Dear Bio-Ontologies Organizers,

We thank both reviewers for their time and attention in reviewing our manuscript. We appreciate the positive and constructive comments that you have provided to us and we believe that these changes will enhance our paper. Below are our responses to the inquiries.

**Reviewer 1**

Opening Paragraph Comment:
How representative are the synthetic patient records?

Response:
Seven synthetic patient records were manually created to capture typical medical record data, such as: demographic information, contact information, family history, life style data, allergies, immunizations, information on conditions, procedures, prescriptions, and encounters with members of the medical community. To clarify this point, we have expanded the main body of the text with the following information:

“These records are by no means complete, or unabridged. In practice, clinicians often base care on similar records when treating patients. Patients typically seen by care providers in one health care network, using one software vendor, may present to another hospital outside their network that uses a completely different vendor. This results in an unfortunate, but common real-world scenario that forces the creation of a duplicate EHR, often simplistic and based largely on the data contained in the previous system. In many instances, vital information is transferred via telephone to the new provider as alternative means are often not yet in place to enable electronic transfer or interoperability in a timely fashion. In this way the patients created for this demonstrative project reflect the type of health record one could expect to see in clinical practice. They are basic, yet contain enough data to demonstrate a foundation for more complex query as standardized systems become more prevalent.

The United States Department of Veterans Affairs currently maintains one of the most comprehensive EHR repositories - Computerized Patient Record System (CPRS) using the related clinical software (VistA). A patient’s medical record within this system will likely contain far more detail than the fake patients we have created for TMO. If a patient presents to any facility within the Veteran’s Network, a complete unabridged medical record is fully accessible and may be updated by all who access this record. Often times, patients never leave this network and all of their details remain in one profile. If the patient chooses to receive care at a hospital outside this network as described above, only relevant details pertaining to their care are transferred. A new, abridged EHR is then re-created at each new institution, in much less detail, and is largely similar to the fake patients designed for TMO.”

R1.1 - How is security/privacy of patient records maintained using Semantic Web technologies?
R1.1 Response:
The following text was added to the Results and Discussion section to address patient record security and privacy:

“One ongoing issue in translational informatics is patient privacy and the security of data. An approach that has been pursued using semantic technologies is to encode data access rules and then check all data accesses against these policies. \cite{weitzner2006transparent} For example, a policy can give a hospital billing specialist access to data about procedures performed at the hospital for the purpose of insurance billing. Then, when procedure data is requested, the requester would need to show that they were a billing specialist and provide the purpose for which they want to access the data. Semantic technologies can be and have been used to encode the policies, recognize compliance (or non-compliance), and explain results.”

R1.2 - How does this work make longitudinal data available?

R1.2 Response:

The following information has been added to the Introduction section of the text to describe how the ontology could be used to monitor a patient's progression of disease:

“Electronic Medical Records now act as main repositories for patient data. As we continue to explore the intricate relationship between phenotype and genotype, these records become a vital source for monitoring patients' progression of disease. The presence of a given variation as it relates to the appearance or absence of disease over time can be mapped as encounters are recorded by clinicians. Every result, encounter, event or diagnosis is recorded as a data item and includes a date. This rich longitudinal data provide trends, improvement or decline in state and occurrence or absence of diagnostic criteria that can be used to guide treatment, provide prognosis, or identify those patients likely to respond to a potential treatment. The following example illustrates the kinds of data we seek to integrate and analyze for clinical research purposes. Carvedilol is prescribed to a given patient, while a number of blood pressures and heart rate recordings are taken sequentially over time. If this patient takes their medication as prescribed, we can easily observe trends and establish alerts to adjust this medication. Alternatively, the simultaneous occurrence of any recorded side effects can be correlated more easily with potential causative agents. Increases or decreases in laboratory parameters can also be viewed graphically and displayed for easy review by clinicians. Rich longitudinal data can also provide the opportunity to validate diagnostic procedures and otherwise catch discrepancies between corresponding clinical reports. This application of longitudinal data is being investigated in the HCLS interest group within the context of breast cancer, where a radiology report is followed by a biopsy and a pathology report. There should be a set of corresponding observations within the two reports, with the pathology report corroborating the findings of the radiology report. \url{http://esw.w3.org/HCLSIG/Terminology/PathRadCorrelation}”
Closing Comment:
Where are the semantics in the SPARQL query that identifies AD patients "without the apoe4 allele"? The example only performs a string match against the value of the apoe4 property.

Response:
The structure of the patient data was revised to eliminate the need for string matching in the following ways:

- typing the Test as an instance of owl:Class (e.g., ADmark_Alzheimers_Evaluation)
- typing the Result as an instance of owl:Class (e.g., SNPResult)
- modeling the tested features using URIs and describing the relation among them (e.g., using skos:relatedMatch)
- incorporating the trans:present attribute to indicate presence or absence of the tested feature.

The following diff highlights the changes in representation after these changes (red minus is removed, green plus is added):

```
2.83  trans:test [
2.85 -    trans:testName "ADmark Alzheimer's Evaluation" ;
2.86 +    a :ADmark_Alzheimers_Evaluation ;
2.87
2.88 -    trans:result [
2.89 -        trans:variant_Synonyms "APOE4, NG_007084.2:g.7903T>C" ;
2.90 -    ] ;
2.91 +    trans:result [
2.92 +        a :SNPResult ;
2.93 +        trans:feature :variant_APOE4 ;
2.94 +        trans:present true ;
2.95 +    ] ;
2.103 +:variant_APOE4 skos:relatedMatch :variant_NG_007084_2_g_7903T_C .
2.104 +:variant_NG_007084_2_g_7903T_C dct:identifier "NG_007084.2:g.7903T>C" .
```

The following SPARQL query replaces the original and has been added to the supplement and is referenced in the Results and Discussion section:

```
PREFIX trans: <tag:eric@w3.org:2009/tmo/translator#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT distinct ?name ?patient
WHERE {
```
?patient
    trans:hasCondition [
        trans:diagnosedWith trans:alzheimers_disease
    ];
    foaf:name ?name .
OPTIONAL {
    ?encounter trans:test [
        a ?testname ;
        trans:result ?result
    ].
    ?result
        trans:feature trans:variant_APOE4 ;
        trans:present true .
}
FILTER ( !bound(?result) ) .

Examples of queries that can now be executed with SPARQL
are listed in Table 4. (footnote to http://www.w3.org/wiki/HCLSIG/PharmaOntology/Queries)

**Reviewer 2**

General comments:
With only 7 made up patient records it is difficult to judge the value of the clinical aspect of the ontology.

Response:
Please see response R1.1 above.

R2.1 - The text in the “Use Case” section does not seem to describe a use case. It introduces Alzheimer’s Disease and suggests that data integration is a good idea – citing the New York Times for support – but does not present the very specific actors and actions one would expect from a typical use case. It would benefit the clarity of the paper substantially if the specific use case employed here (which is referred to in a footnote in the Methods section which follows the Use Case section) was presented in the “Use Case” section of the manuscript.

R2.1 Response:
We have included a summary of the patient use case in the use case section.

**Use Case**

Alzheimer’s Disease (AD) is an incurable, degenerative, and terminal disease with few therapeutic options [14] [15]. It is a complex disease influenced by a range of genetic,
environmental, and other factors [15]. Recently, Jack et al. [16] demonstrated the value of shared data in AD biomarker research. A New York Times article quotes John Trojanowski at U Penn Medical School: “It was unbelievable, ...[we] parked our egos and intellectual-property noses outside the door and agreed that all of our data would be public immediately.” Efficient aggregation of relevant information improves our understanding of disease and significantly benefits researchers, clinicians, patients and pharmaceutical companies. We demonstrate the usefulness of TMO and TMKB in a use case that follows a patient and physician from a first report of symptoms, to diagnosis of AD, selection of an optimal treatment regimen, consideration of alternative treatments following the report of side effects caused by the initial treatment, and finally to the selection of possible appropriate clinical trials for the patient.

The AD patient use case can be summarized in the following way:

1. A patient and family members report symptoms to a physician/clinician. The physician/clinician enters the reported symptoms into an EHR. All concepts are mapped to URI’s with the help of TMO.
2. The physician makes a list of differential diagnoses, with a working diagnosis of AD.
3. The physician arranges for the patient to have a basic biochemical, haematological, and SNP profile undertaken. Biochemistry, haematology, and SNP requests are input directly by the various respective departments into the patient’s EHR. Preliminary SNP and genetic data will be submitted directly to the NIH Pharmacogenetics Research Network (PGRN).
4. A follow up meeting is scheduled to perform a set of diagnostic tests outlined by what the clinician feels initially are most appropriate for disease presentation.
5. The physician continues to add investigations/lab results to the patient's EHR and these are combined with the patient’s medical history information. A disease is chosen as the most likely of the listed differential diagnoses based on all of the information provided.
6. The physician confirms and now has a refined and widely acceptable diagnosis of AD with behavioral assessments, cognitive tests, and appropriate brain scan if indicated and enters the diagnosis data into the patient’s EHR.
7. The physician selects the most appropriate AD drug and clinical protocol from the patient's medical record based on the severity of the disease, the patient’s SNP profile (ADME, efficacy/safety based on presence or absence of receptors), patient’s BMI, and concurrent medication, and availability on Medicare D.
8. Fundamental questions will be answered by the ontology at this stage by sourcing the data sets listed below simultaneously or in a specific order:
   o What are the clinically recommended agents?
   o What products are available for prescription, and which are legally indicated for AD disease?
   o What is the SNP verdict? These agents are sourced with a pharmacogenomics database to determine
   o Will they be efficacious? Is the disease receptor positive?
   o Will they be harmful? Are there toxic metabolites? Is CYP 450 or acetylator
status available?

9. Are the preceding predictive genetic SNP tests covered by the patient's insurance company? Are the resulting pharmaceutical agents covered by the patient's specific insurance?

10. The physician checks with the pharmacist, or consults drug information literature to avoid potential drug interactions.

11. The physician now prescribes Aricept (Donepezil) as it satisfies criteria listed above. It is indicated, safe, effective, available, there are no drug interactions issues with drug delivery, and it is covered by the insurance.

12. In a follow-up visit the patient later reports nausea from donepezil. The physician is aware of this common side effect (other side effects reported include bradycardia, diarrhea, anorexia, abdominal pain, and vivid dreams etc...), and re-consults the literature to ensure this is acceptable and agreeable with patient. The physician documents the side effect for post-marketing adverse event pick-up and future study. He changes medication if necessary or adds another medication to alleviate side effects. The physician considers moving patient to a trial.

13. Physician obtains information on all (local, national, and international trials) for AD. Trials might be listed in data sources from the FDA, WHO, Clinical Trials.gov, Citeline TrialTrove, etc.; academia or pharma may also solicit patients, or the physician may point patient to investigators undertaking a trial.

14. The physician decides whether
   o to enroll the patient in a clinical trial as one of the agents looks very suitable and may benefit patient, or because the patient is interested in participating in the trial;
   o not to enroll the patient because the trial is unsuitable or the patient declines to participate in the trial;
   o to obtain information for the patient on trial appropriate for the patient with potential of future enrollment.

15. The physician checks if the patient meets trial inclusion/exclusion criteria by querying EHR.

16. The patient has a thorough medical assessment (lifestyle, medical history, genomics, proteomics, metabolomics, images, cognition) to supplement and update existing data.

17. The results of the medical exam influence the arm of the trial in which the patient participates. The patient status is updated.

Questions relevant for this use case scenario are listed in table 4. Such questions can be formulated as SPARQL queries (see section SPARQL queries, and supplement 1) and answered using TMKB.

R2.2 - The third sentence in the Ontology Design section points to “research questions” in Table 1, but Table 1 does not list any questions.

R2.2 Response:
We modified the Methods section and the Ontology Design section to clarify that Table 1 contains the user roles and interests, rather than questions.

“The work presented here follows questions asked in the patient care scenario that are related to the user roles and interests summarized in Table 1.”

“TMO terms were obtained from a lexical analysis of sample research questions from 16 types of users, all of whom were involved in aspects of research, clinical care, and/or business (see Table 1 for user types).

R2.3 - The fourth sentence in the Ontology Design section “Terms that refer to real world entities are were then represented…” has a typo (are/were). In addition, if the authors are going to raise the specter of the “real world” here in this section, then they need to specify what their criteria for reality are to make it clear to the reader what they mean by ‘real’. What are examples of terms from the texts that they used to start the ontology engineering process that do not refer to entities in the real world?

R2.3 Response:
We have modified the text to remove references to “real world” as this term is highly contentious and is not necessary for the purpose of our work. Thus, we simplify the description of our methodology by asserting that terms are simply formalized as referring to OWL classes, relations or individuals in the ontology, and that we consider types to be those that can be instantiated.

Below are the changes that were made to the manuscript:

Terms that refer to real world entities are were then represented were formalized as classes, relations or individuals in the OWL ontology.

... refer to instantiable types that can be in the real world and are represented as classes in the ontology.

R2.4 - “Relations are specified using the Relation Ontology” should cite a reference for the R.O.

R2.4 Response:
We have added a reference to http://genomebiology.com/2005/6/5/R46.

R2.5 - Figure 3 is referred to in the sentence “Given the prevalence…” as having something to do with equivalence mapping but the Figure in the current version of the manuscript seems to have nothing to do with mappings. This looks like the wrong figure was uploaded.

R2.5 Response:
In the original submissions, the graphic of Figure 2 was mistakenly replaced with a duplicate of that in Figure 3. These are now figures 1 and 2, but we have fixed the graphics as shown below:

Figure 1:

![Graph of Figure 1: Translational Medicine Ontology]

Figure 2:

![Graph of Figure 2: Adverse drug reaction]

R2.6 - The article would be clearer if the questions referred to in both the Unit Testing and Data Mapping sections were spelled out earlier in the text (see comment 1 again here)

R2.6 Response:
Please see the response to R2.1 above.

R2.7 - The “SPARQL queries” section is very dependent on the existence of this footnoted link (http://esw.w3.org/HCLSIG/PharmaOntology/Queries). It might be better for the journal if these queries were included as supplementary data. (Note that all the queries are operational at the time of this review.)

R2.7 Response:
We have generated a supplementary document, which contains the queries and their results.

R2.8 - The first sentence of the SPARQL queries section discusses 12 questions while 14 are listed.

R2.8 Response:
The sentence was corrected to reflect 14 questions, rather than 12.

Other Changes:
The ontology design section was modified to add permanence by snapshot of data and dependent ontologies. The entire text reviewed and minor revisions were made to improve readability.

Again, we thank you for your review and feedback.

Best regards,
The Translational Medicine Team