Citation: Clin Transl Sci (2019) 12, 86–90; doi:10.1111/cts.12591

COMMENTARY

Toward A Universal Biomedical Data Translator

The Biomedical Data Translator Consortium

Myriad biomedical data are available today and bring with them a need to "translate" those data into meaningful information that can be used to accelerate clinical and translational science and drive innovations in clinical care and drug discovery. We describe the work of The Biomedical Data Translator Consortium to develop a platform that supports the translation of data into knowledge by applying inferential methods to a graph representation of biomedical entities and their relationships.

FROM VISION TO TRANSLATION

The collection and analysis of "big data" are delivering significant benefits across multiple sectors of society, including social media, finance, marketing and advertising, fraud prevention, and the physical sciences.[1–3] Although many of these efforts have yet to demonstrate improvements to public welfare, the promise of big data for the public good is widely anticipated.

This promise is perhaps most broadly felt in translational research and clinical care. However, although the amount and diversity of biomedical data have dramatically increased in recent years, tangible benefits derived from big data to the practice of medicine have been slower to arrive. For example, the Precision Medicine Initiative has made remarkable progress, with several examples of success,[4] but universal approaches for tailoring treatments to individual patients have yet to be developed. In addition, randomized controlled clinical trials, once the "gold standard" of clinical research and practice, are now widely recognized as limited due to their inherent inability to target investigational treatments and therapies to patients who are most likely to demonstrate maximum benefit (and minimum harm) based on individual biomolecular profiles. Indeed, this shortcoming underlies the "basket trial" design that is gaining popularity, and demonstrating early success, in oncology research.[5]

The National Center for Advancing Translational Sciences (NCATS), a center within the National Institutes of Health, launched the Biomedical Data Translator program in October 2016 in an effort to overcome existing challenges and provide a complementary solution (**Table S1**). The goal of the initial feasibility assessment is to design and prototype a "Translator" system capable of integrating existing biomedical data sets or "Knowledge Sources" and "translating" those data into insights that can accelerate translational research, support clinical care, and leverage clinical expertise to drive research innovations.

A HIGH-LEVEL OVERVIEW OF UNIVERSAL TRANSLATION

As envisioned, the Translator system will facilitate the use of disparate biomedical data to solve a range of challenges (**Figure 1**). Among other goals, the Translator aims to enable what has been termed "data-driven clinical regrouping" of patients in an effort to refine current definitions of disease and identify groups of patients who are likely to respond or not respond to specific treatments. These groupings are expected to facilitate the identification of patients with shared molecular and cellular biomarkers and thereby generate new biomedical hypotheses, promote translational discovery, facilitate efficient clinical-trial design, and inform clinical decision making. In this manner, the Translator system is intended to supplement, not replace, traditional approaches to clinical and translational science.

To achieve these ambitious goals, the Translator will identify and target multiple Knowledge Sources deemed relevant to a user inquiry and apply a variety of computational approaches to translate data from those Knowledge Sources into new knowledge, with associated provenance and a level of confidence. The user will then be empowered with new information that can be used directly to answer the inquiry or to generate testable hypotheses. The computational approaches will yield results or "artifacts" that may have obvious clinical or biological relevance, as well as output that may not seem to be relevant, as determined by subject-matter experts. The artifacts of questionable relevance will be maintained by the Translator not only to improve performance of the system over time, but also to facilitate serendipity: unexpected discoveries that are revealed as a result of combining new Knowledge Sources or computational approaches, generating new inquiries, or revealing heretofore unexpected relationships between biological mechanisms. Through this high-level iterative approach, the Translator system is expected to demonstrate improvement over time and ultimately achieve the important goal of disease-agnostic data-driven clinical regrouping—hence, "universality."

DRIVING DEMONSTRATION USE CASES

Demonstration use cases and targeted inquiries have driven research and development of the prototype Translator system. Specifically, two demonstration use cases have been developed thus far: pediatric asthma, a common disease influenced by a plethora of genes and socio-environmental

*Correspondence: Stanley C. Ahalt (ahalt@renci.org)
Received: July 30, 2018; accepted: August 30, 2018. doi:10.1111/cts.12591

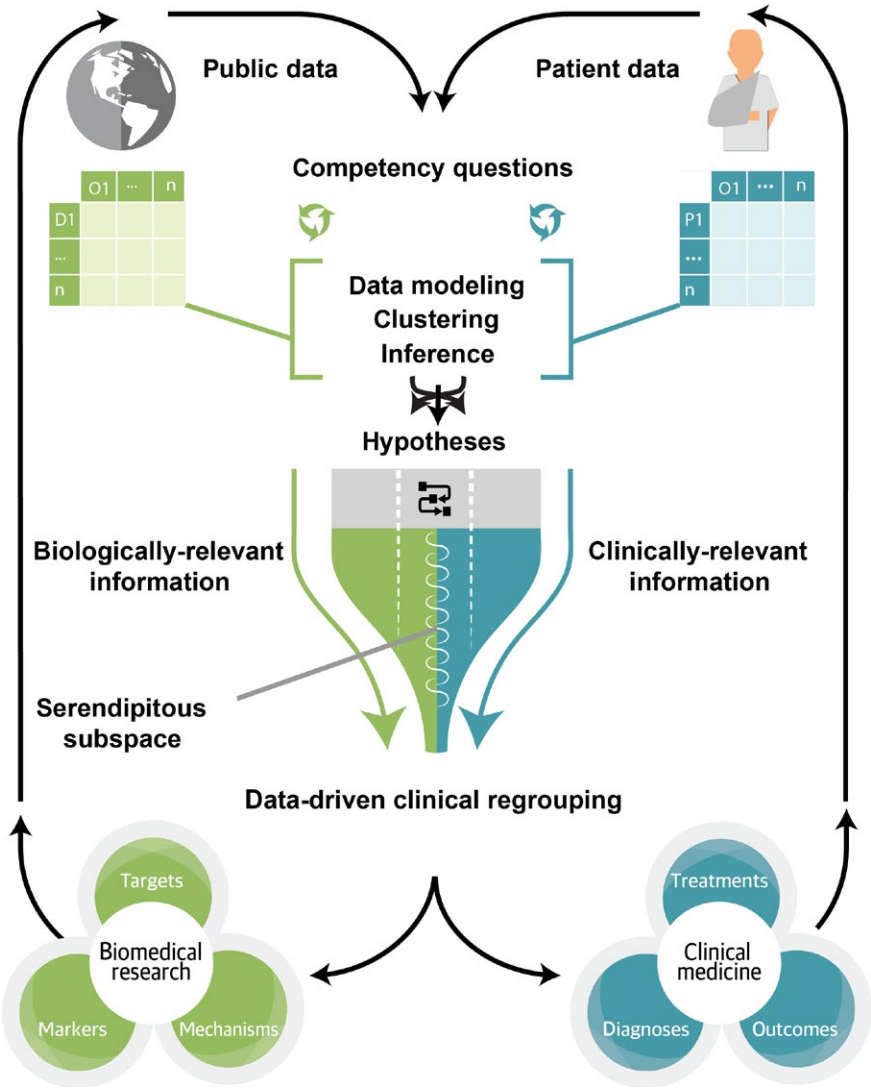

Figure 1 Conceptual overview of Biomedical Data Translator functionality.

exposures; and Fanconi anemia, a rare disease caused by defects in relatively few genes and influenced by select socio-environmental exposures. These two use cases will serve to ensure that the Translator is developed with the flexibility needed to address inquiries of relevance to each extreme of the disease-frequency spectrum. The specific questions that are being asked include questions whose answers are largely known, but are necessary to test the nascent Translator system, such as: Is exposure to airborne particulate matter associated with asthma exacerbations and responsiveness to treatment? Do genes and gene products causal to Fanconi anemia interact with genes and gene products related to aldehyde metabolism? Other questions are more exploratory and are designed to push the boundaries of the prototype system, such as: Which medications are currently prescribed to pediatric patients with an asthma-like phenotype who are responsive to treatment despite high levels of exposure to airborne particulate matter? What drug targets block the biological pathways that are activated by particulate matter? What genes show high phenotypic similarity to the 11 Fanconi anemia core complex genes? What genes and pathways are uniquely targeted by hematopoietic-stem-cell-transplantation-conditioning drugs that are well tolerated vs. poorly tolerated by patients with Fanconi anemia?

Importantly, the use-case questions were selected specifically because answering them requires a variety of data sources and a series of translational tasks, each involving varying degrees of reasoning. Translator team members acknowledge the critical need to identify any "gaps" in data sources and systematically evaluate any answers generated by the prototype Translator system. To this end, select questions will serve as a validation set for the prototype system. These questions will be complex and require a variety of data sources and reasoning-based translational tasks, but they will have "ground-truth" answers that can be used to evaluate output from the prototype system. Moreover, some of the questions will not be related to the use cases,

which will help to prevent overfitting of the prototype system. Furthermore, team members are developing sophisticated analytic approaches to rank and score all answers and quantify uncertainty and confidence in results. Finally, team members are engaging subject-matter experts to provide feedback on the ranking and scoring of answers in order to iteratively improve the prototype system.

KNOWLEDGE SOURCES

To translate these inquiries, the prototype Translator is designed to tap into a rich set of "Knowledge Sources," including (but not limited to): data on clinical signs and symptoms, diagnostic measures and tests, diagnoses and other disease constructs, treatments and drug exposures, and exome or genome pedigrees, available from patient medical records and clinical ontologies; data on environmental exposures (e.g., airborne pollutants and ozone), available from the US Environmental Protection Agency; data on socio-environmental exposures (e.g., household income, access to transportation, and access to health insurance), available from the US Census Bureau; and experimental and mechanistic data on genetic variants, biological pathways, molecular and cellular anatomy, systems biology, and chemical structures and drug targets, available from a variety of open data sources. To date, over 40 such Knowledge Sources have been developed through a collaboration of Translator institutions across the nation. Certain Knowledge Sources have proven to be more challenging to access and integrate than others (e.g., patient data), but Translator team members have worked creatively to overcome the challenges. As a striking example, we recently identified an approach for revealing processed clinical data in a regulatory-compliant manner that enables the open use of clinical data as part of the Translator program (see COHD and ICEES, **Figure 2**). Teams are also developing privacy-preserving analytic methods for summarizing patient records.[6] Successes and insights such as these were only made possible through the collaborative, interdisciplinary culture that the program has fostered and the unique skillsets and perspectives provided by individual team members.

KNOWLEDGE GRAPHS AND REASONERS

A prototype Translator framework and architecture have been developed for integrating semantically annotated Knowledge Sources and creating a data platform to support automated reasoning and serendipitous discovery of new "facts" or interesting and testable hypotheses (**Figure 2**). First, the application programming interfaces (APIs) of Knowledge Sources are catalogued directly with associated metadata in a unified "Translator API Registry" as an extension of the SmartAPI service standard.[7] In addition, Knowledge Sources are wrapped generically as "Knowledge Beacons" using a standardized, semantically typed workflow for discovery of concepts and relationships. The BioLink data model serves to standardize the semantic structure of the data contained within the Knowledge Sources and thereby facilitate integration and automated reasoning. To achieve reasoning, a virtual "Knowledge Graph" integrates the data by connecting "nodes" or entity types with "edges" or predicates in the graph. The Knowledge Graph architecture is similar to the "Blackboard" architecture that was conceptualized more than 4 decades ago by researchers in artificial intelligence research as a general problem-solving approach to overcome the challenges in multimodal problem solving.[8] The

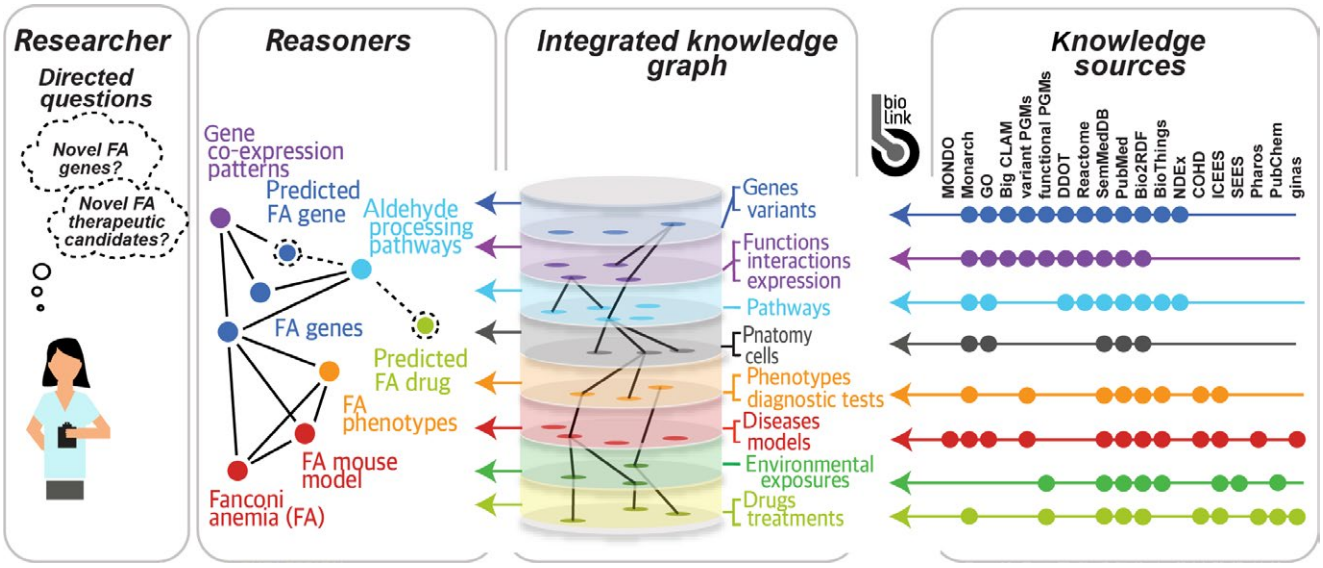

Figure 2 High-level overview of prototype Translator system architecture. Abbreviations for Translator APIs: Big CLAM, Big Cell Line Association Miner; BioThings, BioThings Explorer; COHD, Columbia Open Health Data; DDOT, Data-Driven Ontology Toolkit plus disease ontologies; FA, Fanconi anemia; ginas, global ingredient archival system; ICEES, Integrated Clinical and Environmental Exposures Service; GO, Gene Ontology; MONDO, MONarch Disease Ontology; NDEx, Network Data Exchange; PGMs, probabilistic graphical models; SEES, Socio-Environmental Exposures Service; SemMedDB, Semantic MEDLINE DataBase. For additional information, including hyperlinks to Translator services, see **Table S2**.

central idea is that multiple, distributed Knowledge Sources dynamically propose pieces of a translated user inquiry and posit that information to a virtual Knowledge Graph. One or more "Reasoners" then work collaboratively or independently to apply analytic approaches to infer a complete or partial translation of the inquiry by generating inference "paths" across the nodes and edges contributed by the various Knowledge Sources. The Knowledge Sources themselves do not communicate with each other; rather, the Knowledge Graph is the only "shared knowledge." The Reasoners then return ranked and scored potential translations with provenance and supporting evidence. The user is then able to evaluate the translations and supporting evidence and provide feedback to the Reasoners, thus promoting continuous improvement of the prototype system.

MOTIVATING FACTORS

The Translator program currently is in the feasibility phase. The program's early success will help to ensure the long-term viability of the project, although we recognize that sustaining cyberinfrastructure for science represents a significant challenge across funding agencies and scientific fields. However, the Translator effort is motivated by a number of observations and collective assertions of team members that we believe will encourage continued research and development of the prototype Translator system. First, we assert that a single monolithic data set that directly connects the complete set of clinical characteristics to the complete set of biomolecular features, including "-omics" data, will never exist because the number of characteristics and features is constantly shifting and exponentially growing. Second, even if such a single monolithic data set existed, all-vs.-all associations will inevitably succumb to problems with statistical power (i.e., the curse of dimensionality).[9] Such problems will get worse, not better, as more and more clinical and biomolecular data are collected and become available. We also assert that there is no single language, software or natural, with which to express clinical and biomolecular observations—these observations are necessarily and appropriately linked to the measurement technologies that produce them, as well as the nuances of language. The lack of a universal language for expressing clinical and biomolecular observations presents a risk of isolation or marginalization of data that are relevant for answering a particular inquiry, but are never accessed because of a failure in translation.

Based on these observations, our final assertion is that automating the ability to reason across integrated data sources and providing users who pose inquiries with a dossier of translated answers coupled with full provenance and confidence in the results is critical if we wish to accelerate clinical and translational insights, drive new discoveries, facilitate serendipity, improve clinical-trial design, and ultimately improve clinical care. This final assertion represents the driving motivation for the Translator system.

MOVING FORWARD

Why is the time right to build a Translator system? First, we have access to vast amounts of valuable clinical and biomolecular data, yet managing the data, integrating the data, analyzing the data, and extracting value from the data have proven to be both challenging and expensive. Second, the cultural and political climate is such that the public is demanding that researchers, clinicians, and funders deliver on their promise to use all data sources that have been generated at considerable public cost for the public good. A final, critical reason to move forward with the Translator system is that only recently have we had the ***data***, the ***standards***, the ***tools***, the ***networks***, the ***computing power***, and the ***expertise*** to even conceptualize a Translator, let alone research and develop one.

Supporting Information. Supplementary information accompanies this paper on the *Clinical and Translational Science* website (www.cts-journal.com).

Table S1. The Biomedical Data Translator Consortium: teams and team members.

Table S2. Translator architecture: select Knowledge Sources.

Acknowledgments. Karamarie Fecho prepared the first draft of the manuscript and served as lead author. Stanley C. Ahalt served as senior author. Russ B. Altman, Jim Balhoff, Chris Bizon, Richard Brukiewich, Noel Burtt, Christopher G. Chute, Paul A. Clemons, Steve Cox, Vlado Dancik, Michel Dumontier, Gustavo Glusman, Rajarshi Guha, Melissa Haendel, Trey Ideker, Christopher Mungall, David B. Peden, Andrew Su, Nicholas Tatonetti, Alexander Tropsha, Chunhua Weng, and Michael Yu provided critical feedback and editorial input. Julie McMurry created **Figures 1 and 2** and provided critical feedback. All members of The Biomedical Data Translator Consortium (**Table S1**) contributed to the vision of the manuscript and conducted the work described therein.

Funding. Support for the preparation of this manuscript was provided by the National Center for Advancing Translational Sciences, National Institutes of Health, through the Biomedical Data Translator program (awards 1OT3TR002019, 1OT3TR002020, 1OT3TR002025, 1OT3TR002026, 1OT3TR002027, 1OT2TR002514, 1OT2TR002515, 1OT2TR002517, 1OT2TR002520, 1OT2TR002584). Any opinions expressed in this document are those of The Biomedical Data Translator Consortium and do not necessarily reflect the views of the National Institutes of Health, individual Translator team members, or affiliated organizations and institutions.

Conflict of Interest. The authors declared no competing interests for this work.

1. World Economic Forum. Big Data, Big Impact: New Possibilities for International Development. VITAL WAVE CONSULTING™ on behalf of World Economic Forum <http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf> (2012). Accessed June 21, 2018.
2. Marr, B. 4 Ways Big Data Will Change Every Business. Forbes. <https://www.forbes.com/sites/bernardmarr/2015/09/08/4-ways-big-data-will-change-every-business/#54c156ec2729> (September 8, 2015). Accessed June 21, 2018.
3. Schmitt, C. *et al.* Scientific Discovery in the Era of Big Data: More than the Scientific Method. RENCI White Paper, Vol. 3, No. 6. RENCI, University of North Carolina at Chapel Hill <https://doi.org/10.7921/g0c82763> (2015). Accessed June 21, 2018.
4. Marcus, M.B. Precision Medicine Initiative aims to Revolutionize Health Care. CBS NEWS, CBS Interactive, Inc. <https://www.cbsnews.com/news/presidents-precision-medicine-initiative-is-propelling-forward/> (February 25, 2016). Accessed June 21, 2018.

5. Mandrekar, S.J., Dahlberg, S.E. & Simon, R. Improving clinical trial efficiency: thinking outside the box. *Am. Soc. Clin. Oncol. Educ. Book* e141–e147 <https://meetinglibrary.asco.org/record/104032/edbook> (2015). Accessed June 21, 2018.
6. Robinson, M. *et al.* Fast and simple comparison of semi-structured data, with emphasis on electronic health records. bioRxiv. Preprint available online <https://doi.org/10.1101/293183> (April 2, 2018).
7. Dastgheib, S. *et al.* The 16th International Semantic Web Conference, Vienna, Austria, Vol. **1931**, pp. 1–4 <http://hdl.handle.net/1854/LU-8540886> (October 21, 2017 – October 25, 2017). Accessed June 21, 2018.
8. Corkill, D.D. Blackboard systems. *AI Expert* **6**, 40–47 <https://pdfs.semanticscholar.org/bc95/eab1002f6d5497ffa3a204736233e4d92b80.pdf> (1991). Accessed June 21, 2018.
9. Donoho, D.L. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Aide-Memoire. Department of Statistics, Stanford University <https://pdfs.semanticscholar.org/63c6/8278418b69f60b4814fae8dd15b1b1854295.pdf>. Accessed June 21, 2018.