





**MINI-REVIEW**

# Progress toward a universal biomedical data translator

**Karamarie Fecho**<sup>1,†</sup>  | **Anne E. Thessen**<sup>2,†</sup>  | **Sergio E. Baranzini**<sup>3</sup> | **Chris Bizon**<sup>1</sup> | **Jennifer J. Hadlock**<sup>4</sup>  | **Sui Huang**<sup>4</sup> | **Ryan T. Roper**<sup>4</sup> | **Noel Southall**<sup>5</sup> | **Casey Ta**<sup>6</sup> | **Paul B. Watkins**<sup>7</sup> | **Mark D. Williams**<sup>5</sup>  | **Hao Xu**<sup>1</sup> | **William Byrd**<sup>8</sup>  | **Vlado Dančik**<sup>9</sup>  | **Marc P. Duby**<sup>10</sup> | **Michel Dumontier**<sup>11</sup> | **Gustavo Glusman**<sup>4</sup>  | **Nomi L. Harris**<sup>12</sup>  | **Eugene W. Hinderer**<sup>13</sup> | **Greg Hyde**<sup>14</sup> | **Adam Johs**<sup>15</sup>  | **Andrew I. Su**<sup>16</sup>  | **Guangrong Qin**<sup>4</sup> | **Qian Zhu**<sup>5</sup> | **The Biomedical Data Translator Consortium**<sup>‡</sup>

<sup>1</sup>Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

<sup>2</sup>Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

<sup>3</sup>Weill Institute for Neurosciences, Department of Neurology, University of California at San Francisco, San Francisco, California, USA

<sup>4</sup>Institute for Systems Biology, Seattle, Washington, USA

<sup>5</sup>Division of Preclinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland, USA

<sup>6</sup>Department of Biomedical Informatics, Columbia University, New York, New York, USA

<sup>7</sup>Division of Pharmacotherapy and Experimental Therapeutics, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

<sup>8</sup>Hugh Kaul Precision Medicine Institute, University of Alabama at Birmingham, Birmingham, Alabama, USA

<sup>9</sup>Chemical Biology and Therapeutics Science Program, Broad Institute, Cambridge, Massachusetts, USA

<sup>10</sup>Medical and Population Genetics Program, Broad Institute, Cambridge, Massachusetts, USA

<sup>11</sup>Institute of Data Science, Maastricht University, Maastricht, The Netherlands

<sup>12</sup>Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, California, USA

<sup>13</sup>Tufts Clinical and Translational Science Institute, Tufts Medical Center, Boston, Massachusetts, USA

<sup>14</sup>Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire, USA

<sup>15</sup>Department of Information Science, College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania, USA

<sup>16</sup>Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA

**Correspondence**

Karamarie Fecho, c/o Renaissance Computing Institute, 100 Europa Drive, Suite 540, Chapel Hill, NC 27517, USA.  
Email: [kfecho@copperlineprofessionalsolutions.com](mailto:kfecho@ copperlineprofessionalsolutions.com) and [kfecho@renci.org](mailto:kfecho@renci.org)

**Funding information**

This work was supported by the National Center for Advancing

**Abstract**

Clinical, biomedical, and translational science has reached an inflection point in the breadth and diversity of available data and the potential impact of such data to improve human health and well-being. However, the data are often siloed, disorganized, and not broadly accessible due to discipline-specific differences in terminology and representation. To address these challenges, the Biomedical Data Translator Consortium has developed and tested a pilot knowledge graph-based

<sup>†</sup>These authors served as co-first authors.

<sup>‡</sup>Consortial/collaborative authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Clinical and Translational Science* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

Translational Sciences, Biomedical Data Translator Program (Other Transaction Awards OT2TR003434, OT2TR003436, OT2TR003428, OT2TR003448, OT2TR003427, OT2TR003430, OT2TR003433, OT2TR003450, OT2TR003437, OT2TR003443, OT2TR003441, OT2TR003449, OT2TR003445, OT2TR003422, OT2TR003435, OT3TR002026, OT3TR002020, OT3TR002025, OT3TR002019, OT3TR002027, OT2TR002517, OT2TR002514, OT2TR002515, OT2TR002584, and OT2TR002520; Contract number 75N95021P00636). Additional funding was provided by the National Center for Advancing Translational Sciences, Intramural Research Program (ZIA TR000276-05) and the National Institute of Diabetes and Digestive and Kidney Diseases (5U01DK065201).

“Translator” system capable of integrating existing biomedical data sets and “translating” those data into insights intended to augment human reasoning and accelerate translational science. Having demonstrated feasibility of the Translator system, the Translator program has since moved into development, and the Translator Consortium has made significant progress in the research, design, and implementation of an operational system. Herein, we describe the current system’s architecture, performance, and quality of results. We apply Translator to several real-world use cases developed in collaboration with subject-matter experts. Finally, we discuss the scientific and technical features of Translator and compare those features to other state-of-the-art, biomedical graph-based question-answering systems.

## INTRODUCTION

The breadth and diversity of biomedical data available today hold great promise in the application of such data into actionable outcomes aimed at accelerating translational science and ultimately improving human health and well-being. Indeed, advancements in computing and storage capabilities have fostered a wealth of large data sets across translational domains. Translational scientists now have unprecedented access to data and knowledge on genes, biological pathways, chemicals, metabolites, drugs, diseases, environmental exposures, clinical health-care records, and more. However, the inherent power of the available data has not been fully harnessed due to long-recognized challenges related to the compartmentalization of data into separate domains, the lack of widely adopted standards or the adoption of standards that are domain-specific, and noncompliance with the principles of findability, accessibility, interoperability, and reusability (FAIR).<sup>1</sup>

The Biomedical Data Translator program (“Translator program”) was launched in Fall 2016 by the National Center for Advancing Translational Sciences (NCATS) in an effort to overcome the many challenges that have long hindered translational science. The vision of the Translator program is to augment human reasoning and accelerate scientific discovery “through an informatics platform that enables interrogation of relationships across the full spectrum of data types.”<sup>2</sup> To achieve this goal, NCATS rapidly and adeptly established a diverse community of nearly 200 basic and

clinical scientists, informaticians, ontologists, software developers, and practicing clinicians distributed over 11 teams and 28 institutions to form the Biomedical Data Translator Consortium (“Translator Consortium”). The Translator Consortium adheres to several core principles that have allowed the program to make considerable progress toward a shared vision: namely, team science; a bottom-up management approach; and open-source community-contributed software development. (See Figure S1 for complete timeline and notable milestones.)

The Translator Consortium last reported on the program in two 2019 publications.<sup>3,4</sup> The aim of this review is to provide an update on the Translator program. We first review approaches for knowledge representation in translational science. We then describe the technical solution that the Translator program has converged on. We demonstrate real-world use-case applications of the prototype Translator system (“Translator”). Finally, we end with a discussion of next steps and a comparison between Translator and similar systems.

## KNOWLEDGE REPRESENTATION IN TRANSLATIONAL SCIENCE

### “Knowledge” versus “data”

The distinction between “knowledge” and “data” is most often captured as the data-to-information-to-knowledge-to-wisdom transformation or DIKW pyramid.<sup>5</sup> Although the origins of this hierarchical representation model are

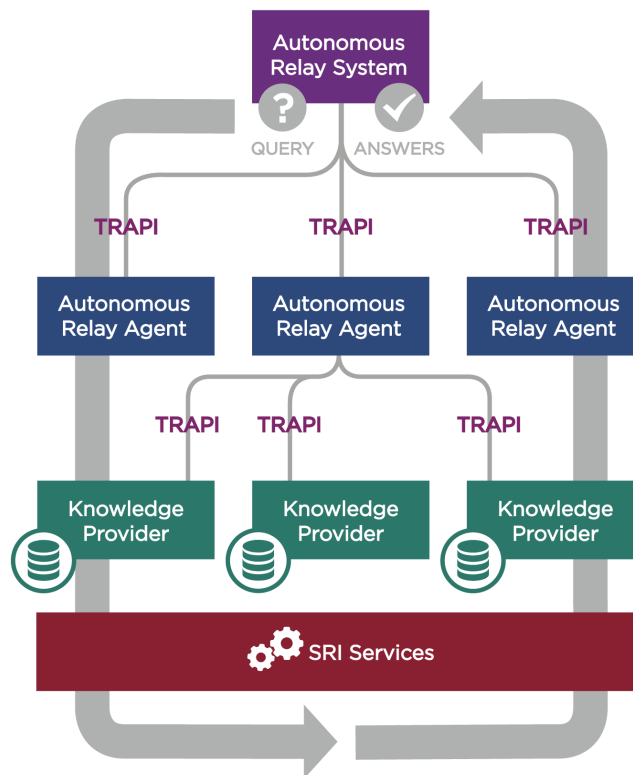
uncertain, and other knowledge representations exist,<sup>6</sup> the DIKW framework has been widely used in fields like information science, communications science, and library science. Within this hierarchical framework, data are viewed as abundant and characterized as discrete objective facts or observations; information is considered to be assertions derived from data and intended to provide interpretation of the data; knowledge is viewed as generally accepted, universal assertions derived from the accumulation of information; and wisdom is considered to be the most abstract layer of understanding derived from assertions and insights into acquired knowledge.<sup>7</sup>

## Approaches for knowledge representation

Application of the conceptual DIKW framework has focused primarily on knowledge discovery, or the systematic process whereby observations or data are organized and interpreted into information that is then scrutinized or tested in the context of existing knowledge, with any subsequent assertions disseminated for peer consensus and adjudication before being accepted as new knowledge. Approaches for knowledge discovery date back to ancient times and form the foundation of the scientific method.<sup>8</sup> Approaches for knowledge representation likewise date back to ancient times.<sup>8</sup> Early forms of modern peer-reviewed publication represent one approach to knowledge representation that remains in use today.

## Knowledge graphs

In recent years, “knowledge graphs” (KGs) have become a common approach for knowledge representation in a variety of fields.<sup>9,10</sup> In a KG, entities or data types are represented as nodes and connected to each other by edges with predicates that describe the relationship between entities. A “schema” is used to constrain the KG by specifying how knowledge can be represented; as such, it provides a framework for validating specific instances of knowledge representation through rules that dictate the syntax and semantics. KGs allow users to pose questions that can then be translated into query graphs and applied to identify subgraphs within the KG that match the general structure of the query graph, thereby producing answers to user queries and generating new knowledge.<sup>11</sup> KGs have had many successful applications, with Google’s KG<sup>10</sup> perhaps the most widely known.



**FIGURE 1** Overview of the Translator architecture. Note that while the high-level architecture depicted in the figure is accurate, certain components may deviate slightly from the architecture in their approach to implementation. Abbreviations: SRI, Standards and Reference Implementation; TRAPI, Translator Reasoner Application Programming Interface. (Graphic prepared by Kelsey Uργο).

## THE TRANSLATOR SOLUTION

The Translator Consortium has adopted a federated KG-based approach for biomedical knowledge representation and discovery (Figure 1).

Translator comprises four main components: Knowledge Providers (KPs); Autonomous Relay Agents (ARAs); an Autonomous Relay System (ARS); and a Standards and Reference Implementation Component (SRI).

The objective of KPs is to contribute domain-specific, high-value information abstracted from one or more underlying “knowledge sources,” which may be raw data as defined by the DIKW framework or information that has been abstracted from the data. ARAs build upon the knowledge contributed by KPs by way of reasoning and inference and in response to user-defined queries. In addition, ARAs may independently expose information abstracted from data. The ARS functions as a central relay station between ARAs and broadcasts user queries to the ARAs. The SRI services are responsible for the

development, implementation, and community adoption of the standards needed to achieve the overall goals of the Translator Consortium.

Translator leverages integrated data from over 250 knowledge sources, each exposed via open application programming interfaces (APIs). The knowledge sources include, among others, highly curated biomedical databases such as Comparative Toxicogenomics Database,<sup>12</sup> and ontologies such as Mondo, the Monarch Disease Ontology.<sup>13</sup>

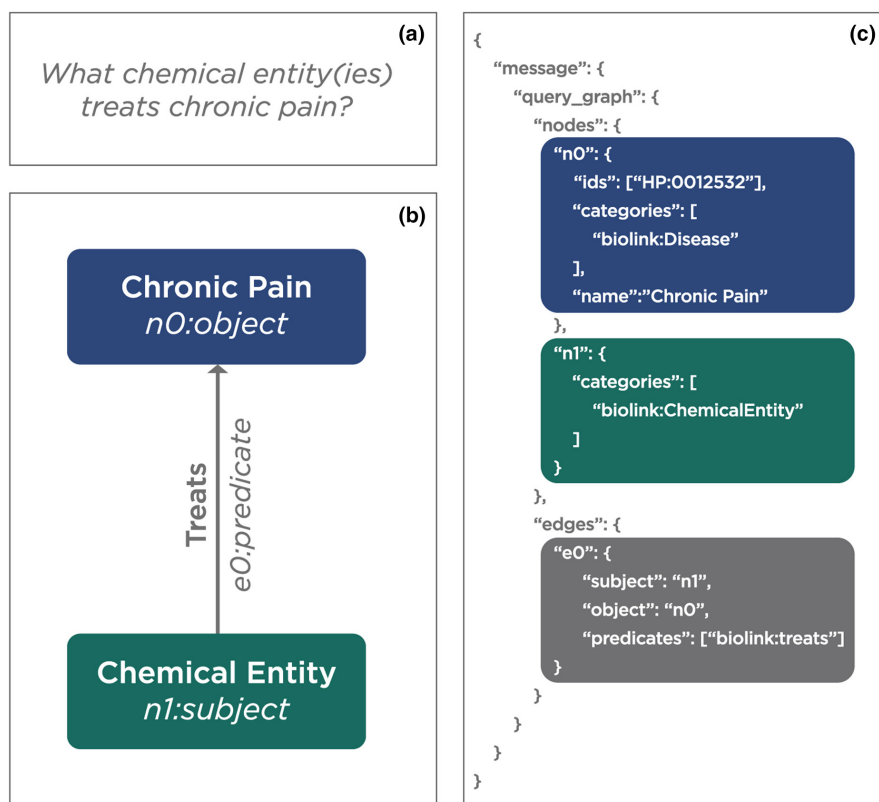
In addition, Translator openly exposes data derived from several electronic health record (EHR) systems, clinical registries, and clinical studies, from which future medical knowledge can be generated: Columbia University Irving Medical Center; UNC Health; the nonprofit Providence Health System; the Drug Induced Liver Injury Network (DILI Network); the Personalized Environment and Genes Study within the National Institute of Environmental Health Sciences; the Institute for System Biology's Wellness cohort; and select cancer cohorts from within The Cancer Genome Atlas. Of importance, the Translator clinical KPs do not expose raw clinical data, but rather aggregated or semi-aggregated data and statistical associations or machine learning predictions derived from clinical data, in full compliance with all federal and institutional regulations.<sup>14</sup>

The Translator Consortium has adopted several tools and approaches to support standardization,

harmonization, and interoperability across the diverse Translator system. First, all Translator services are accessible via APIs. The APIs are standardized in their metadata, structure, and operations using the Translator Reasoner API (TRAPI) standard,<sup>15</sup> which defines a standard HTTP protocol for transmitting queries and receiving answers, with both structured as graphs. Second, all Translator services are registered in the SmartAPI registry,<sup>16</sup> thus adhering to FAIR principles. Third, the open-source Biolink Model<sup>17–20</sup> provides an upper-level graph-oriented universal schema that facilitates semantic harmonization and reasoning across disparate knowledge sources.

With these standards in place, users can query across the numerous data sources that are accessible via the federated Translator system. To demonstrate, we provide a simple example. Suppose a user asks *what chemical entities treat chronic pain?* The user is thus asking about approved drugs and other chemicals that may treat chronic pain. To answer this question, the user question must first be translated into a TRAPI-compliant directed query graph, structured in JSON format, with Biolink Model node and edge types specified and a compact unique resource identifiers (CURIE) used to constrain one node (Figure 2).

In this query, “chronic pain” is specified as a *biolink:Disease* type node *n0* with the CURIE HP:0012532, which is defined by the Human Phenotype Ontology as “chronic pain.” A second node *n1* is specified only as a *biolink:ChemicalEntity* type. Nodes *n0* and *n1* are related by



**FIGURE 2** An example of a natural language question translated into a TRAPI directed query graph in JSON format. (a) the natural language question: *what chemical entity(ies) treats chronic pain?* (b) the natural language question represented as an object-predicate-subject “triple.” (c) the TRAPI query that was executed by Translator. TRAPI, Translator Reasoner Application Programming Interface. (Graphic prepared by Kelsey Urgo).



an edge with the relation defined by a predicate specified as *biolink:treats*. The query graph is thus structured to ask *what chemical entity(ies) treats chronic pain?* The query graph is then sent to the ARS, which parses the query and distributes it to the ARAs. The ARAs then distribute it to those KPs that have provided a meta-graph within the SmartAPI registry indicating that they are able to respond to queries of this type. The ARAs may apply a variety of sophisticated reasoning and inference algorithms to the answers returned by the KPs, including different approaches for ranking and scoring answers such as weighting by supporting publications or abstract co-occurrence of subject and object nodes. Finally, the ARS compiles the ARA results for the user.

A review of the answers to the query finds expected answers such as oxycodone, hydrocodone, codeine, lidocaine, and ibuprofen. There are also answers that are accurate but may not be responsive to the user's query such as methadone, which is used to treat opioid dependence,<sup>21</sup> and caffeine, which is an adjuvant in certain pain medicine formulations.<sup>22</sup> In addition, the answer set includes perhaps unexpected answers such as naloxone and naltrexone, which are opioid antagonists. An examination of the evidence and provenance that Translator returns in support of these answers identifies publications in the form of PubMed identifiers (PMIDs), with links to PubMed abstracts that suggest that these compounds may

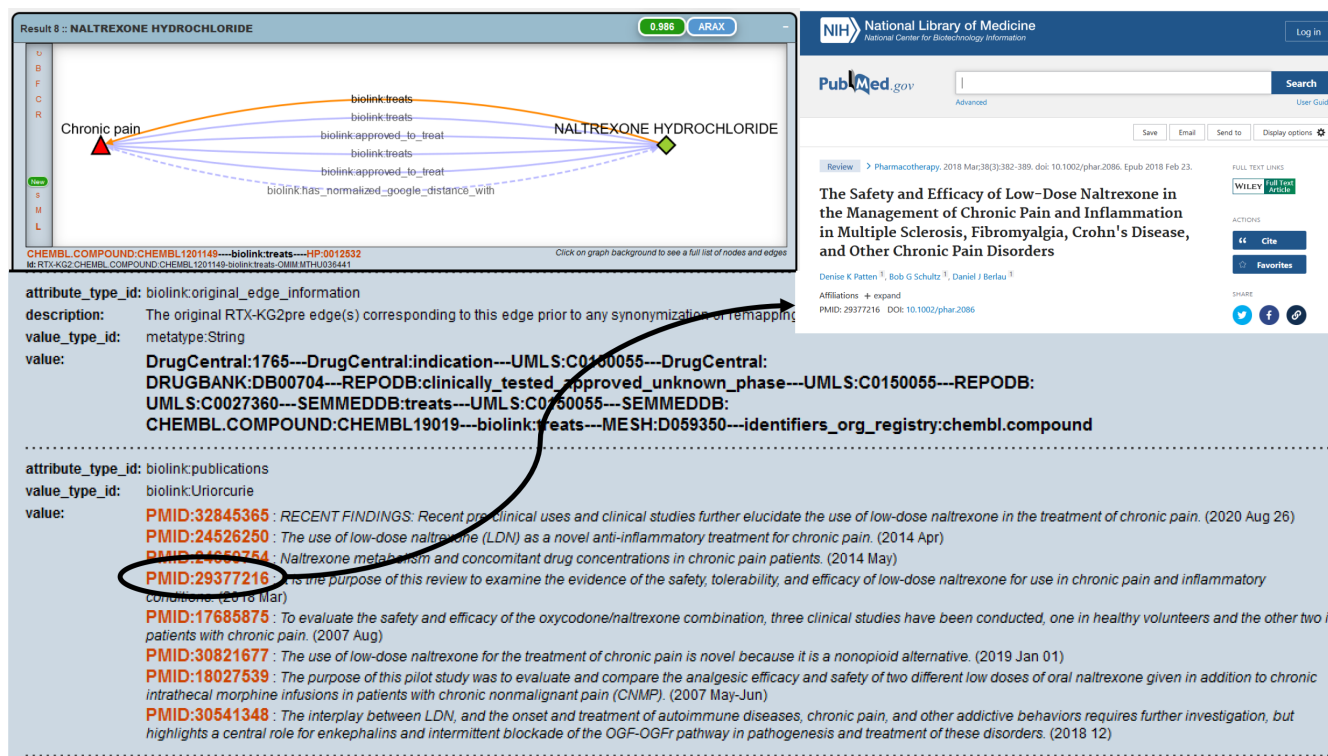
be effective in the treatment of chronic pain conditions such as fibromyalgia and inflammatory bowel conditions (Figure 3). Although a pain specialist may not find these findings surprising, many users likely would be surprised to find that there are cases in which an opioid antagonist is beneficial in the treatment of pain, for which opioid antagonists are often administered.

## APPLICATION USE CASES

The chronic pain use case illustrates basic Translator functionalities in the context of a simple “one-hop” Translator query (i.e., two nodes connected by one edge) and the types of insights and discoveries that the Translator Consortium intends to achieve. Here, we provide an overview of three additional use cases (Figure 4).

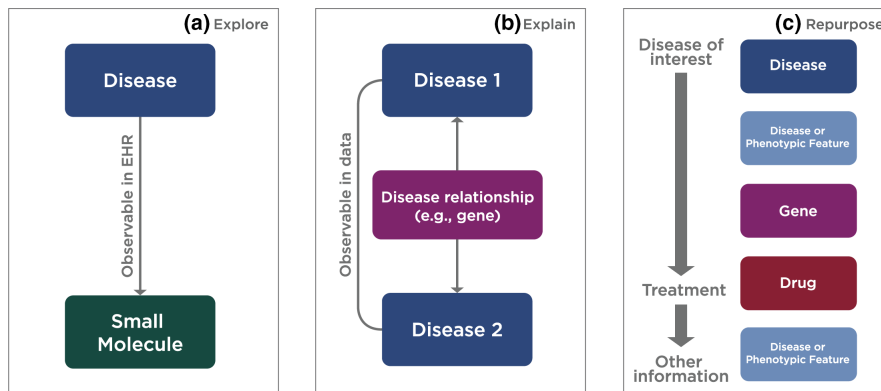
### Explore: Immune-mediated inflammatory diseases

The immune-mediated inflammatory disease (IMID) use case was motivated by an interdisciplinary team that was interested in learning more about immunomodulatory drugs that are used to treat IMIDs, including systemic sclerosis, which is a spectrum of rare diseases involving excess



**FIGURE 3** Screenshots demonstrating an example of Translator evidence and provenance in support of naltrexone hydrochloride as an answer to the query in Figure 2.

**FIGURE 4** Schematic of three generalizable Translator workflows applied to support specific use-case queries on (a) immune-mediated inflammatory diseases, (b) Crohn's–Parkinson's disease relationship, and (c) drug-induced liver injury. (Graphic prepared by Kelsey Urgo).



collagen that can lead to fibrosis of the skin and/or internal organs. The team was interested in many classes of drugs, including Janus kinase inhibitors (JAK-Is), which have been suggested in the literature as a potential treatment for systemic sclerosis. The team thus approached the Translator Consortium with the following question: *what real-world evidence is there for the use of JAK-Is in patients with systemic sclerosis?*

Structured EHR data do not track the condition for which a medication is prescribed to a given patient. An investigator can examine co-occurrence rates between diagnoses and medications, but those rates can be deceptive due to the prevalence of commonly prescribed drugs such as acetaminophen among the general population. Translator clinical KPs have overcome this limitation of EHR data by allowing users to openly explore both co-occurrence rates and relative frequencies of medications, as well as information on whether a medication is contemporaneously predictive for a given disease or phenotype, thus provisioning informative EHR data and assertions without regulatory hurdles.

In this case, the Translator Consortium approached the user's question by executing a one-hop query that targeted Translator clinical KPs (Figure 4a). They first queried on a set of multiple IMIDs simultaneously. Translator answer sets comprised between 360 and 905 specific answers each and included drugs commonly used to treat IMIDs such as methotrexate and dexamethasone. A subsequent query focused specifically on the IMID systemic sclerosis. For this more restrictive query, Translator answer sets comprised between 128 and 366 specific answers each, including expected results such as mycophenolate, cyclophosphamide, and rituximab. Real-world evidence also was returned in the answer sets. For example, the observed-expected frequency ratio for co-occurrence of mycophenolate and systemic sclerosis was 3.91 (99% confidence interval: 3.67–4.11). When examining JAK-Is, Translator found evidence of co-occurrence in patients with systemic sclerosis, although the results were not among the top-ranked

answers. However, Translator reported that the JAK-I tofacitinib was predictive of systemic sclerosis in a real-world logistic regression model, indicating that JAK-Is have been prescribed to certain patients with systemic sclerosis. In addition, Translator provided PubMed abstracts suggesting mechanisms by which JAK-Is might treat systemic sclerosis, including evidence from mouse models and case studies. One example publication was titled: "Generation of a novel CD30+ B cell subset producing GM-CSF and its possible link to the pathogenesis of systemic sclerosis."<sup>23</sup>

The investigative team is now using Translator to further explore mechanistic evidence connecting JAK-Is and IMID disease processes.

### Explain: Crohn's disease and Parkinson's disease

This use case was motivated by clinical observations that patients with Crohn's disease are at higher risk of Parkinson's disease—two apparently unrelated diseases. Specifically, the investigative team approached the Translator Consortium with the following question: *why do patients with Crohn's disease have a higher risk of developing Parkinson's disease?*

The Translator Consortium addressed this question by constructing a two-hop query that sought biomedical entities that might be shared by both Crohn's disease and Parkinson's disease (Figure 4b). The query was structured with two specified *biolink:Disease* nodes, each connected to an unspecified *biolink:NamedThing* node (i.e., a root class for all things and informational relationships).

Due to the open structure of the query, Translator returned a variety of biomedical entities, including genes, diseases, chemicals, and drugs. Five genes were found to be associated with both Crohn's disease and Parkinson's disease, supporting the initial observation and suggesting at least partial common susceptibility pathways between these diseases. The identified genes were: *LRRK2* (leucine rich repeat kinase 2); *PARK7* (Parkinsonism associated deglycase);

*NOD2* (nucleotide binding oligomerization domain containing 2); *GPR65* (G protein-coupled receptor 65); and *MUC19* (mucin 19). Moreover, Translator provided quantitative publication support for each gene's involvement in both Crohn's disease and Parkinson's disease. In addition to genes, Translator found that the antibiotic rifaximin was associated with both diseases. Whereas the association between rifaximin and Crohn's disease was not surprising to the investigative team, given that antibiotics are used to control bacterial overgrowth in patients with Crohn's disease,<sup>24</sup> the association between rifaximin and Parkinson's disease was surprising. In fact, Translator provided publication support showing that rifaximin reduced motor fluctuations in a small clinical trial on Parkinson's disease, with a publication titled: "Small intestinal bacterial overgrowth in Parkinson's disease: tribulations of a trial."<sup>25</sup>

The investigative team had expected *LRRK2* to be among the answers returned to the query, so the fact that this gene indeed was returned by Translator provided the team with confidence in the accuracy and sensitivity of Translator answers and convinced them that a convergence of evidence, even if modest, such as the evidence exposed in this use case, can guide the emergence of unknown or unconventional KG paths and thereby assist with the identification of new treatment approaches to disease. The investigative team now plans to take a deeper dive into the supporting evidence and generate new queries to determine if there are common biological processes that might explain how these shared genes contribute to two diseases that were not previously thought to be related. The team also plans to search for additional data sources to incorporate into Translator, including specialized data sources on gene expression, functional genomics, and pharmacogenomics.

## Repurpose: Drug-induced liver injury

The DILI use case was motivated by shared interests between the Translator Consortium and the DILI Network. A high priority for the DILI Network, which is the longest running cohort-based study funded by the National Institutes of Health, is to initiate a DILI clinical trial. This priority is motivated by the fact that the only consensus treatment for DILI is to discontinue the causal agent, leaving patients with few therapeutic options until the drug injury resolves, and leaving underlying diseases and conditions untreated. DILI Network investigators have been unable to identify a suitable therapeutic, namely, one that is generally safe, with sufficient biological justification to support a clinical trial.

Hence, one of the investigators of the DILI Network approached the Translator Consortium with this goal in

mind. The specific question that was asked was *what drug candidate(s) might be repurposed for the treatment of DILI, and is there sufficient biological plausibility to justify the use of that candidate(s) in a clinical trial?*

The Translator Consortium approached this question with a two-fold solution (Figure 4c): (1) implement a complex asynchronous three-hop query to identify candidate drugs, leveraging the knowledge provided by Translator clinical KPs; and then (2) implement a simple one-hop query to find additional support for any candidate drugs thus identified, leveraging the real-world and curated knowledge provided by all KPs.

Translator successfully executed both queries and identified two candidate drugs, both antioxidants that are available over-the-counter and in prescription formulation: resveratrol and quercetin. Translator provided additional evidence to justify the use of these candidates in a clinical trial, including: the identification of intermediary genes that suggest biological plausibility; evidence of effectiveness in rodent models of DILI; and clinical trial precedence in other diseases and conditions such as chronic obstructive pulmonary disease. Moreover, Translator provided real-world evidence that these drugs are prescribed to patients.

To exemplify the knowledge and data that Translator reasons over, we highlight the answers and additional evidence that Translator provided in support of quercetin. Specifically, for the initial three-hop query, Translator provided real-world evidence that DILI is associated with a variety of other diseases, including autoimmune hepatitis, psoriasis, and osteoarthritis. One answer subgraph indicated that toxic liver disease (equivalent to DILI) co-occurs with infectious bacterial disease with sepsis in patients, with an observed-expected frequency ratio of 4.48 (99% confidence interval: 3.63–5.00). Tumor necrosis factor (TNF), a proinflammatory cytokine, was identified as the gene in the path between infectious bacterial disease with sepsis and quercetin, with Translator indicating that the evidence was derived from a resource called SemMedDB. Translator provided more than two dozen publications, including PubMed abstracts, supporting a relationship between TNF and quercetin, with most publications derived from primary rodent studies. The first publication was titled: "Quercetin inhibits LPS-induced nitric oxide and tumor necrosis factor-alpha production in murine macrophages"; and the abstract suggests that quercetin inhibits TNF.<sup>26</sup> The second one-hop query then asked for additional evidence related to quercetin. Translator provided evidence that quercetin is effective in the treatment of DILI, drug-induced dyskinesia, and drug-related side effects and adverse reactions in rodent models. The first publication<sup>27</sup> in one answer subgraph was titled: "Involvement of P450s and nuclear receptors in

the hepatoprotective effect of quercetin on liver injury by bacterial lipopolysaccharide"; and the abstract contained the sentence: "In this study, we used liposomal nanoparticles to entrap quercetin and evaluated its protective and therapeutic effects on drug-induced liver injury in rats." In addition, Translator provided real-world evidence that quercetin was prescribed to patients with a variety of diseases, including allergic rhinitis, with an observed-expected frequency ratio of 2.24 (99% confidence interval: 1.30–2.79). Moreover, Translator provided evidence that quercetin is in clinical trials as a treatment for chronic obstructive pulmonary disorder,<sup>28</sup> thus establishing precedence for a clinical trial on DILI.

Having met the criteria for viable drug candidates in clinical trials of DILI, members of the Translator Consortium now plan to prepare a formal report on Translator's findings for consideration by the DILI Network Steering Committee.

## DISCUSSION

The Translator program is in its third year of development, having first demonstrated feasibility. (See Figure S1 for complete timeline and notable milestones.) Several key advancements have been achieved since we first described the Translator system in 2019.<sup>3,4</sup> For example, at the time of our first report, a unified Translator "system" functionally did not exist; rather, Translator was comprised of many individual tools and services that were not truly integrated or harmonized. This is in contrast to the prototype Translator system that now exists, which functions as a truly unified system. This achievement is due, in part, to the consortium-wide adoption of ontologies and standards, such as Biolink Model and TRAPI, as well as tools to support their adoption and continued use. These ontologies and standards allow for the seamless integration and harmonization across completely disparate "knowledge sources," including observational clinical datasets and curated biomedical datasets. The Translator program has also moved beyond its initial two use cases on Fanconi anemia and asthma to include the use cases described here on IMIDs, Crohn's disease/Parkinson's disease, and DILI, as well as others. Moreover, the Translator program now has a nontechnical component, the SRI, that aims to create and maintain the collaborative framework required to support the adoption and implementation of standards and references, including services to support technical Translator components and teams. Through these standards and services, Translator has been able to readily expand the number of knowledge sources from which it draws data and knowledge and the number of use cases that it is able to support. A final achievement worth

mentioning is that the Translator Consortium has maintained a unique culture of open collegial collaboration and communication, despite the addition of new teams and the inevitable turnover of team members.

Whereas a prototype Translator system now exists, with demonstration of its success in returning valid answers to user questions, there are several areas of improvement required to truly achieve a production-level Translator system.

First, the scoring and ranking algorithms that are invoked by the ARAs are intentionally varied to provide breadth in answer sets and associated evidence. We acknowledge a need to refine the scoring and ranking algorithms in order to prioritize those answers with strong evidence, more complete provenance, and high confidence, thereby enriching for answers that are likely to provide the greatest insights to users.

Second, the TRAPI standard and Biolink Model are critical to standardize queries and answers across the federated Translator system. However, standardization can result in a lack of granularity and an inability to pose nuanced queries. For instance, workflow operations are only minimally supported in the current TRAPI standard. We are working to provision a variety of logical operations such as a graph overlay operation. We are also extending the Biolink Model to support nuanced statements by developing a core set of qualifiers (e.g., disease severity) that can be used to capture semantic richness.

Third, the clinical insights provided by the Translator system should be interpreted with caution. For instance, in our IMID use case, we provided real-world EHR evidence that JAK-Is co-occurred with systemic sclerosis and were predictive of systemic sclerosis in a logistic regression model, thus supporting the assertion that they are prescribed to patients with systemic sclerosis. However, we did not provide evidence of clinical benefit when prescribed to treat systemic sclerosis. Translator clinical KPs rely primarily on structured EHR data. Structured EHR data can be used to derive information on clinical benefit, for example, by examining the frequency of emergency department visits for condition X among patients with a diagnosis of disease Y who were prescribed medication Z compared to those who were not prescribed the same medication. However, TRAPI currently does not support such nuanced queries, although efforts are underway to adapt TRAPI to allow for more sophisticated queries. For certain use cases (e.g., DILI), Translator clinical KPs expose study data, which do support TRAPI-compliant assertions regarding clinical outcomes, but such data are not available for all use cases. At present, the approach that we are taking is to use curated knowledge sources to explore mechanistic evidence for how JAK-Is might reduce inflammation in systemic sclerosis.



Finally, whereas several Translator teams have developed user interfaces (UIs) that support TRAPI queries and answers, a uniform cross-component UI is not yet available, although NCATS recently funded a team to develop one (see timeline in Figure S1). We recognize the urgent need for such an interface, which will allow us to more effectively engage users, thus serving a broader community and promoting long-term sustainability. We note that a mock-up Translator UI has been developed and is now being vetted by users, with an early-phase prototype UI expected to be deployed by the end of calendar year 2022.

We note that the Translator system is one of several available biomedical KG-based question-answering systems. Others include Causaly,<sup>29</sup> Elsevier's Biology Knowledge Graph<sup>30</sup> and related Pathway Studio,<sup>31</sup> and Google's Knowledge Graph.<sup>10</sup> We emphasize a few differences among these systems. First, the Translator system is the only open-source, community-contributed system; Causaly and Elsevier's systems are commercial, and Google's Knowledge Graph is largely proprietary. For the IMID and DILI use cases reported here, the open-source nature of Translator allowed us to run queries that openly explored EHR evidence on co-occurrence rates of observations, relative frequencies, and disease risk predictions, without regulatory hurdles. Second, these systems are narrower in scope than Translator. Elsevier's systems are highly specific to basic biology and do not span the translational spectrum. Causaly's system supports a broader set of translational questions, but only a subset of those supported by Translator. Thus, our use cases included queries that spanned multiple biomedical entities (e.g., genes, chemical entities, small molecules, drugs, phenotypes, diseases) and numerous knowledge sources, including clinical knowledge sources. Third, Translator supports a more sophisticated set of queries than the other systems. For instance, Google's Knowledge Graph only supports simple "lookup" operations, albeit with highly sophisticated natural language parsing of user questions. Causaly's system is currently limited to linear two-hop queries. Neither Causaly's nor Elsevier's systems support batch or asynchronous queries, in contrast to the Translator system. Our DILI use case leveraged Translator's advanced capabilities, including three-hop, batch, and asynchronous queries. Finally, none of the other systems support clinical knowledge, such as EHR data, which provided key support for two of the three use cases reported herein.

In conclusion, we have developed a biomedical KG-based Translator system capable of integrating a wide range of data sets and translating those data into insights intended to augment human reasoning and accelerate translational science. We are now working on refinements to the prototype Translator system.

## ACKNOWLEDGEMENTS

The authors are grateful to members of the Publications Committees at the National Center for Advancing Translational Sciences (NCATS), the National Institute of Environmental Health Sciences, and the National Institute on Aging, as well as Dr. Naga P. Chalasani, for their review and approval of the manuscript for publication. They further thank Kelsey Urgo for assistance with graphics design, and Stanley C. Ahalt of the Renaissance Computing Institute for financially supporting the graphics design work. Moreover, the authors are appreciative of the unwavering leadership and support provided by the Translator Extramural Leadership Team and the Intramural Research Program at NCATS.

## CONFLICT OF INTEREST

S.E.B. and S.H. have received support from the NSF Convergence Accelerator Open Knowledge Networks to develop applications related to the SPOKE KG. All other authors declared no competing interests for this work.

## ORCID

Karamarie Fecho  <https://orcid.org/0000-0002-6704-9306>

Anne E. Thessen  <https://orcid.org/0000-0002-2908-3327>

Jennifer J. Hadlock  <https://orcid.org/0000-0001-6103-7606>

Mark D. Williams  <https://orcid.org/0000-0001-8020-916X>

William Byrd  <https://orcid.org/0000-0003-4730-5293>

Vlado Dančik  <https://orcid.org/0000-0002-5970-6660>

Gustavo Glusman  <https://orcid.org/0000-0001-8060-5955>

Nomi L. Harris  <https://orcid.org/0000-0001-6315-3707>

Adam Johs  <https://orcid.org/0000-0001-7570-4326>

Andrew I. Su  <https://orcid.org/0000-0002-9859-4104>

## REFERENCES

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18
2. Austin CP, Colvis CM, Southall NT. Deconstructing the translational tower of babel. *Clin Transl Sci*. 2019;12(2):85. doi:10.1111/cts.12595
3. Biomedical Data Translator Consortium. Toward a universal biomedical data translator. *Clin Transl Sci*. 2019;12(2):86-90. doi:10.1111/cts.12591
4. Biomedical Data Translator Consortium. The biomedical data translator program: conception, culture, and community. *Clin Transl Sci*. 2019;12(2):91-94. doi:10.1111/cts.12592
5. Ackoff RL. From data to wisdom. *J Appl Syst Anal*. 1989;16(1):3-9. <https://softwarezen.me/wp-content/uploads/2018/01/datawisdom.pdf>
6. Alavi M, Leidner DE. Review: knowledge management and knowledge management systems: conceptual foundations and research issues. *Miss Q*. 2001;25(1):107-136. doi:10.2307/3250961
7. Rowley J. The wisdom hierarchy: representations of the DIKW hierarchy. *J Inf Sci Eng*. 2007;33(2):163-180. doi:10.1177/0165551506070706

8. Schmitt CP, Cox S, Fecho K, et al. *Scientific Discovery in the Era of Big Data: More than the Scientific Method*. Vol 3. RENCI; 2015. Accessed December 13, 2021. <https://renci.org/wp-content/uploads/2015/11/Sci-Discovery-BigData-FINAL-11.23.15.pdf>.
9. Wilcke X, Bloem P, de Boer V. The knowledge graph as the default data model for learning on heterogeneous knowledge. *Data Sci*. 2017;1(1–2):39–57. doi:10.3233/ds-170007
10. Singhal A. *Introducing the Knowledge Graph: things, not strings*. Google; 2012. Published May 16, 2012. Accessed January 18, 2022. <https://blog.google/products/search/introducing-knowledge-graph-things-not/>.
11. Huang Z, Zheng Y, Cheng R, Sun Y, Mamoulis N, Li X. Meta structure: computing relevance in large heterogeneous information networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:1595–1604. doi:10.1145/2939672.2939815
12. Davis AP, Grondin CJ, Johnson RJ, et al. The comparative Toxicogenomics database: update 2019. *Nucleic Acids Res*. 2019;47(D1):D948–D954. doi:10.1093/nar/gky868
13. Wg OT. *Mondo Disease Ontology*. Accessed December 13, 2021. <http://www.obofoundry.org/ontology/mondo.html>.
14. Ahalt SC, Chute CG, Fecho K, et al. Clinical data: sources and types, regulatory constraints, applications. *Clin Transl Sci*. 2019;12(4):329–333. doi:10.1111/cts.12638
15. Github. *ReasonerAPI: NCATS Biomedical Translator Reasoners Standard API*. Accessed December 13, 2021. <https://github.com/NCATSTranslator/ReasonerAPI>
16. SmartAPI. SmartAPI. Accessed December 13, 2021. <https://smart-api.info/registry?tags=translator>.
17. Unni DR, Moxon SAT, Bada M, et al. and The Biomedical Data Translator Consortium. Biolink model: a universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin Transl Sci*. 2022;15:1848–1855. doi:10.1111/cts.13302
18. Tree Viz of Biolink. Accessed February 4, 2022. <http://tree-viz-biolink.herokuapp.com/>
19. Github. *Biolink-Model: Schema and Generated Objects for Biolink Data Model and Upper Ontology*. Accessed February 4, 2022. <https://github.com/biolink/biolink-model>.
20. Biolink model. *Biolink Model*. Accessed January 18, 2022. <https://biolink.github.io/biolink-model/>.
21. Bell J, Strang J. Medication treatment of opioid use disorder. *Biol Psychiatry*. 2020;87(1):82–88. doi:10.1016/j.biopsych.2019.06.020
22. Derry CJ, Derry S, Moore RA. Caffeine as an analgesic adjuvant for acute pain in adults. *Cochrane Database Syst Rev*. 2012;3:CD009281. doi:10.1002/14651858.CD009281.pub3
23. Higashioka K, Kikushige Y, Ayano M, et al. Generation of a novel CD30+ B cell subset producing GM-CSF and its possible link to the pathogenesis of systemic sclerosis. *Clin Exp Immunol*. 2020;201(3):233–243.
24. Castiglione F, Rispo A, Di Girolamo E, et al. Antibiotic treatment of small bowel bacterial overgrowth in patients with Crohn's disease. *Aliment Pharmacol Ther*. 2003;18(11–12):1107–1112. doi:10.1046/j.1365-2036.2003.01800.x
25. Vizcarra JA, Wilson-Perez HE, Fasano A, Espay AJ. Small intestinal bacterial overgrowth in Parkinson's disease: tribulations of a trial. *Parkinsonism Relat Disord*. 2018;54:110–112. doi:10.1016/j.parkreldis.2018.04.003
26. Manjeet KR, Ghosh B. Quercetin inhibits LPS-induced nitric oxide and tumor necrosis factor-alpha production in murine macrophages. *Int J Immunopharmacol*. 1999;21(7):435–443. doi:10.1016/s0192-0561(99)00024-7
27. Zhao L, Chen F, Zhang Y, et al. Involvement of P450s and nuclear receptors in the hepatoprotective effect of quercetin on liver injury by bacterial lipopolysaccharide. *Immunopharmacol Immunotoxicol*. 2020;42(3):211–220. doi:10.1080/08923973.2020.174215428
28. Translator evidence for DILI use case. *Search of: "NCT01708278"OR"NCT03989271" – list results – ClinicalTrials.gov*. Accessed April 27, 2022. <https://clinicaltrials.gov/search?id=%22NCT01708278%22OR%22NCT03989271%22>
29. Causaly – Accelerate your research. Accessed December 13, 2021. <https://www.causaly.com/>
30. Elsevier. *Biology Knowledge Graph*. Accessed December 13, 2021. <https://www.elsevier.com/solutions/biology-knowledge-graph>
31. Elsevier. *Pathway Studio*. Accessed December 13, 2021. <https://www.elsevier.com/solutions/pathway-studio-biological-research>

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Fecho K, Thessen AE, Baranzini SE, et al. Progress toward a universal biomedical data translator. *Clin Transl Sci*. 2022;15:1838–1847. doi:10.1111/cts.13301