient is query set, against which we can compare Translator results over time, for example, ranking of variants.

Synthetic Data

We are in the process of creating generalizable tooling to generate synthetic clinical data and to make those data accessible. We intend to provide a repository for synthetic datasets that can enable end-end testing and evaluation of query scenarios, specifically demonstrator queries. We aspire to enable this resource to address training and evaluation needs of all members of the Translator consortium, support open-access to the tooling and derivative synthetic datasets. Such open, transparent access would not be possible with de-identified or limited datasets deriving from real clinical sources. What distinguishes our synthetic data is that it is drawn from the statistical distributions that underlie real clinical data.

Initially this tooling will emulate clinical parameters found in typical EHR environments. We are leveraging major national efforts for synthetic clinical data, such as the Mitre Synthea project. We expect to support generating datasets according to specified parameters for demographics and disease distributions. We plan to contribute datasets that emulate the natural history and clinical course of translator demonstrator conditions (as one type of synthetic dataset).

Finally, pertinent to the NCATS Translator, we will document specifications for generating synthetic datasets for common conditions and those clinical conditions specific to Translator projects such as Fanconi Anemia, Asthma, and Diabetes. Documentation will include parameters for disease distribution and will be deposited in a publicly accessible repository/library.

Evaluation plans

- Review and approval by TransMed team: synthetic datasets are easily accessible and have clear documentation sufficient for their use
- Review and approval by TransMed team: ensure that training materials are appropriate for the audience (e.g., prereq knowledge & skills needed to use)
- Review and approval by Translator Teams: ensure that synthetic datasets are realistic (i.e., simulate the natural history and clinical course of demonstrator conditions)
- Test synthetic datasets based upon requirements (i.e., can our tooling be used to generate useful synthetic datasets?). Motivation for use:
    - To assess how many varying conditions & circumstances are there?
    - To assess what is the generalizability of clinical insights from Translator demonstrators to different EHR patient populations?

DATA CONTENT

Documentation

- The metadata for each component data source should be robust with respect to generic attributes, such as timestamps, versions, counts, etc. The metadata should be specific with respect to source content, provenance, and attribution. These attributes will be assessed for each constituent data source within the Translator.
- The APIs are well-documented, and populated with examples.
- Data versioning and change history is well documented, and policies for change management are documented

Domain Content and connectivity

Having diverse data in the same warehouse is a good starting point, but it does not make the data inherently more usable or integrated. Data connectedness can be a measure of computational power across diverse data, and is the ultimate goal of the Translator.

- The number of sources integrated into the Translator will be documented over time, and will follow the timeline set out in the milestones.
- Graph metrics such as size of graph, number of sources, number of inter and intra data source connections (triples across sources), number of relationship types between different entity types, Information Content (IC) and other graph or link association measures of granularity and complexity, etc. For example, each included data resource can be conceived of as a sub-graph, and can be measured in terms of nodes, edges, and quantified how well-connected it is to other sub-graphs (sources). Less connectivity is only an indicator of quality in cases where the information is highly novel.
- For data sources, qualified links between related entities in other systems are provided. For instance, unqualified database cross references ("DB xrefs") could mean that an entry is related to another, derived from another, more general or more specific than another, etc. Lack of description of why/how records are related has led to others integrating based on false assumptions. Evaluating these for each source is an important aspect to the methods utilize for each source's integration into the Translator.
- T-score. We will experiment with combining different data attributes, such as graph metrics (as above), performance statistics (e.g. query time, time to refresh graph from sources), and connectivity across sources classified in different ways (semantic data type, qualified DB xrefs, and source) to create an innovative Translational (T) Score, where the T-Score measures connectivity of clinical sources with basic science sources, for example, how well diseases and symptoms connect to basic mechanisms.

Quality of data content

Both quantitative and qualitative methods will be utilized to evaluate quality of data content.

Usability testing by the evaluation team will be utilized to quantify more nuanced determinations, such as “easy to find” or “well documented” should be encouraged.

- **Comprehensiveness:** Resource is as comprehensive as it needs to be in order to be answer the use cases and competency questions. These will be automatically benchmarked against the Translator over time, as new sources, data integration methods, and algorithms are added.
- **The data model complexity** is appropriate to the described demonstrator use cases and the API allows the data complexity be put to its full use
- **Unit tests and data reports:** A suite of unit tests and data reports will be developed and enhanced with each release, and will result on a compliance report for each. Examples of unit tests: a) Are there any entities from two sources where their primary IDs are equivalent, but in each source they map to multiple non-equivalent concepts? b) Are there any equivalent IDs with different semantic types (e.g. Gene vs. Disease)? Examples of data reports are: a) Report all deprecated IDs and compare to prior load. b) Report all new triples between existing entities (e.g. the relationship has changed).
- **Data Status:** Transparency regarding presence of data that may be included but not yet quality assured is indicated.
- **Probability**: Transparent with regard to probabilities, where relevant (for example, text mined associations).

Text-mined BEL pathway fragment data

Traditional approaches to extracting knowledge from text include expert-based, supervised learning, and pattern learning—all of which can be expensive to apply. Here we will apply supervision, active learning and topic modeling to take advantage of a large amount of raw text, information redundancy across sources, and existing knowledge to semi-automatically acquire NLP semantic parsing intelligence. Causal pathway fragments will be captured using BELExtractor, a pattern-based system to extract knowledge and codify it into Biological Expression Language (BEL) representation. BELMiner achieved an overall best F-measure of 44.6% in extracting the complete BEL statement from evidence sentences provided in Biocreative Shared Task V in 2015. For relation extraction, the system achieved an F-measure of 69.54% on a blind test data set.

For our evaluation, we will first create a gold standard set of annotations for the Fanconi Anemia demonstrator, including annotations to anatomy, genes, gene functions, gene interactions, and phenotypes. A corpus of manuscripts will be identified, of these, 10% will be randomly chosen for markup as the gold standard. We will then analyze the remaining 90% of manuscripts with BELExtractor, and compare results against the prior Biocreative results. We will extend the framework to also capture other kinds of associations, e.g. between alcohol consumption, cellular mechanisms, and developmental defects; these will be added to the gold standard annotation and again compared against the unannotated corpus.

Provenance and Attribution

The data's provenance and attribution policies are well documented:

- Derived content is credited using its original identifier(s) and linked using some persistent mechanism (eg. PURL, identifiers.org, etc.)
- Data processing/transportation provenance is tracked using systems such as PROV or the W3C Dataset description.
- For the complex integrated data within the Translator, provenance information should be available via APIs, as graphs, or other mechanisms. We will incrementally evaluate new methods for providing this information via the APIs.
- Contributions to the content (data, tools, algorithms, sources, etc.) are clearly declared.
    - The contributor, author or data source's organizations are attributed using identifiers, logos, and other references to source content
    - Individual people / institutions / grants etc. are referenced with identifiers where relevant, such as from Wikidata; some examples are also: ORCID or ResearcherID for people; Digital Science GRID or OCLC for organizations

DATA ACCESS

The Translator should provide data access options for all or parts of the data corpus via one or more well documented mechanisms. For the API, the following highly-recommended implementation guidelines are recommended. Direct database endpoints (e.g. MySQL, SPARQL etc) can be valuable; however, expertise in using these varies. Therefore, it is important to also wrap these with an API wherever possible. A summary of important REST principles is below; see also SSI REST best practices here.

- **API:** Application Programming Interface (API) for the data exists and is well documented (above)
- **RESTful**: Follow RESTful API pattern
- **JSON:** Return JSON if possible, TSV if not
- **Retrieval:**
    - Allow retrieval of a single record by using its identifier
    - Allow batch retrieval of a list of data entities using a list of identifiers
- **Indexed**: Contents are indexed and optimized to support the most common types of queries (e.g. those of the competency questions)
- **Dumps:** Whole/partial database dumps are available (where appropriate)
- **Downloads:** Slices of the database and individual records can be downloaded (e.g. as JSON/XML/tab delimited, etc.)
- **Versioned:** A versioned URL pattern is provided to support future API changes
- **Uptime:** API uptime is reported.

IDENTIFIERS AND INTEROPERABILITY

How the data is identified, both at an individual entity level (for example, a “gene”) as well as larger datasets or subsets (for example, all human genes from Entrez), is crucially important (see this community declaration). Note that each of the following attributes will be evaluated for each source as well as for the Translator itself. It is possible that in the time allotted that lack of documentation from the sources will lead to gaps in what the Translator can provision. We see this as a positive outcome in that the Translator will be able to provide this feedback back to the data providers.

- **Semantics/data structure:**
    - A data dictionary
    - Defined schema or data model
    - Services are well aligned to the model and consistent across various access mechanisms
    - Structure, format, architecture, and metadata for the repository is consistent with community norms or shared specifications (for example, use of the W3C Dataset Description)
- **Exchange standards:**
    - Data are made accessible using common exchange formats, if applicable (for example, use of the HL7 FHIR standard or OMOP for exchanging healthcare information electronically)
    - Data elements are well-defined using metadata standards (e.g., ISO/IEC 11179, DDI and SDMX/ISO17369)
    - Value set services and value set definition services use the Common Terminology Services 2 (CTS2) standard
- **Ontologies:** All ontologies in use are documented and are consistently applied to the data.
    - Novel ontologies, if any, are registered in public standards repositories (such as the OBO Foundry Library) and released via standard well documented mechanisms (for examples ROBOT or the OBO Starter Kit)
    - Appropriate community standards/vocabularies are used to record metadata; preferably standards that are: a) designated or *de facto* standards within the relevant domain, and b) free to use, see also Licensure section
    - Version of the ontologies used is indicated
    - Ontologies are attributed according to community best practice
- **Identifiers:**
    - Identifier strategy, prefix, and patterns are well documented and consistently used
    - No embedded meaning or reliance on it for uniqueness
    - Simple, durable web resolution. We will create a web endpoint that accepts any common identifier and returns all properties/edges within the Translator.

- ○ An identifier version-management policy exists
- ○ URIs are clear and findable
- ○ IDs are not reassigned or deleted
- ○ Derived content references original identifier

LICENSURE

Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available. This is true for almost all existing NIH-funded resources. Some widely-used examples of resources that are commonly thought of as “open” but in practice cannot easily be derived and redistributed are:

> *ClinVar: “This site contains resources which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws…..All persons reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder.”*
>
> *OMIM: “This site contains resources which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws…..All persons reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder.”*
>
> *PharmGKB: “PharmGKB grants use of its contents for research purposes. The use of this data and knowledge is NOT available for redistribution... This content is freely available to researchers in academia and industry for RESEARCH PURPOSES. Absent the issuance of a license by Stanford, the content shall not be used in any non-research commercial application in any form."*

We believe that there needs to be better awareness of the impacts of data license choices among both resource providers and NIH program staff. Moreover, few databases produce just data; most also produce software source code, algorithms, and applications. There should be licenses explicitly covering each of these products. Critically, for the Translator to be successful, the barriers to data reuse and re-dissemination MUST be overcome.

We propose a license rating of 1 to 5 stars for each data source based on the following issues. Our goal for the Translator itself will be to provide a 5 star rating for each.

- **Documented:** Explicit data use terms (ideally formal licenses) should be defined by the resource providers and easy to find and understand.A variety of specific examples of data use/reuse conditions should be included.
- **Minimally restrictive:** The licenses and/or data use agreements should explicitly permit downstream data reuse, derivation, and re-dissemination. Licenses should not require negotiation and licenses themselves should be legally re-distributable without engaging legal counsel. CC0 licensure is an example of a minimally restricted license.

- **Standard licenses.** We note that considerations for data are significantly different than those for software and they must be considered separately (see this blog for example).
    - **Standard data license:** For example, public domain (PDDL or CC0), GPL, ODBL, CC-BY, CC-ND, GNU, etc.
    - **Standard software license:** For software, ideally Apache version 2. Note that software license choices are the subject of much community discussion especially regarding “copy-left” approaches and there are other valid standard options available (such as GPLv2, GPLv3, AGPLv3, etc.)
    - **Flowthrough:** Documentation about which source resources/data, if any, come with flowthrough implications. Links to the original licenses/data use terms of all redistributed content are provided. It is currently commonplace that such terms do not exist; in such cases, it should be clearly stated that license/terms could not be found. If specific authorization has been obtained for redistribution, this should be indicated.
- **Contactable:** There should be an appropriate person available for contact with questions about licensure; this person’s contact information should be easy to find.