Applying knowledge-anchored hypothesis discovery methods to advance clinical and translational research: the OAMiner project

Philip R O Payne, 1 Rebecca D Jackson, 2 Thomas M Best, 3 Tara B Borlawsky, 1 Albert M Lai, 1 Stephen James, 1 Metin N Gurcan 1

ABSTRACT
The conduct of clinical and translational research regularly involves the use of a variety of heterogeneous and large-scale data resources. Scalable methods for the integrative analysis of such resources, particularly when attempting to leverage computable domain knowledge in order to generate actionable hypotheses in a high-throughput manner, remain an open area of research. In this report, we describe both a generalizable design pattern for such integrative knowledge-anchored hypothesis discovery operations and our experience in applying that design pattern in the experimental context of a set of driving research questions related to the publicly available Osteoarthritis Initiative data repository. We believe that this ‘test bed’ project and the lessons learned during its execution are both generalizable and representative of common clinical and translational research paradigms.

INTRODUCTION
Clinical and translational research programs regularly produce large amounts of heterogeneous data, information, and knowledge. For example, the NIH-funded Osteoarthritis Initiative (OAI) is a multi-center, longitudinal study that seeks to identify predictive clinical characteristics, environmental exposures, and biomarkers associated with the development and progression of knee osteoarthritis (OA). 1 2 A core activity of the OAI program is the collection of demographic, anthropometric, exposure, physical performance, biochemical, genetic, and imaging data from a cohort of over 4000 participants. These data are made publicly available as a scientific resource for the broad biomedical research community. However, due to a number of factors, including inconsistent data representation schemata and a paucity of informatics methods for hypothesis discovery and testing in such multi-dimensional data sets, the ability to reuse such resources is often extremely limited.3 One potential solution to these challenges is the use of knowledge-anchored reasoning methods to discover and explore hypotheses spanning multiple, heterogeneous variables of interest that can be mapped back to their originating data sets. In this report, we describe a body of work conducted as part of the NLM-funded OAMiner project, focusing on two primary goals: (1) the development of a generalizable design pattern for the application of integrative, knowledge-anchored hypothesis discovery methods to heterogeneous data sets; and (2) the application and evaluation of that design pattern, utilizing the OAI data repository as a test bed.

CASE DESCRIPTION
The work described in this report has been conducted within the experimental context of a collaborative effort spanning the Department of Biomedical Informatics and a team of OA investigators at The Ohio State University. This context was selected to demonstrate the applicability of integrative informatics methods for hypothesis discovery in large-scale and heterogeneous data—a scenario frequently encountered in the clinical and translational science domain.

METHODS OF IMPLEMENTATION
A primary objective of the OAMiner project was to develop a design pattern for knowledge-anchored hypothesis generation in large-scale and heterogeneous public data sets, drawn from current best practices and emerging methods.4–9 This design pattern is illustrated in figure 1. Of note relative to our experimental context, namely the ability to reason upon linkages between:

Phase 1: information needs identification
A series of semi-structured interviews and focus group discussions were conducted with a convenience sample of OA investigators at The Ohio State University (n=4) in order to identify recurring hypothesis-centric information needs encountered by those individuals. Three recurring needs were identified relative to our experimental context, namely the ability to reason upon linkages between:
Image derived markers and clinical measurements;
> Image derived markers and patient reported outcomes; and
> Image derived markers, clinical measurements, and functional status indicators.

Phase 2: feature extraction
Informed by the information needs articulated in phase 1, we pursued two parallel and complementary approaches to feature extraction, focusing on image-based and structured symbolic data, respectively.

Feature extraction from unstructured data: image-derived feature generation
In order to fully elucidate the progression of any disease including OA, it is critical to precisely define variations in structural phenotypes so that one can rigorously select specific physical features to explore both cross-sectionally and longitudinally. Computer-assisted image-derived feature extraction methods are promising in terms of addressing such information needs. Therefore, throughout this project, we emphasized a rigorous approach to understanding and describing OA in a discrete manner by analyzing several structural groups in the knee area (e.g., cartilage, muscle, or meniscus). Specifically, we targeted the meniscus, bones (femur, tibia, fibula), and quadriceps muscles (vastus intermedius, vastus lateralis, vastus medialis, and rectus femoris). These structures were automatically or semi-automatically detected and segmented, and their characteristics (e.g., volume, cross-sectional area, etc) were measured. These measurements constitute the first-order image-derived features that were used to characterize each of these structures and their components in quantifiable form. The natural variation between participants regardless of the incidence or progression of the disease was minimized through statistical normalization techniques. Algorithms for the extraction of first-order features also led to the generation of second-order image-derived features (e.g., statistical properties such as kurtosis of intensity values) both at a single time point as well as for longitudinal analysis of temporal change in the characteristics of the disease. An example of such second-order image-derived features is provided in figure 2. In total, the implementation of first- and second-order feature extraction algorithms relative to the imaging data present in the OA data repository allows for the creation of a set of structured and computable participant-associated image-based markers that can then be semantically annotated and integrated with other correlative data types.
Feature extraction from structured and semi-structured data: knowledge discovery in databases

Building upon the image-derived features described above, we then focused on the extraction of high priority phenotypic features from the OAI data set. The information resource utilized in this project phase was comprised of variables related to case report form questions extracted from the OAI data dictionary. A knowledge engineer with over 10 years of experience in the biomedical informatics domain abstracted the data dictionary entries and curated those concepts into a computationally tractable format. The case report form questions (or labels) and imaging markers generated during this and the preceding image-derived feature analysis methods were then annotated with SNOMED CT concepts using the MetaMap annotation engine and UMLS Terminology Services Metathesaurus Browser. Conceptual post-coordination of concepts was necessary in order to adequately capture the context of many of these variables. For example, the phenotypic variable containing the text ‘Knee pain: in bed’ can be represented using the SNOMED CT concepts for (Knee Pain:30989003) and (Lying in bed:17535004).

Phases 3–4: feature aggregation and knowledge synthesis

Hypotheses concerning potentially novel relationships between the aforementioned variable types were induced using a component-based biomedical knowledge synthesis platform known as TOKEn (Translational Ontology-anchored Knowledge Discovery Engine). This platform uses conceptual knowledge engineering techniques to support knowledge discovery in databases, and in particular, a method known as constructive induction. This approach leverages domain-specific knowledge found in both publically available ontologies as well as complementary knowledge extracted from literature using text mining and machine learning methods, in order to identify knowledge-anchored relationships of interest between sets of variables in a targeted data set.

In order to provide flexibility in terms of the use of TOKEn in our design pattern, components were developed to transform all knowledge sources to an OWL 2.0 representation. The OWL standard was chosen due to its widespread adoption and use in the knowledge engineering and semantic web communities. This approach was the basis for the implementation of a computational pipeline including the following steps (figure 3):

Figure 2 Output of meniscus segmentation (outlined), with corresponding histogram and Gaussian curve fitting. Higher order statistical measures derived from the histogram, such as skewness and kurtosis, are examples of second-order image-derived features.

Figure 3 System design overview of the OAMiner hypothesis generation pipeline. The knowledge source component pipeline exists to extract computable information from unstructured and semi-structured knowledge using natural language processing (NLP) techniques and subject matter expert (SME) involvement, the result of which is represented in the web ontology language (OWL). The UMLS based component pipeline serves as the basis for the constructive induction framework for the discovery of transitive relationships using the ontological information that exists in the UMLS.
(1) semantic annotation of heterogeneous data sets; (2) induction of relationships between identified concepts; (3) generation of OWL representations of such data; and (4) use of the TOKEn engine to induce transitive relationships between conceptual entities.

Phase 5: evaluation
A structured survey instrument was implemented using the REDCap platform, allowing subject matter experts (SMEs) to visualize graph-based visualizations of hypotheses generated in phase 4, and evaluate those hypotheses based upon three axes as introduced earlier, specifically: validity, usefulness, and novelty. A random subset of the hypotheses generated in phase 4 have been selected for this process, which is actively being evaluated in a user-centric and iterative manner at the time of submission of this manuscript. Preliminary results from this evaluation process have indicated that: (1) hypotheses generated spanning image-derived markers and clinical measurements or performance status indicators are regularly found to be valid, useful, and novel; (2) the complexity of such hypotheses (in terms of the number of concepts, relationships, and information resources involved in their generation and presentation) can affect the ability of SMEs to readily evaluate such constructs; and (3) the use of post-coordinated conceptual entities to comprise such hypotheses remains an open area of investigation, yielding variable results in terms of the validity and usefulness of resulting hypotheses. We intend to report upon the full spectrum of these evaluative results, which extend beyond the scope of this case report, in a subsequent manuscript.

DISCUSSION
Based upon our experiences in implementing and evaluating the aforementioned design pattern in a prototypical use case, we have identified a number of critical lessons that we believe are applicable to analogous projects, as summarized below.

- Each disease has its own detection, diagnosis, and treatment regimen and in many cases imaging is a critical component of these steps. However, scalable methods that allow such imaging data to be integrated with other heterogeneous data types have not been well developed. In order to apply the approaches we have described in this report to other diseases, it will be very important to understand: (1) the type of imaging modalities (eg, MRI, CT, histopathology, etc); and (2) imaging stage (eg, detection, radiation therapy, treatment monitoring) and the imaging-informed key decision factors (eg, detection of tumors, accurately quantifying the size and morphology of structures in longitudinal studies). Imaging should inform and extend the current knowledge with all its capabilities (accuracy, consistency, and measurement of phenomena of interest (eg, texture characteristics)). A combination of the consistent and objective measurement tools combined with newly developed ones, opens up the possibilities of imaging-based biomarker generation and validation.

- The lack of standards surrounding the annotation of public research data sets are a barrier to using existing knowledge sources consistent with the design patterns described in this report. Without conceptual metadata, usage of publically available data is limited to traditional syntactic analysis. Annotation of the data is required in order to process and infer knowledge over the data collection at a conceptual level. However, many logistical issues concerning conceptual annotation of generic data sets have not been standardized. This includes mapping and storage of the annotations, bidirectional query support from the data to its metadata, and inaccurate automated named-entity-recognition software products. Because of these issues, we encountered many barriers while trying to reuse existing knowledge collections in OAMiner.

- There are a number of sources of potential bias with respect to using SMEs when evaluating the validity of integrative hypotheses. SMEs typically have very deep knowledge in a single relatively narrow domain and in essence their knowledge is ‘silied.’ The hypotheses generated using TOKEn will frequently span across these silos of knowledge, while SMEs will tend to be inherently anchored in their domains. When evaluating the concept maps linking the phenotypic markers to the image-derived biomarkers, an SME may dismiss certain hypotheses because the pathway linking them may go outside of their area of expertise. In addition, it is also possible that an SME’s established mental model for the relationship between the presented phenotypic marker and the image-derived biomarker is that they are not related when a plausible pathway is presented. In this situation, it can be difficult for an SME to ignore the anchoring in their reasoning in order to consider other possible outcomes.

CONCLUSIONS
In this report, we have described a design pattern for the integrative and knowledge-based discovery of hypothesis in a high-throughput manner. We also have presented a number of lessons learned from the application of this design pattern in a prototypical research use case. In doing so, we hope to inform future and analogous research and development efforts, and to catalyze further innovation in this timely and critical area of applied biomedical informatics.

Acknowledgments The authors wish to acknowledge to contributions of Dr Peter Embi to the design and evaluation phases of the OAMiner project, as well as Dr David Flanigan for useful discussions and participating as an SME in the validation of generated hypotheses. We also wish to acknowledge the contributions of Mr Omkar Lele to the design of the TOKEn platform.

Contributors PROP, RDJ, TMB, and MNG contributed to the conceptualization and planning of the work summarized in this case report. PROP, TBB, AML, SJ, and MNG designed, implemented, and executed the data generation and analysis pipelines as described. PROP, RDJ, TMB, TBB, AML, SJ, and MNG participated in the preparation of the final manuscript, as submitted.

Funding This work was supported by the National Library of Medicine (R01LM010119, PI: M Gurcan) and the NCRR-funded OSU Center for Clinical and Translational Science (U54RR024394, PI: R Jackson).

Competing interests None.

Ethics approval Ethics approval was provided by The Ohio State University Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES