Response to reviewers of

Enhancing the Representational Power of i2b2 through Referent Tracking

Jonathan C. Blaisure, MSc & Werner M. Ceusters, MD

AMIA 2018

Reviewer Comments (Please note that your submission was reviewed by at least three reviewers and an SPC member. Numbers 1, 2, 3., etc. below represent different reviewers):

SPC member

Thank you for giving us the opportunity to consider your paper for AMIA 2018. The paper demonstrates the development of an i2b2 (Informatics for Integrating Biology and Bedside software platform) modifier system to enhance representations of observational relationships and their referents by applying principles of ontological realism and referent tracking. I agree with the reviewers that the paper is well-written and the study is well-executed. As the reviewers have highlighted the demonstration using a detailed example of a case is a valuable contribution to AMIA. I suggest the authors also add a short paragraph on the lessons learned from this study that can be generalized, and limitations of the approach in their revised submission.

→ **We added this paragraph at the end of the Discussion section.**

Reviewer 1

Thank you for the submission.

In your introduction, you do not mention ICPC, international classification of primary care, which allows primary care doctors to perform referent tracking. Please add.

→ **The ICPC is a coding and classification system, one of the 154 structured vocabularies which are currently integrated in the UMLS meta-thesaurus. As these other systems, it has unique codes for its 'concepts', but it does not have – neither should it have – unique codes for the entities on the side of the patient. Of course, once these entities have been given a unique identifier by a referent tracking system, they can then be annotated by means of concept codes from any system that covers the relevant domain. Following the reasoning of this reviewer, we should then add references to each of these 154 systems. Hence, we did not follow this suggestion.**

In you discussion, you need to talk about the difference between diagnosis that are acute and resolve, and chronic conditions. The fact that some diagnosis resolve or disappear could be portrayed as an error.

→ **We demonstrated in this paper how our proposed approach can be used for a chronic disease and discussed in the last paragraph of the discussion of our original submission the possibility of the clinicians having made a mistake and how this possibility could be represented. We agree with this reviewer that the differences between acute and chronic diseases should be taken into account when clinicians register data in the EHR and that this is not discussed – because of the selected scenario – in our paper. We added this as a limitation of this paper since adding a second scenario, involving an acute disease, is not feasible within the page limitations.**

Another concepts that exists in medicine but you do not discuss, is different illness having similar presentation of symptoms. In your final discussion before the conclusion you talk about not deciding who is correct, however, you could discuss how both could also be correct.

→ **It is of course true that different types of diseases might have similar presentations of symptoms. But the focus of our work here is about construing a representation about what**

clinicians reported in an EHR and how what is reported relates to the actual reality. There is no point here in discussing what is not reported, hence no action taken.

<u>Reviewer 2</u>

This is an extremely dense paper describing an evaluation of the i2b2 data model to accommodate data represented in referent tracking structures. I believe that a general informatics and data science audience will find it difficult to comprehend the more specialized and philosophical content.

→ **We agree, but one of the goals of a scientific conference is continued education. So we hope that with this contribution the general informatics and data science audience will pay more attention to these philosophical principles. The need to take realist ontological principles into account was already pointed out by one such scholar in general informatics in 1978: https://www.amazon.com/Data-Reality-Perspective-Perceiving-Information-ebook/dp/B0086WGJ7W. A detailed explanation of the principles in order to make the paper more easy to understand is unfortunately not possible within the page limitations put forward by AMIA. But we added the requirement to be familiar with these principles as a limitation of our approach.**

The authors have chosen a use case with diagnosis data, and they focus on the inconsistencies and disagreements present in these data. However, I am concerned about this choice.

→ **But then you should also be very concerned about any i2b2 data repository, because diagnoses are precisely the core content of such repositories. In any case, we are concerned too, and that is one of the reasons why we performed our research, i.e. to assess the extent to which we can make such problems explicit.**

The feasibility of representing certain data in a model is an important consideration, but a more intractable issue is the state of data in the real world; just because a model can accommodate a certain representation of data does not mean that these data exist.

→ **The data DO exist as they were extracted from the EHR of which the data are governed by our institution. That the data might not be faithful to what actually was the case in reality, and how our approach handles this, was thoroughly discussed in the last paragraph of the original submission. No action taken.**

The use case appears to be based in health care diagnosis data, but these data frequently contain inconsistencies. The scenario laid out in Table 2, where a patient is represented as having both type 1 and type 2 diabetes, would be a very unlikely scenario in truth. The clinical implausibility of this scenario makes me uncomfortable when it is presented in the context of a "faithful representation of reality."

→ **The scenario discussed was real, not invented: we found in our EHR system actually 650 of such cases with both diabetes type 1 and 2 being ascribed to a patient, sometimes even both being stated active during the same time period! We made this clearer in the final submission. We also highlighted that the goal of 'faithful representation of reality' is not only with respect to what is happening in the patient's body (1st order reality), but also to that what is stated about the patient (2nd order reality).**

Later, the authors state in Table 3 that "this entity only exists if the provider's observations are faithful to reality, which might be doubted." I agree that healthcare data are often messy and inconsistent, but it is unclear to me how this solution addresses the reality.

→ **The entity this reviewer is rightfully arguing against was in the original submission categorized as a 'disease course'. We changed this into 'part of a disease course' which addresses the reality better. With 'disease course' there had to be another disease-entity different from IUI-2. Indeed, since till to date DM type 1 is not curable, we must assume that there never was another disease**

entity of type DM than IUI-2, but that it was this entity that was misdiagnosed as DM1, and then, 2nd mistake, being declared as resolved, rather than having been entered in error. Thus, in reality, there was only one disease course (IUI-3) and it is of this disease course that IUI-5 is a part. This is now explained in the discussion.

In the results section, the authors describe a solution is to store the referent tracking data points in the PATIENT_DIMENSION table because this table allows optional columns to be added. I don't understand this solution because the grain of data do not appear to match and I'd assume that multiple referent tracking instances can be associated with one patient; would this not necessitate multiple rows?

→ **Indeed, that is why there are already many more rows in our solution than what a 'mainstream' i2b2 conversion would lead to: for a diagnosis, there would simply be only one row! No action taken.**

Or, if the intention is to store an array in the PATIENT_DIMENSION table allowing multiple associated instance, this would be helpful to describe in additional detail.

→ **no, that is not the case. No action taken.**

Another solution, to include identifiers for the provider referent tracker entries in the PROVIDER_PATH field, seems problematic.

→ **no, it is allowed according to the i2b2 specifications used at the time of writing this paper. No action taken.**

The paper describes a great deal of effort to fit referent tracking into the i2b2 structuring, but I find it unclear whether it is advisable to do so with healthcare data.

→ **We agree. It would be much better to use only a referent tracking system rather than i2b2 augmented with RT-data. Feel free to be an advocate for this! Many papers have been written about this approach to healthcare data, some of which are referenced in this paper. No action taken.**

Reviewer 3

This paper presents an approach based on principles of Referent Tracking and Ontological Realism to enable additional queries on the i2b2 software platform. The paper is well written in an engaging style and the structure is adequate. The approach is relevant since it makes it possible for the i2b2 platform to deal with a new range of relevant queries that help to avoid ambiguities and confusions. The paper would benefit from (1) some comments on how generalizable the approach is to other platforms; and

→ **Since we did not research this, we cannot comment on it. This would be a different project. No action taken.**

(2) a short description of the next steps of this research.

→ **we did so in the newly added last paragraph of the discussion.**

Reviewer 4

I2b2 is a successful approach for a research data repository. This paper describes a creative method for extending i2b2 to contain additional meta-data that represents information about the assertions (observations) in the i2b2 database using its modifier system. In particular, the modifiers represent information about the relationship between an observation and its referents (either explicit or implied). This allows standard i2b2 clients to query this meta-data in a standard way and answer questions such as whether and when diagnosis were revised or invalidated.

The paper is well written and does a good job of explaining ontological concepts along the way (i.e. Referent Tracking, Ontological Realism, etc) and includes a detailed example to illustrate the approach to augmenting i2b2 using modifiers.

My only recommendation would be to add a short discussion of the limitations of this approach.

→ **We did so.**

The authors should mention whether they have actually implemented this approach in a working i2b2 system and, if they have, what issues they encountered.

→ **We did so**

Also, it would seem that there could be an exponential explosion of additional data / modifiers needed to represent referents of observations which could impact processing time and storage.

→ **We agree. Our preferred solution is using a data-repository managed by a referent tracker server. An example can be found here: https://pennturbo.github.io/Turbo-Documentation/**