Journal of Biomedical Semantics

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis.

--Manuscript Draft--

Manuscript Number:	
Full Title:	Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis.
Article Type:	Research
Funding Information:	National Center for Advancing Translational Sciences (UL1TR001427) — Not applicable
Abstract:	Background The ontological analysis of disease and diagnosis has been the subject of much ontological inquiry, but applying the insights gained when making diagnostic assertions in electronic healthcare records (EHR) suffers from workarounds due to the limitations of the current state-of-the art in EHR design. These limitations include difficulties accounting for misdiagnosis, the various types of entities that diagnoses as information content entities can be about, and the status of coincidentally correct statements about disease. Methods We applied recent advances in our understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created and analyzed six scenarios using the method of Referent Tracking, to identify all the entities and their relationships necessary for each scenario to hold true. We noted deficiencies in any existing ontological definitions and attempted revisions of them to account for any improved understanding resulting from our analysis. Results Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are about the configuration of an organism, a disease inhering in the organism, and the type the disease instantiates. Misdiagnoses are ICEs whose concretizations succeed in aboutness at the level of reference but fail in aboutness at the level of compound expression (i.e., they are not about the configuration). Provenance of diagnoses and their concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief. Conclusions Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR systems that allow diagnostic assertions to be entered with less ambiguity.
Corresponding Author:	William R. Hogan, MD, MS University of Florida Gainesville, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Florida
Corresponding Author's Secondary Institution:	
First Author:	William R. Hogan, MD, MS
First Author Secondary Information:	
Order of Authors:	William R. Hogan, MD, MS Werner Ceusters, MD
Order of Authors Secondary Information:	

Opposed Reviewers:	

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis.

William R. Hogan (corresponding)
University of Florida
P.O. Box 100219
2004 Mowry Rd
Gainesville, FL 32610-0219
hoganwr@ufl.edu
(352) 294-4197

Werner Ceusters
University at Buffalo
921 Main Street
Buffalo, NY 14203
ceusters@buffalo.edu

Abstract

Background

The ontological analysis of disease and diagnosis has been the subject of much ontological inquiry, but applying the insights gained when making diagnostic assertions in electronic healthcare records (EHR) suffers from workarounds due to the limitations of the current state-of-the art in EHR design. These limitations include difficulties accounting for misdiagnosis, the various types of entities that diagnoses as information content entities can be about, and the status of coincidentally correct statements about disease.

Methods

We applied recent advances in our understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created and analyzed six scenarios using the method of Referent Tracking, to identify all the entities and their relationships necessary for each scenario to hold true. We noted deficiencies in any existing ontological definitions and attempted revisions of them to account for any improved understanding resulting from our analysis.

Results

Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are about the configuration of an organism, a disease inhering in the organism, and the type the disease instantiates. Misdiagnoses are ICEs whose concretizations succeed in aboutness at the level of reference but fail in aboutness at the level of compound expression (i.e., they are not about the configuration). Provenance of

diagnoses and their concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief.

Conclusions

Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR systems that allow diagnostic assertions to be entered with less ambiguity.

Keywords

Biomedical ontology

Disease

Diagnosis

Information content entity

Representation

Ontological realism

Background

The nature of information, especially in a biomedical and healthcare context, has been the subject of significant ontological inquiry. Much of this work strictly views information as being about one entity, such as one particular or one universal, although Ceusters [1] and more recently Smith and Ceusters [2] clarify that the ontological theory behind the Information Artifact Ontology never intended such a restriction, and that information can also be about a portion of reality that is composed of multiple particulars, universals, and the relations among them.

Nevertheless, the most popular logical formalism for ontology development, the Web Ontology Language or OWL, further constrains aboutness to a single particular, because it does not allow the assertion of any relation between OWL individuals and OWL classes other than one relation akin to "instance of" (i.e., the Resource Description Framework "type" property). The work of Schulz et al. explores options and their consequences for working around this constraint of OWL to assert that an information content entity is about a universal [3].

Research on the aboutness of clinical statements like diagnoses has also been constrained by the view that nothing is an information content entity (ICE) if it misrepresents reality. Restated, this view considers only perfectly correct statements as meeting the criteria for being an ICE. For example Martínez Costa and Schulz use the universal quantifier when defining 'information entity about a clinical situation' *…to avoid asserting the existence of an entity the existence of which cannot be guaranteed* [4]. In other words, if they had used the

existential quantifier, then if a particular description of a "clinical situation" were not completely correct, it would presumably not meet the criteria of being an *information entity about a clinical situation*. Because the universal quantifier does not imply the existence of something that the information entity is (correctly) about, its use served as a workaround to perceived limitations in the theory of ICEs at the time. Researchers working in areas other than diagnosis encountered similar issues. For example, Hastings et al. note that chemical graphs and diagrams are not always about types of molecules that exist [5]. They, too, used the workaround of replacing existential quantification with universal quantification to avoid asserting that every chemical graph/diagram has to be a completely correct representation of some type of molecule in order to be an ICE.

In our own, previous ontological analysis of diagnosis, using the methodology of referent tracking, we identified what entities must exist or must have existed for a particular diagnostic statement to hold true [6, 7]. A key result of this work is that a diagnosis is minimally about *both* the patient and the type of disease asserted to exist. In addition, building on previous work on the Ontology for General Medical Science, the foundations of which were laid down in Scheuermann et al. [8], we also noted that for a diagnosis to exist (at least in medicine and under the assumption that the diagnosis was made *lege artis*), there must also have existed a diagnostic process, a person who carried out that process, and a clinical picture which was used as input into the process.

In this work, we advance our work on disease and diagnosis, while at the same time addressing the strong restriction placed on ICEs that they be completely correct, by

building on two key aspects of Smith and Ceusters' recent results in their work on aboutness [2]. The first result is that they hold that a statement such as a diagnostic assertion can succeed or fail in aboutness at two levels: (1) the level of denotating single entities (i.e., the level of *reference*) and (2) the level of veridical representation of a configuration of multiple entities (i.e., the level of *compound expression*). The second result is that whenever an aboutness relation holds, it does not hold between some information content entity (ICE) and some portion of reality (POR), but rather between the quality which concretizes some ICE and some POR.

Our hypothesis is that these advances can address limitations encountered in previous ontological work on disease and diagnosis, where ontological workarounds such as those described above are common, and consequently improve our ontological representations of them.

Methods

To test this hypothesis, we analyzed a set of scenarios that we created and that involve correct and incorrect diagnoses, lucky guesses, and justified layperson belief in the existence of a disease of a certain type. The goal was to explore the ideal case of a correct diagnosis, and systematically from there, deviations from the ideal.

Materials

In our analysis we used as input (1) Smith and Ceusters' work on aboutness and their definitions of representation, mental quality, cognitive representation, and information

quality entity (Table 1) and (2) definitions of disease, disorder, and diagnosis from the Ontology for General Medical Science (Table 2).

Smith and Ceusters stressed that the relation of aboutness includes any portion of reality, rather than being limited to just a single particular or a single universal. A portion of reality can be a particular, a universal, a relation, or a configuration. A configuration is a combination of particulars and/or universals and certain relation(s) that hold among them.

A representation, then, that is intended to be about a portion of reality but fails in its aboutness because it misrepresents that POR in some way, is misinformation. The sentence *Bob Dylan was in the Beatles* fails to represent not because Bob Dylan or the Beatles did not exist, but because such a configuration involving Bob Dylan and the Beatles in the way as expressed, never existed. The sentence fails in aboutness at the level of compound expression, but nevertheless is about Bob Dylan and the Beatles individually (level of reference) and thus is still an information content entity.

In addition to Smith and Ceusters' work, we also founded our ontological analysis on the Ontology for General Medical Science or OGMS [8]. This work distinguishes disease, disorder, and diagnosis, and we used definitions from OGMS as starting points for our analysis (Table 2). Note that in OGMS, a diagnosis refers to the existence of a disease of a given type. In clinical medicine, however, diagnoses also refer to (1) disease courses (e.g., acute hepatitis vs. chronic hepatitis), (2) disorders (e.g., fractures and tumors), and (3) the absence of any disease (i.e., a conclusion that a person is healthy also is a diagnosis). It was

not our intent to address this issue in this work, as it was not our goal to refine the OGMS definition of diagnosis. We revisit the implications of this issue in the discussion.

The scenarios

All the scenarios have in common a particular patient, Mr. Adam Jones, who suffers from type 2 diabetes mellitus. Thus in every scenario, there exists Mr. Jones, his disease, and its instantiation of the type *Type 2 diabetes mellitus* (Figure 1).

Scenario 1: correct diagnosis by physician (ideal case)

Dr. Anne Smith sees Mr. Jones in the office. She takes a history and physical, performs certain laboratory testing, and based on her analysis of the findings, correctly concludes that Mr. Jones has type 2 diabetes mellitus. She subsequently writes her diagnosis in the patient's medical record.

Scenario 2: subsequent correct diagnosis by physician using first diagnosis

A second doctor, Dr. John Brown, sees Mr. Jones in the office at some later date. Mr. Jones has released his records from Dr. Smith to Dr. Brown, who subsequently sees Dr. Smith's diagnosis prior to seeing Mr. Jones. He uses that diagnosis plus his own findings to infer a new clinical picture of Mr. Jones, which he subsequently uses to make another correct diagnosis of Mr. Jones' disease. He writes his diagnosis in Mr. Jones' medical record.

Scenario 3: incorrect diagnosis by physician

Mr. Jones is traveling on vacation, when he falls ill. He sees Dr. Jane Miller who does not have any of his past records available, and thus she is not aware of the previous diagnoses of Drs. Smith or Brown. She infers a new clinical picture of Mr. Jones, and based on it incorrectly concludes that Mr. Jones has *type 1 diabetes mellitus* (as opposed to type 2). She records a diagnosis of type 1 diabetes mellitus in her medical record for for Mr. Jones.

Scenario #4: coincidentally correct conclusion by layperson (lucky guess)

A friend of Mr. Jones is a "seer". Mr. Jones asks his friend what is in his future. Having no prior knowledge of Mr. Jones medical conditions, the "seer" concludes based on Mr. Jones horoscope and the position of the moon that he has type 2 diabetes mellitus. He subsequently predicts that Mr. Jones will be hospitalized for his diabetes and miss his daughter's wedding.

Scenario #5: layperson's justifiable conclusion

Mr. Jones' daughter, upon learning of her father's type 2 diabetes mellitus, adds this information to her letter to her brother, writing "Dad has type 2 diabetes mellitus".

Scenario #6: correct diagnosis by computer-based expert system

A medical student is seeing Mr. Jones in the clinic. He performs a history and physical, and types his findings into a diagnostic expert system. The diagnostic expert system infers based on these findings that Mr. Jones has type 2 diabetes mellitus. The medical student writes this diagnosis in Mr. Jones medical record.

The analysis

Our analysis follows the method of Referent Tracking, which we have found to be a stringent test of ontologies and their definitions [6]. This approach proceeds in three main steps. First, we systematically identify all the relevant particulars that must exist for the scenario to be true, regardless of whether the scenario explicitly mentions them or implies their existence. We assign each particular an instance unique identifier (IUI), of the form 'IUI-n', where 'n' is any integer. Second, we identify for each particular the type it instantiates and the temporal interval during which it exists (and assign an identifier of the form *tn* to that interval). Lastly, we identify all the relationships that hold between the particulars as well as all relevant relations particulars have to universals other than instantiation, including situations where a particular lacks a given relation to any instance of a certain type (for example, a statement that a patient has had no cough in the last two weeks means that the patient does not stand in the agent_of relation to any instance of the type *Coughing event*, indexed temporally to the two-week interval).

This approach identifies problems in ontologies and their definitions in two major ways. First, it identifies problems that occur when the scenario explicitly rules out the existence of a particular whose existence is implied by an ontological definition (and vice versa). Second, it helps identify exceptions to existing definitions and situations that should not fall under a definition but are erroneously captured by it. Definitions in ontologies can subsequently be adjusted to avoid the errors so identified.

Although our approach is to identify particulars implied by sentences in natural language, the ontological analysis of language and the mechanism(s) by which it makes implicit reference to certain entities is not the focus of this work. Therefore, we convert a sentence like "Mr. Jones has type 2 diabetes mellitus" to Referent Tracking Tuples (e.g., as in Tables 3 through 7) and it is these tuples in which inhere representations that are the objects of our analysis.

To simplify our analysis somewhat, we wrote the scenarios such that humans record diagnoses on paper. However, concretization of ICEs also occurs by pixels on monitors, binary switches in memory and processor chips, and magnetic fields on hard disks. But a detailed account of these concretizations and transformations among them is not central to our analysis of what is a diagnosis. Our analysis can be extended to these concretizations without modification.

Results and discussion

In each scenario, Mr. Jones (IUI-1) and his disease (IUI-2) exist, the latter inhering in the former (Table 3). Furthermore, his disease is an instance of type 2 diabetes mellitus. Mr. Jones (IUI-1) exists through a certain period of time (*t1*) of which we do not know the exact beginning or end. We use temporal identifiers of the form '*t*n' to clearly distinguish such identifiers from IUIs: where IUIs are always intended to be globally and singularly unique, distinct temporal identifiers may denote a unique period of time which is also denoted by another temporal identifier. We also assign an identifier to the time interval during which his disease (IUI-2) exists (*t2*). Diseases usually begin to exist after the organism does, but

in the case of congenital genetic diseases, the two intervals might be coextensive. Also, we assume that disease IUI-2 existed at the time of diagnosing, but we recognize that diagnosing a disease thousands of years after it existed is possible, such as in the case of archaeologists' recent diagnosis of Tutankhamun's malaria [9].

Note that the configuration of organism, disease, and disease type is anchored at a particular location in spacetime, as is the diagnosis. But note also that the diagnosis additionally has an implicit or explicit reference to the location of the configuration in spacetime. To be a correct diagnosis, this reference must also be correct (it has to refer to some part, not necessarily the entirety of spacetime, occupied by the configuration). Thus, for example, to say that Tutankhamun had malaria in 1000 C.E. or today is incorrect, as it would be to say that Mr. Jones had type 2 diabetes mellitus before his parents were born.

***Scenario 1: correct diagnosis*.**

In this scenario, numerous entities must also exist and stand in certain relationships (Tables 4-6) in addition to Mr. Jones and his disease. Before Dr. Smith (IUI-3) writes (IUI-13) her diagnosis (IUI-8), she first has a cognitive representation (IUI-6) that concretizes it in some anatomical part (IUI-5) of her cognitive system (IUI-4). Note that we follow Ceusters and Smith [10] in asserting that all anatomical entities in which cognitive representations inhere are part of a person's cognitive system (that is, any entity used in cognition, including the bearing of cognitive representations, are necessarily within a person's cognitive system) at least during the temporal interval that the cognitive representation exists. If, for example, it would be the case that some white blood cell

flowing through some brain capillary would through some of its molecules take part in the concretization of a cognitive representation, then that white blood cell would be part of the cognitive system at least during the existence of that concretization. It would not anymore be part of the cognitive system once it continues its journey through the body without participating in thought formation. Additionally, Ceusters and Smith take the position (which we also follow) that the cognitive system is not necessarily strictly limited to the brain or even to the entire neurological system of a person: the current state-of-the-art of neuroscience is yet searching for answers to questions such as "what is it in which cognitive representations inhere?" but until it reaches such answers, we remain in our representations agnostic.

IUI-9 denotes the sentence Dr. Smith wrote, as it exists on the particular piece of paper she used to write it: "The patient has type 2 diabetes mellitus". This writing on paper (IUI-9) bears an information quality entity or IQE (IUI-10) that concretizes her diagnosis (IUI-8). The cognitive representation (IUI-6) and IQE (IUI-10) that concretize the diagnosis are both about the configuration (IUI-7) in Table 1 (the level of compound expression), as well as about Mr. Jones, Mr. Jones' disease, and the universal *Type 2 diabetes mellitus* individually (the level of reference). The cognitive representation (IUI-6) and the diagnosis (IUI-8) are the output of Dr. Smith's diagnostic process (IUI-11), which had as input Dr. Smith's clinical picture (IUI-12) of Mr. Jones. Because the cognitive representation and IQE concretize the same ICE, the latter is conformant to the former (see Table 1).

A correct diagnosis is thus fundamentally an information content entity that is concretized by a representation that stands in an is_about relation to the configuration of an organism, its disease, the relation of inherence between the disease and the organism, and a type that the disease instantiates (Figure 2). Furthermore, diagnoses are additionally differentiated from other ICEs by the fact that they are generated by a diagnostic process that has a clinical picture as input. Note that it is trivial to state that the particular disease inhering in the organism is an instance of *entity* or even *disease*. Thus, there is an expectation that a diagnosis be as precise (the most specific type) as possible and as is relevant to the treatment of the patient.

Scenario 2*: *second diagnosis.

The second physician, Dr. Brown, makes a second diagnosis at a later point in time, using the first diagnosis in addition to clinical and possibly other findings to infer a new clinical picture of Mr. Jones. With the exception of the configuration of Mr. Jones/his disease/Type 2 diabetes mellitus (IUI-7), there is a one-to-one correspondence of entities as in Scenario 1, numbered IUI-23 through IUI-33 (Additional file 1 : Tables S1-S3). That is, there is no IUI-27 because the configuration is the same entity across scenarios. Similarly, there is no IUI-21 or IUI-22 because Mr. Jones (IUI-1) and his disease (IUI-2) are the same entities across scenarios.

In this scenario, Dr. Brown (IUI-23) makes a new diagnosis (IUI-28), concretized both by his cognitive representation (IUI-26) in some part (IUI-25) of his cognitive system (IUI-24) and by the IQE (IUI-30) inhering in the sentence in his note (IUI-29). Dr. Smith's previous

diagnosis (IUI-8) can be viewed as either (*view1*) being in the aggregate of things that Dr. Brown uses to infer his clinical picture (IUI-32) that serves as input into his diagnostic process (IUI-31), or (*view2*) as something which serves as extra input – alongside his clinical picture – for the diagnostic process. The cognitive representation and the IQE are about the configuration (IUI-7) as well as Mr. Jones (IUI-1), his disease (IUI-2), and type 2 diabetes mellitus (UUI-1).

The current definition of 'clinical picture' in OGMS (see Table 2) seems to conflict with *view1* about this scenario, because the definition seems to exclude using a past diagnosis to infer a clinical picture: although the current OGMS definition of 'clinical picture' is inclusive of clinical findings, diagnosis as currently defined is not an explicit subtype of clinical finding in OGMS. Furthermore, it is common for clinicians to elicit a previous provider's past diagnosis from the patient or the patient's caregiver during an interview (for example, if Mr. Jones in scenario #2 would have said: 'Dr. Smith says I have type 2 diabetes mellitus'). But the current OGMS definition of 'clinical history' (Table 2) conflicts with this possibility. It refers to health-relevant features of a patient, but features as elucidated by OGMS include only qualities, processes, and physical components of the organism—not dispositions of which disease is a subtype. Therefore, a representation of a disease such as a diagnosis is currently excluded from the OGMS definition of 'clinical history'.

We also note that the OGMS definition of 'clinical picture' is ambiguous in that it is not clear whether it *requires* that laboratory and image findings must always be used to infer a clinical picture, or that they are the only things that can be used. Regardless, it would be a

mistake to do so, because diagnoses can and frequently are made from clinical findings alone. Laboratory and image findings are not necessary components of a clinical picture in reality.

Given that clinical history taking elicits past diagnoses routinely in clinical medicine, we propose modifying the definition of 'clinical history' to accommodate this reality (bolded sections represent changes to the definition):

Clinical history =def. – *A series of statements representing one or more health-relevant features**, diseases, and their configurations** of a patient.*

Note that the definition already allows—under the broader heading of 'feature'—representations of disorders (kinds of physical component) and disease courses (kinds of process). Thus, the definition already accommodates these aspects of clinical histories. We also allow the statements to represent configurations, in line with Smith and Ceusters [2]. These configurations might or might not include various relevant types (for example, "The patient has not participated in any instance of vomiting in the last two weeks."). Finally, note that by using the word 'representing', the definition also accommodates per Smith and Ceusters [2] that some statements might fail in aboutness despite their intention to be about such features. In other words, some statements in the clinical picture might be wrong: for example, a statement that the patient has a disease or pain that she does not in fact have.

To clarify that laboratory and imaging findings are not always required inputs into the diagnostic process, and to capture realistic scenarios compatible with *view2* (for example, Dr. Brown reads Dr. Smith's note in the chart), we also propose a modified definition of 'clinical picture' (changes in bold) :

Clinical Picture =def. – *A representation of a clinical phenotype that is inferred from the combination of,* ***for example, diagnoses and*** *laboratory, image, and clinical findings about a given patient.*

These changes to the definitions of 'clinical history' and 'clinical picture' now properly capture situations where past diagnoses are elicited from the patient and/or her caregiver during a clinical history taking: these diagnoses are now clinical findings in the clinical history that was generated by the clinical history taking (see the definition of 'clinical finding' in Table 2).

Scenario 3: Misdiagnosis.

The third physician, Dr. Miller, misdiagnoses Mr. Jones' type 2 diabetes mellitus as type 1 diabetes mellitus (Figure 3). Per Smith and Ceusters, because the misdiagnosis is still about Mr. Jones, his disease, the relationship between them, and type 1 diabetes mellitus (it is about these entities at the level of reference), it is an information content entity. However, it fails to about the configuration IUI-7 (level of compound expression) as a whole.

Again, in this scenario there exist entities in one-to-one correspondence (except the configuration and its components) numbered IUI-43 through IUI-53 (Additional file 2 : Tables S4-S6). Dr. Miller (IUI-43) writes (IUI-53) his misdiagnosis (IUI-48) in Mr. Jones' chart, and the IQE (IUI-50) inhering in the ink (IUI-49) is conformant to his cognitive representation (IUI-46), and both are about—at the level of reference—Mr. Jones, his disease, and type 1 diabetes mellitus. But neither one is about the configuration (IUI-7). To capture the relation both (1) between the cognitive representation and the configuration and (2) between the IQE and the configuration, we define a new relation:

> **is-misrepresentation-of**: domain: representation, range: portion of reality.
>
> Def: *x* is-misrepresentation of *y* iif *x* is a representation and *x* is intended to be about *y* and it is not the case that *x* is about *y*.

Then we assert that the representations (IUI-46 and IUI-50) are misrepresentations of the configuration (Table 7 and Additional file 2 : Table S6). Note that our definition precludes the cognitive representation (IUI-46) and IQE (IUI-50) being about any configuration other than IUI-7, because they are not intended to be about, for example, the configuration of the sun, earth, and moon on a particular date and time.

Note that asserting the incorrect disease type is not the only way to make a misdiagnosis. There are at least six possibilities where a diagnosis fails to be about a configuration at the level of compound expression (Table 8). If a representation fails at the level of reference, it also fails at the level of compound expressions, because a configuration can not be made up of that which does not exist. These six possibilities could also exist in combination, but

if the 2nd, 3rd, and 4th possibilities are all present (for example, "Ron Weasley has spattergroit"), then there is not a diagnosis, or even any information content entity at all, because the representation is not about anything, even at the level of reference. Of course, if the organism itself does not exist, then there cannot be a clinical picture inferred, and thus it would not be a diagnosis or misdiagnosis, although it could still be an ICE if it is about a really-existing disease type (for example, "James Bond has influenza").

It is likely that the most common kind of misdiagnosis today is of the first type, where the disease and disease type exist, but the disease does not instantiate the type (Figure 3). Also, as medical knowledge evolves, the profession comes to understand that certain types of disease thought to exist in fact do not. Thus past diagnoses of *dropsy* and *consumption* we now understand to be misdiagnoses.

Scenario 4: the lucky guess.

In this scenario, a layperson (the "seer"—IUI 63) correctly concluded coincidentally that Mr. Jones had type 2 diabetes mellitus based on the position of the moon and Mr. Jones' horoscope (Additional file 3 : Tables S7-S9). It would be wrong to say the seer's reasoning (IUI-71) constituted a diagnostic process. To avoid coincidentally correct statements from qualifying as diagnoses, we additionally require as input into the diagnostic process cognitive representations of the disease type and its typical sequalae, signs, symptoms, and any clinical, laboratory, or imaging findings or phenotypes. This is a minimal requirement: clinicians often additionally include in their diagnostic reasoning cognitive representations

of other disease types and associated entities when considering alternative possibilities for the disease type.

Because the seer had no cognitive representations of type 2 diabetes mellitus, let alone used them as input into his "reasoning", his conclusion (IUI-68), although an ICE, is not a diagnosis. Similarly, if a physician makes a lucky guess based not on his cognitive representations of the stated disease type but instead by flipping a coin or some such, that too would not be a diagnosis.

To Table 3 we add an aggregate of cognitive representations of disease types and associated entities as input into the diagnostic process (Table 9).

We propose to redefine diagnostic process as follows:

Diagnostic process =def. *An interpretive PROCESS that has as input (1) a CLINICAL PICTURE of a given patient AND (2) an aggregate of REPRESENTATIONs of at least one type of disease and at least one type of phenotype whose instances are associated with instances of that disease, and as output an assertion to the effect that the patient has a DISEASE of a certain type.*

Scenario 5: layperson's justifiable conclusion.

Mr. Jones' daughter wrote a sentence in her letter to her brother based on reading Dr. Smith's progress note saying that that her father has type 2 diabetes mellitus (Additional

file 4 : Tables S10-S12). Of course, the daughter has not made a diagnosis. She is communicating to her brother what she believes to be the case.

Had she merely written "Dr. Smith says" and then copied Dr. Smith's sentence word for word into her letter, then her writing would concretize Dr. Smith's diagnosis (IUI-8). Similarly, if a person who does not know German copies German texts, then she reproduces the representations in the texts and they concretize the same ICEs as the original texts. This is the case of hearsay ("so-and-so said it was the case that…").

As Smith and Ceusters showed, however, the same sentence written by two different people does not guarantee they concretize the same ICE. ICEs are further differentiated by the provenance of their concretizations, including who created them and when, and to what POR they intend to be about. In their example, two people writing the sentence *Barack Obama has never been President of the United States*—one before and one after Obama's inauguration as President—generate two different ICEs. The one written after fails at the level of compound expressions but not at the level of reference, whereas the one written before succeeds in aboutness at both levels (it remains true that at the time when the sentence was written, he had never been President).

We therefore distinguish between a human (1) merely copying a representation, in which case the copy concretizes the same ICE as the original text and (2) creating her own cognitive representation of the POR—which involves forming a belief that the POR really existed as represented—and then subsequently creating an IQE that is conformant to the

cognitive representation. In the former case, a new ICE does not come into being. It does not even require in the cognitive system of the copier any representation of the POR that the original representation is about (as in the case of copying German text that one does not understand at all). In the latter case, by contrast, a new ICE does come into being.

In Scenario 5, the daughter did not merely repeat Dr. Smith's diagnosis. She communicated to her brother *her* belief about her father's disease. She deliberately chose not to merely convey Dr. Smith's diagnosis, but rather her belief that her father has type 2 diabetes mellitus. She heard the opinion of an expert, in whom she had trust. Based on (1) her observations of her father, (2) Dr. Smith's diagnosis, and (3) her trust in Dr. Smith, she reached the conclusion herself that her father suffers from type 2 diabetes mellitus. Because she lacks the clinical skill and knowledge to infer a clinical picture and make a diagnosis, her conclusion however is not a diagnosis.

However, consider the scenario where she is given the clinical picture and has enough knowledge to arrive at a conclusion, which could be the case either if she were a physician or somehow other acquired or were given the necessary knowledge: it is analagous to Scenario #6, where she takes the place of the expert system (see analysis of that scenario below). Thus, here in Scenario #5 it is important to note that she does not reason from a clinical picture to the diagnosis.

In this scenario, therefore, the daughter has created a new ICE (IUI-88) that is not a diagnosis. She has concretized it in the sentence (IUI-89) in her letter.

Scenario 6: diagnosis by non-human.

The diagnostic decision support system has made a diagnosis (or misdiagnosis depending on whether it is correct), because it (1) takes as input a clinical picture and representations of the relevant disease type and one or more types of phenotypes with which it is associated; (2) participates in a process of making a conclusion based on this input; and (3) outputs from this process a statement about a configuration involving an organism, a disease, and a disease type.

In this case, there are no cognitive representations. In their place are digital representations on hard drives, memory chips, and central processing units. If we assume the system generates a sentence and prints it on paper, then we have an analagous IQE to the written diagnosis of the physician and ICE of the sister.

Nothing in our proposed definitions conflicts with this scenario. Replacing Dr. Smith and associated representations and diagnostic process with various components of the computer and its digital representations as well as inferential process (which is an instance of diagnostic process) is straightforward.

Returning briefly to a point made in Scenario #5, Mr. Jones' daughter could follow the exact same algorithm(s) of the diagnostic expert system using the exact same clinical picture as input, and she would arrive at (or make) a diagnosis, in contrast to scenario #5 where her conclusion was an ICE but not a diagnosis.

Conclusions

We applied Smith and Ceusters' results on aboutness [2] to diagnosis in order to develop an account of diagnosis, misdiagnosis, lucky guesses, hearsay, a layperson's justified belief about disease configurations, and a diagnosis made by an expert system. Our key result is that a correct diagnosis, as defined by OGMS, is about a configuration of an organism, its disease, and the type the disease instantiates (level of compound expression). The most common forms of misdiagnosis are information content entities whose concretizations succeed in aboutness at the level of reference, individually referring to an organism, disease, and/or type of disease, but fail in aboutness at the level of compound expression. Also, the provenance of the ICE and its concretizations are critical: lucky guesses, hearsay, and laypersons' conclusions about disease (when not arrived at through a diagnostic process) do not constitute diagnoses and therefore are different types of ICE than diagnoses.

Smith and Ceusters' results on aboutness and our extension of them here to diagnosis eliminate the need for the workarounds reported by Martínez Costa and Schulz [4] and Hastings et al. [5] It is perfectly legitimate to define 'information entity about a clinical situation' with the existential quantifier: if an instance of this type is not completely correct, it is still an ICE that is individually about the patient and any types of disease, disorder, or other entity referenced by it. Similarly, chemical graphs and diagrams are ICEs about individual types of atoms such as carbon, oxygen, hydrogen, and so on, even when they fail to be about any type of configuration (molecule) of such atoms. There is therefore

no need to use universal quantification when defining ICEs to avoid failure of aboutness at the level of compound expression. Note that our work here, however, does not address the limitation of OWL that prevents asserting that an ICE is about a type (as opposed to a particular), without using workarounds such as those assessed by Schulz et al. [3]

Future work includes (1) an account of differential diagnosis, where a clinician or expert system generates a list of likely types of disease for further investigation to identify the actual type the organism's disease instantiates; (2) proposing to the OGMS community to clarify the definitions of 'clinical history' and 'clinical picture' as suggested here, and to expand the definition of diagnosis to include disorders, disease courses, and absence of disease (i.e., healthy); and (3) extending our analysis as reported here to this expanded definition of 'diagnosis'.

List of Abbreviations

BFO	Basic Formal Ontology
GDC	Generically dependent continuant
IAO	Information Artifact Ontology
ICE	Information Content Entity
IQE	Information Quality Entity
OGMS	Ontology for General Medical Science
POR	Portion of Reality
RT	Referent Tracking
RTT	Referent Tracking Tuple

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

The authors contributed equally to the ontological analysis and development of results. Author WRH created the first version of the manuscript. Both authors had full access to all materials and analysis and participated in revising the manuscript. Both authors approved the final version of the manuscript.

Acknowledgments

This work was suppoted in part by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1TR001427.

References

1. Ceusters W: **An information artifact ontology perspective on data collections and associated representational artifacts**. *Stud Health Technol Inform* 2012, **180**:68-72.
2. Smith B, Ceusters W: **Aboutness: Towards Foundations for the Information Artifact Ontology**. In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015.
3. Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A: **An Ontological Analysis of Reference in Health Record Statements**. In: *Formal Ontology in Information Systems: Proceedings of the Eighth International Conference (FOIS 2014): 2014*. IOS Press: 289.
4. Martínez-Costa C, Schulz S: **Ontology-based reinterpretation of the SNOMED CT context model**. In: *Proceedings of the Fourth International Conference on Biomedical Ontology: July 7th-12th, 2013; Montreal*: Edited by Dumontier M, Hoehndorf R, Baker CJO. 2013: 90-95.
5. Hastings J, Batchelor C, Neuhaus F, Steinbeck C: **What's in an 'is about' link? Chemical diagrams and the information artifact ontology**. In: *Proceedings of the*

2nd International Conference on Biomedical Ontology; Buffalo, New York: Edited by Bodenreider O, Martone ME, Ruttenberg A. 2011: 201-208.

6. Ceusters W, Hogan WR: **An ontological analysis of diagnostic assertions in electronic healthcare records** In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015.
7. Hogan WR: **To what entities does an ICD-9-CM code refer? A realist approach.** In: *Bio-ontologies; Boston, MA*: Edited by Shah N, Sansone S-A, Stephens S, Soldatova L. 2010.
8. Scheuermann RH, Ceusters W, Smith B: **Toward an ontological treatment of disease and diagnosis**. In: *AMIA Summit on Translational Bioinformatics: 2009*. 116-120.
9. Hawass Z, Gad YZ, Ismail S, Khairat R, Fathalla D, Hasan N, Ahmed A, Elleithy H, Ball M, Gaballah F *et al*: **Ancestry and pathology in King Tutankhamun's family**. *JAMA* 2010, **303**(7):638-647.
10. Ceusters W, Smith B: **Foundations for a realist ontology of mental disease**. *J Biomed Semantics* 2010, **1**(1):10.

Table 1. Definitions based on Smith and Ceusters [2].

Term	**Definition**
INFORMATION CONTENT ENTITY	An ENTITY which is (1) GENERICALLY DEPENDENT on (2) some MATERIAL ENTITY and which is (3) concretized by a QUALITY (a) inhering in the MATERIAL ENTITY and (b) that is_about some PORTION OF REALITY
INFORMATION QUALITY ENTITY	A REPRESENTATION that is the concretization of some INFORMATION CONTENT ENTITY
REPRESENTATION	A QUALITY which is_about or is intended to be about a PORTION OF REALITY
MENTAL QUALITY	A QUALITY which specifically depends on an ANATOMICAL STRUCTURE in the cognitive system of an organism
COGNITIVE REPRESENTATION	A REPRESENTATION which is a MENTAL QUALITY
Relation	**Explanation**
x ***is_about*** *y*	*x refers to or is cognitively directed towards y*. **Domain:** representations; **Range:** portions of reality
x ***concretizes*** *y*	*x* is a QUALITY and *y* is a GENERICALLY DEPENDENT CONTINUANT (GDC) and for some MATERIAL ENTITY *z, x* ***specifically_depends_on*** z at *t* and *y* ***generically_depends_on*** *z* at *t*, and if *y* migrates from bearer *z* to another bearer *w* then a copy of *x* will be created in *w*.
x ***is_conformant_to*** *y*	=def. *x* is an INFORMATION QUALITY ENTITY and *y* is a COGNITIVE REPRESENTATION and there is some GDC *g* such that *x* ***concretizes*** *g* and *y* ***concretizes*** *g*.

Table 2. Key definitions from OGMS used in the analysis

Term	Definition
DISEASE	A DISPOSITION (i) to undergo PATHOLOGICAL PROCESSes that (ii) exists in an ORGANISM because of one or more DISORDERs in that ORGANISM.
DISORDER	A causally relatively isolated combination of physical components that is (a) clinically abnormal and (b) maximal, in the sense that it is not a part of some larger such combination.
DIAGNOSIS	A conclusion of an interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion (diagnostic statement) to the effect that the patient has a DISEASE of such and such a type.
DIAGNOSTIC PROCESS	An interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion to the effect that the patient has a DISEASE of a certain type.
PATHOLOGICAL PROCESS	A bodily PROCESS that is a manifestation of a DISORDER.
PHENOTYPE	A bodily feature or combination of bodily features of an organism determined by the interaction of the genetic make-up of the organism and its environment.
CLINICAL PHENOTYPE	A clinically abnormal PHENOTYPE.
CLINICAL PICTURE	A representation of a CLINICAL PHENOTYPE that is inferred from the combination of laboratory, image and clinical findings about a given patient.
CLINICAL FINDING	A REPRESENTATION that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof.
MANIFESTATION OF DISEASE	A QUALITY of a patient that is (a) a deviation from clinical normality that exists in virtue of the realization of a disease and (b) is observable.
CLINICAL HISTORY TAKING	An interview in which a clinician elicits a clinical history from a patient or from a third party who is authorized to make health care decisions on behalf of the patient.
CLINICAL HISTORY	A series of statements representing health-relevant features of a patient.

Table 3. Referent tracking tuples true in every scenario

IUI	**Entity**	**Existence period**	**Type**	**Notes**
IUI-1	Mr. Adam Jones	*t1* – the period during which IUI-1 exists	Material Entity	
IUI-2	IUI-1's disease	*t2*	Disposition	
Relationships among particulars				**Notes**
IUI-2	**inheres in**	IUI-1	at *t2*	
IUI-2	**instance of**	UUI-1	at *t2*	UUI-1 is a universal unique identifier that denotes *type 2 diabetes mellitus.* We assume that if something is at any time of its existence an instance of type 2 DM, it is instance of type 2 DM at all times it exists.

Table 4. The entities in Scenario 1

IUI	Entity	Existence period	Type	Notes
IUI-3	Dr. Anne Smith	t3	Human being	
IUI-4	Cognitive system of IUI-3	t4		
IUI-5	An anatomical entity that is part of IUI-4	t5	Anatomical entity	Which anatomical entity and its lifetime cannot be easily specified given current state of neuroscience.
IUI-6	Quality that inheres in IUI-5 and is about IUI-7	t6	Cognitive representation	
IUI-7	The POR that is truth-maker for IUI-8	t7	Configuration	Mr. Jones, his disease, their relationship, and disease's instantiation
IUI-8	Dr. Smith's diagnosis	t8	Diagnosis	ICE concretized by IUI-6 and IUI-10
IUI-9	That which is written down on paper and forms the sentence.	t9	Material entity	*I conclude therefore that Mr. Jones has type 2 diabetes mellitus.*
IUI-10	IQE that inheres in IUI-9.	t10	Information quality entity	The sentence began to exist as soon as ink was laid down on paper, but the IQE did not begin to exist until the sentence was finished.
IUI-11	Dr. Smith's interpretive process	occupies t11	Diagnostic process	Dr. Smith's diagnostic process that led to her diagnosis IUI-8
IUI-12	The clinical picture input into IUI-11	t12	Clinical picture	Dr. Smith's clinical picture as ascertained prior to t6
IUI-13	Dr. Smith writing her diagnosis in the note	occupies t13	Process	

Table 5. Additional temporal entities in Scenario 1.

Temporal identifier	Description	Notes
t14	The interval during which the anatomical entity (IUI-5) is part of the cognitive system (IUI-4)	This interval is not easily specified given the current state of neuroscience. It could be different than t3 and t4.
t15	The interval during which the clinical picture (IUI-12) is used in the interpretive process (IUI-11)	Could be shorter than t11
t16	The point in time at which the cognitive representation (IUI-6) and diagnosis (IUI-8) begin to exist	t16 ends t11. Because the ICE does not exist until the cognitive representation—its first concretization—exists, this is also the point in time at which the diagnosis begins to exist.
t17	The interval during which the cognitive representation (IUI-6) participates in the writing process (IUI-13)	
t18	The interval during which the diagnosis (IUI-8) participates in the writing process (IUI-13)	It is possible that the original cognitive representation (IUI-6) gets copied elsewhere in the brain for reasoning and thus that the ICE continues to participate after the initial cognitive representation
t19	The interval during which that which is written on paper (IUI-10) begins to exist until it exists in full	The writing process begins earlier than the time at which the sentence begins to exist: the author starts the process with getting a pen and paper, any preparation necessary ("clicking" the pen), etc.

Table 6. Relationships among particulars in Scenario 1.

IUI	Relation	IUI	When relation holds in reality	Notes
IUI-4	**part of**	IUI-3	at t4	
IUI-5	**part of**	IUI-4	at t14	All anatomical components in which the cognitive representation inheres are part of the cognitive system. We do not assume the cognitive system is limited to the brain, as the state of neuroscience does not permit such an assumption.
IUI-6	**inheres in**	IUI-5	at t6	
IUI-6	**is about**	IUI-7	at t6	The cognitive representation stands in aboutness to IUI-7 as long as it exists
IUI-6	**is about**	IUI-1	at t6	It is also about Mr. Jones
IUI-6	**is about**	IUI-2	at t6	And about Mr. Jones' disease
IUI-6	**is about**	UUI-1	at t6	And about Type 2 diabetes mellitus
IUI-6	**concretizes**	IUI-8	at t6	It also concretizes the diagnosis
IUI-10	**inheres in**	IUI-9	at t9	The IQE inheres in the sentence on paper
IUI-10	**is about**	IUI-7	at t10	The IQE stands in aboutness to IUI-7
IUI-10	**is about**	IUI-1	at t10	It is also about Mr. Jones
IUI-10	**is about**	IUI-2	at t10	And about Mr. Jones' disease
IUI-10	**is about**	UUI-1	at t10	And about Type 2 diabetes mellitus
IUI-10	**concretizes**	IUI-8	at t10	
IUI-10	**is conformant to**	IUI-6	at t10	Is conformant to the cognitive representation as long as it exists
IUI-3	**agent in**	IUI-11	at t11	
IUI-12	**input into**	IUI-11	at t15	Clinical picture input into IUI-11
IUI-6	**output of**	IUI-11	at t16	Cognitive representation output from IUI-11
IUI-8	**output of**	IUI-11	at t16	Both the diagnosis and its concretization are outputs of IUI-11
IUI-8	**input into**	IUI-13	at t17	The diagnosis is input into writing
IUI-6	**input into**	IUI-13	at t18	As is its cognitive representation
IUI-10	**output of**	IUI-13	at t19	The sentence is output of writing

Table 7. Relationships of representations to portions of reality in Scenario 3: *Incorrect diagnosis*.

Relationships among particulars				**Notes**
IUI-46	**is about**	IUI-1	at t46	Dr. Jane Miller's cognitive representation is about Mr. Jones
IUI-46	**is about**	IUI-2	at t46	And Mr. Jones' disease
IUI-46	**is about**	UUI-2	at t46	And Type 1 diabetes mellitus (denoted by UUI-2)
IUI-50	**is about**	IUI-1	at t50	Likewise with the IQE inhering in the ink on paper
IUI-50	**is about**	IUI-2	at t50	
IUI-50	**is about**	UUI-2	at t50	
IUI-46	**is misrepresentation of**	IUI-7	at t46	But the cognitive representation is a misrepresentation of the configuration, i.e., it is intended to be about the configuration but fails at the level of compound expression
IUI-50	**is misrepresentation of**	IUI-7	at t50	The same is true of the IQE

Table 8. Six possibilities for a diagnosis failing in aboutness at the level of compound expressions.

Problem	**Where it fails *first***	**Description**
Noninstantation, asserted type exists	Level of compound expression	Disease instantiates a different type than the stated type, but the stated type exists
Noninstantation, asserted type does not exist	Level of reference	Disease instantiates a different type than stated, while the stated type of disease does not exist
Disease nonexistence	Level of reference	The disease instance does not exist
Organism nonexistence	Level of reference	The organism instance does not exist. In this case, there could not be a clinical picture properly inferred and thus it is not a misdiagnosis although it could still be an ICE.
Disease non-inherence	Level of compound expression	The disease inheres in a different organism than the one stated. For example, the doctor mistakenly ascribes Mr. Johnson's hypertension to his twin.
Configuration is not located in that part of spacetime where the diagnosis says it is located.	Level of compound expression	A diagnosis of type 2 diabetes mellitus 5 years ago is wrong because the patient didn't have the disease at that time, even though the patient has type 2 diabetes today. Also, a diagnosis that the patient has an upper respiratory tract infection today when in reality the infection resolved two weeks ago.

Table 9. Additional tuples required to distinguish diagnosing from a lucky guess.

IUI	Entity	Lifetime	Type	Notes
IUI-14	The aggregate of Dr. Smith's cognitive representations of various disease types and their associated types of phenotypes including type 2 diabetes mellitus that he used in the diagnostic process	t20	Aggregate of cognitive representations	
Relationships among particulars				**Notes**
IUI-14	**input into**	IUI-11	at t21	t21 refers to the temporal interval during which IUI-14 participated in the reasoning process. It could start at the same time as t11 or after t11, and end at the same time as or before t11.

Figure legends

Figure 1. The configuration of Mr. Jones, his disease, and type 2 diabetes mellitus

Figure 2. Diagram of diagnostic process, its inputs, a correct diagnosis, its concretization, and the configuration that that the concretization is about

Figure 3. Misdiagnosis of type of disease. The diagnosis is individually about the patient, the disease, and the incorrectly diagnosed disease type Y, but it is not about the configuration of patient, disease, and disease type X.

Figure 1

Click here to download Figure Fig 1.png

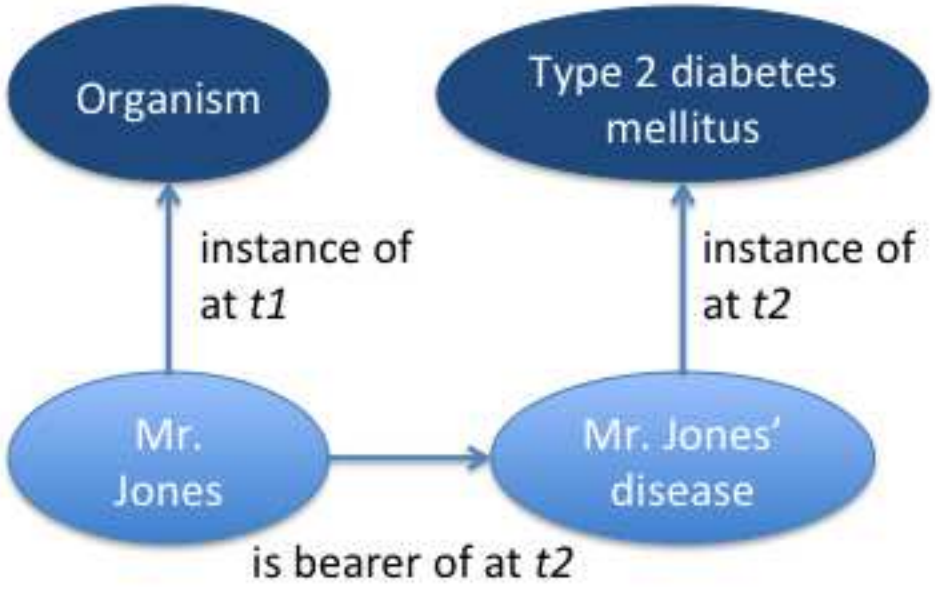


Figure 2

Click here to download Figure Fig 2.png

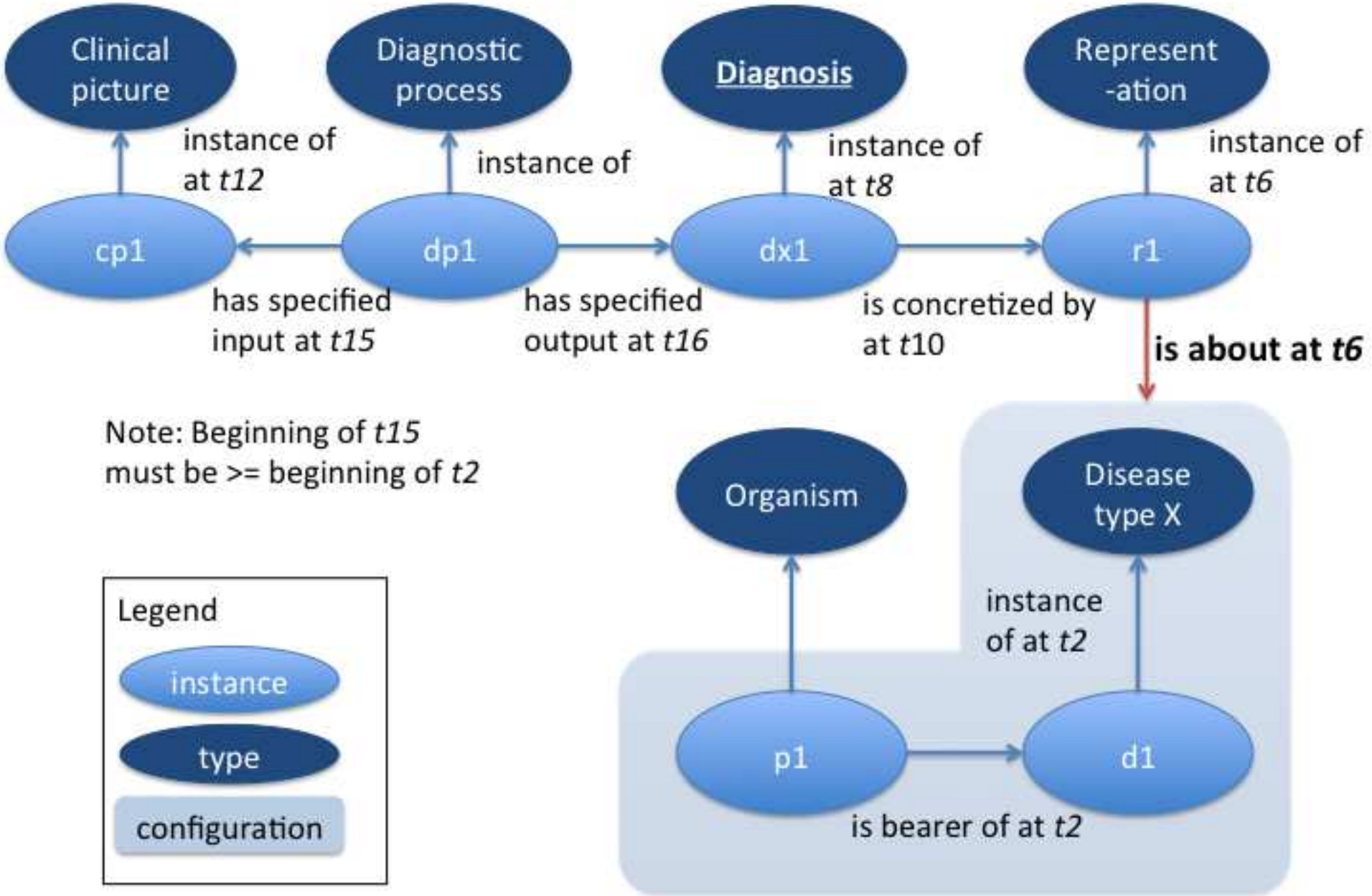
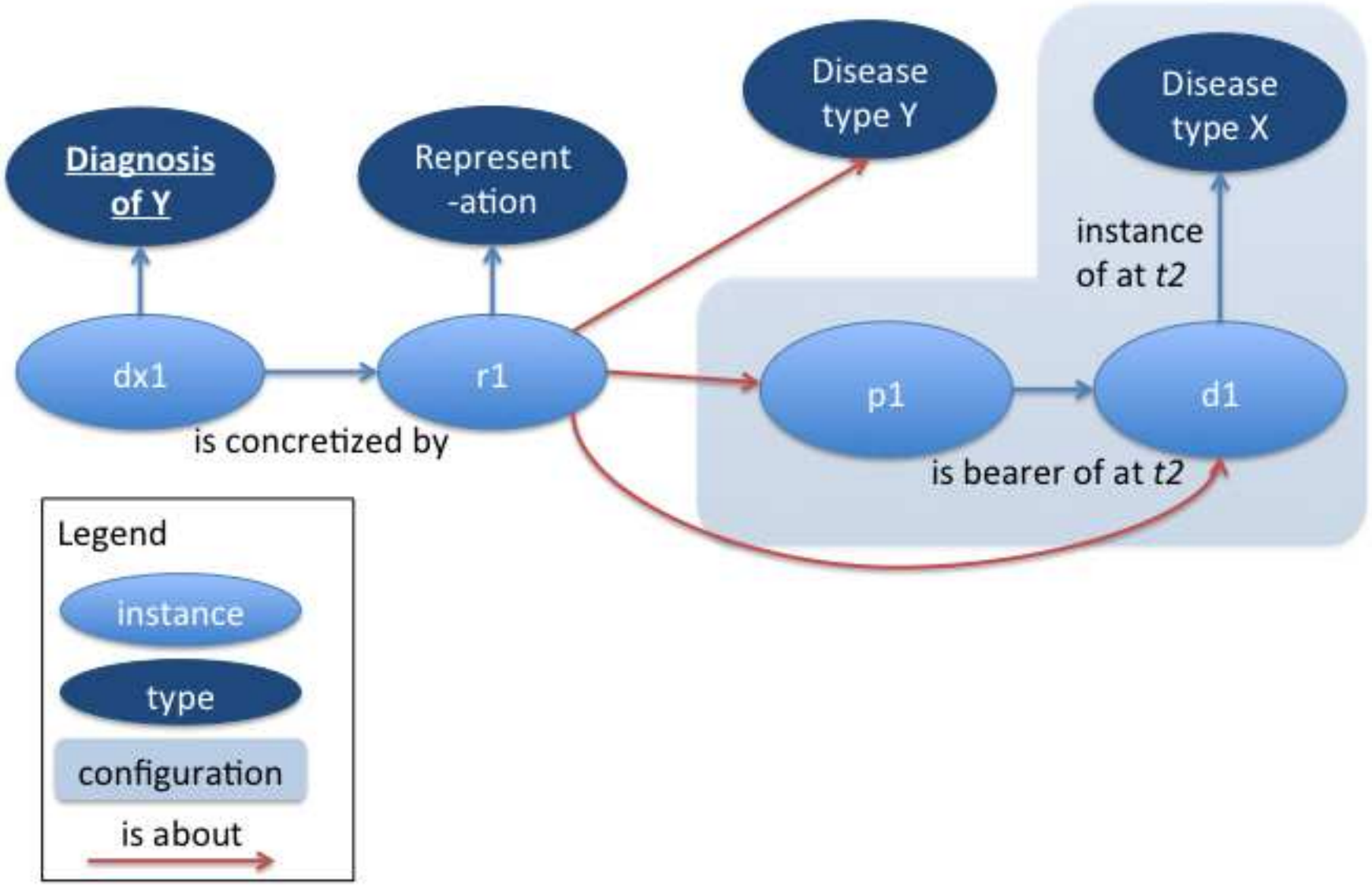


Figure 3

Click here to download Figure Fig 3.png



Additional file 1

Additional file 2

Additional file 3

Additional file 4

Journal of Biomedical Semantics

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis.

--Manuscript Draft--

Manuscript Number:	JBSM-D-15-00018R1
Full Title:	Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis.
Article Type:	Research
Funding Information:	National Center for Advancing Translational Sciences (UL1TR001427) — Not applicable Patient-Centered Outcomes Research Institute (CDRN-1501-26692) — Dr. William R. Hogan
Abstract:	Background Disease and diagnosis have been the subject of much ontological inquiry. However, the insights gained therein have not yet been applied to the study, management, and improvement of data error in electronic health records (EHR) and administrative systems, whose data suffer from workarounds due to limitations in the current state-of-the art in system design. These limitations include difficulties accounting for misdiagnosis, the various types of entities that diagnoses as information content entities can be and are about, and the status of coincidentally correct statements about disease. Methods We applied recent advances in the ontological understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created six scenarios that we analyzed using the method of Referent Tracking, to identify all the entities and their relationships which must be present for each scenario to hold true. We discovered deficiencies in existing ontological definitions and proposed revisions of them to account for the improved understanding that resulted from our analysis. Results Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are typically about a configuration in which there exists a disease that inheres in an organism and instantiates a certain type (e.g., hypertension). Misdiagnoses are ICEs whose concretizations succeed in aboutness on the level of reference for individual entities and types (the organism and the disease), but fail in aboutness on the level of compound expression (i.e., there is no configuration that corresponds in total with what is asserted). Provenance of diagnoses as concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief. Conclusions Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR and other systems that allow diagnostic assertions to be captured with less ambiguity.
Corresponding Author:	William R. Hogan, MD, MS University of Florida Gainesville, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Florida
Corresponding Author's Secondary Institution:	
First Author:	William R. Hogan, MD, MS

First Author Secondary Information:	
Order of Authors:	William R. Hogan, MD, MS
	Werner Ceusters, MD
Order of Authors Secondary Information:	
Response to Reviewers:	Dietrich Rebholz-Schumann Journal of Biomedical Semantics March 15, 2016 Dear Dr. Rebholz-Schumann, We greatly appreciate the reviewers' insightful feedback and the opportunity to revise the manuscript accordingly. We have significantly revised the manuscript and clarified a number of confusing and ambiguous points. We have also cited some of the literature on the diagnostic process recommended to us by reviewer #2. The revised manuscript also illustrates the use of the existential vs. universal quantifier, and how this is enabled by our analysis, with specific examples. Below, we address each comment of the reviewers individually. We very much look forward to the outcome of further reviews of and decisions about our manuscript. Sincerely, William Hogan and Werner Ceusters In the below, paragraphs written by us in response to the reviewers begin with an arrow as such ->. All other paragraphs are written by the reviewers. Reviewer reports: Reviewer #1: The manuscript "Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis" describes six scenarios in which a diagnosis is involved, and their ontological analysis. The analysis mainly uses the framework of referent tracking, developed by (one of) the authors, and consists of creating identifiers for individuals and concepts. The point of the six scenarios is mainly to demonstrate that referent tracking is able to describe these scenarios. Major comments: The manuscripts lacks appropriate context or any biomedical applications, -> We now include in background the increasing exchange of health data (eg institutes sharing data for the advance of medical science), patients' willingness to share data, and the dangers of EHR secondary use research if data are inaccurate, erroneous, etc. -> We also now discuss need for better data provenance -> Thus better, unambiguous representation of diagnostic assertions is required. Hopefully this context meets the reviewers requirements. it is primarily a lengthy discussion on how to represent six scenarios in a database using the referent tracking approach. Six of the ten references given in the paper are to previous work of the authors(!), -> This shows the key feature of novelty. Did the reviewer find references on referent tracking NOT by the authors? one irrelevant reference to a diagnosis of malaria in Tutankhamun, and only three to related work (two to the work of Schulz et al., one to Hastings et al.). -> point being? Does the reviewer not like Schulz? Also, see our response on the next point about paucity of related work. -> Also, the reference to the diagnosis of malaria in Tut proves the claim that a disease

can be diagnosed long after it ceases to exist. Although some individuals are not as concerned with time as we are and thus don't take extraordinary pains to keep detailed track of times of disease and diagnosis, it is a key feature of referent tracking to be extremely precise with the times of diagnostic statements, the times at which diseases began to exist, the times at which diseases ceased to exist, the time intervals of diagnostic processes, the beginning and ending boundaries of those intervals of diagnostics processes, and so on. We also note that this reviewer has asked in other comments that we avoid making unsupported claims. It is not prima facie "obvious" that it is possible to diagnose a disease after it has ceased to exist to those who are not familiar with medical diagnosis, autopsy, etc. and even then such distant diagnoses are likely to be somewhat surprising to astute readers.

There is no discussion in the manuscript of related work
-> related work has not been done, to our knowledge.
-> What has been done: study of adequate biomedical terminology to express what a diagnosis is about, how to come to a diagnosis, but not how to formulate a diagnosis, nor the implicitness of stating just a diagnostic code in the problem list.
-> Querying for "structure of diagnosis", "diagnostic expressions", etc. gives no similar work, but rather returns results that exemplify the general confusion between a diagnosis as an assertion and the diagnostic process.

or why the work presented would be important, or even how it could be applied; there is also no demonstration that the specific results presented in the paper (description of the six sitations) would be applicable in a biomedical context.
-> Here is the answer, which we now discuss (and even quote) in the extensively rewritten Background: "Researchers who do not consider data provenance risk compiling data that are systematically incomplete or incorrect. For example, researchers who are not familiar with the clinical workflow under which data were entered might miss or misunderstand patient information or procedure and diagnostic codes."

How the provenance of electronic health record data matters for research: a case example using system mapping. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ.
EGEMS (Wash DC). 2014 Apr 16;2(1):1058. doi: 10.13063/2327-9214.1058. eCollection 2014.

-> Also, we note that neither Einstein's original paper on special relativity, nor Heisenberg's paper on quantum mechanics, nor Shannon's paper on the theory of information, did or could possibly have fruitfully explored the entire space of applications that has subsequently been realized (even the more immediate ones). We therefore disagree that it is an absolute requirement of a scientific paper to have immediate practical benefit, beyond providing an improved explanation of phenomena in reality. These papers did explore the implications for theories that proceeded them (e.g., Einstein noted that the theory of a luminiferous ether was now dead), but beyond that nothing. Heisenberg did not foresee the computer, for example. Should he have? Our paper, although much more modest in its scope (and we are certainly not so vain as to think the current manuscript does or will stand next to the ones mentioned, we merely use them as some of the strongest possible examples that one could use), does exactly what these papers did: develop an improved account and explanation of phenomena in reality (in our case, of diagnosis as an ICE). That makes it a perfectly valid and useful scientific paper, even if there's no obvious "application". However, we have better placed it into context as to why we are doing this work now.

2. A main result or conclusion from the work presented is that "[i]t is perfectly legitimate to define 'information entity about a clinical situation' with the existential quantifier". But nowhere in the manuscript is this actually done or demonstrated.
-> We now demonstrate this in the Conclusion section. We also have greatly improved our explanation of the two levels of aboutness that motivates it, beginning with the rewritten Background section.

Instead, the authors merely state that "if an instance of this type is not completely correct, it is still an ICE that is individually about the patient and any types of disease, disorder, or other entity referenced by it". While this may be true in the specific

framework the authors use here, it misses the point about the use of the universal quantifier entirely, because it is explicitly one to a concept that has as instances the configurations that may or may not exist.
-> We did not sufficiently explain what we meant by the term ‘completely correct’, a problem we have fixed beginning with the extensively rewritten Background. The point is that prior authors would say an incorrect sentence like “Obama is President of Russia” is either not an ICE or is not about anything at all. That is why they use universal quantification, because it allows for the latter possibility. Instead the Obama sentence is about Obama, Russia, and his President role. It is not about nothing merely because it is wrong on the level of compound expression.

And if "an instance of this type is not completely correct" then it is just not an instance of this type, and therefore an existential quantifier to instances of the type would be incorrect.
-> Again, the issue is that we failed to explain what we mean by ‘completely correct’. We have elaborated and rephrased the text. We meant by not ‘completely correct’ correctness on the level of compound expression: The Obama is president of Russia case. It is now better explained in the paper.

3. The other main results, according to the authors, are that a correct diagnosis "is about a configuration of an organism, its disease, and the type the disease instantiates", and that "[t]he most common forms of misdiagnosis are information content entities whose concretizations succeed in aboutness at the level of reference, individually referring to an organism, disease, and/or type of disease, but fail in aboutness at the level of compound expression". For the latter, no evidence is given besides speculation.
-> We agree, and in fact when we went to search the literature on the problem, we could find no data about the different kinds of misdiagnosis that we identified. The fact that we could find no study of the incidence of diagnostic error that divided misdiagnoses into these categories ultimately supports our claim that better ontological understanding of diagnosis has the potential to benefit research into misdiagnoses.

With regard to the former, this may be a useful result if it was evidenced with an application or use case that demonstrates how this is actually applied (outside some informal discussion of six hypothetical scenarios) in an ontology such as the OGMS.
-> we presented a formal discussion. We also proposed improved definitions for OGMS. Lastly, we think the term misdiagnosis as we have defined it should be included in OGMS.
-> Also, see our prior discussion about what constitutes a valid scientific paper. Immediate application or practical utility is not an absolute requirement.

Finally, a minor result is that provenance is important and "lucky guesses, hearsay, and laypersons' conclusions about disease (when not arrived at through a diagnostic process) do not constitute diagnoses and therefore are different types of ICE than diagnoses." While I completely agree on the importance of provenance, the authors do not provide information on how to distinguish diagnosis processes from non-diagnosis processes.
-> We state that diagnostic process requires a clinical picture and medical knowledge as input, and cite reviews of the extensive literature on diagnostic process as support for the latter. Non-diagnostic processes would lack either as input. Perhaps the reviewer is confusing diagnostic process with diagnosis? Not again that the object of our study was not diagnostic processes. As the reviewer notes, our addition to the definition of ‘diagnostic process’ was a secondary result (and also we derive it from the literature on that process, which we now cite). We also added text in Conclusion that clarifies this point, in response to reviewer 2’s comments. Also we did comment on the literature, including on the diagnostic process.

Reviewer #2: In the present work, the authors provide an ontological theory of medical diagnosis. In their work they focus on diagnosis as encountered in electronic health records. They also cover associated entities as misdiagnosis and lucky guess as related to medical diagnosis.
-> Correct, but qualified by the fact that our primary object of analysis was diagnosis as an information content entity, not diagnosis as a process.

	As a key process diagnostic reasoning and diagnoses as its results, the topic of the paper is of high importance. The authors deduce the objective of their work from prior ontological work and what is missing in it. The paper is well structured and written. The authors provide a sound ontological theory based on the methodology of reference tracking and prior ontological models. As ontological model, it must prove itself in its use and application. However, the reviewer has the following concerns with the representation of „diagnostic process" (diagnostic reasoning at least when observed in a humans) and „diagnosis": For the reviewer, the perspective on diagnosis and diagnostic processes originating mainly from electronic health records seems heavily biased. For over three decades, cognitive scientists and medical educators have gained insights in phenomena around clinical and especially diagnostic reasoning. There is a large body of evidence on how medical expertise can be observed and measured and described in terms of cognitive models. As a realistic theory on diagnosis and processes leading to diagnosis, in the reviewers view, this insight should be included in the ontological work and can not be completely neglected; it will be scientists from the domain of diagnostic reasoning which can provide insights in their domain which should be included in ontological theory (they will be the scientists with the clearest view on the ‚reality' of diagnostic reasoning and what a diagnosis in terms of diagnostic reasoning might be). Especially, when it comes to different types of reasoning on diagnosis: the higher the level of expertise a reasoning physician has, the *less* is the amount of deductive reasoning (evidence is abundant). In addition, medical and healthcare professional provide a large body of educational, evidence based and regulatory material on how diagnosis must be derived and how the diagnostic processes have to be structured and conducted. The authors seem not to be interested to relate their work with what has been provided in this regard by the healthcare professions. They fall much to short with their ontological theory ignoring foundational scientific and professional literature in the domain. The reviewer might provide a comparison: e.g., ontological work on anatomy will be based on anatomical knowledge (FMA) otherwise it will be ignored. So this work should have a focus on diagnoses and diagnostic process as described by the corresponding domain experts and scientists. -> We agree that further review of the literature on diagnosis and further study of the diagnostic process from ontological process has high potential to be very valuable. However, we think that what the reviewer suggests is the further categorization and development of what in OGMS is ‘diagnostic process’. That is NOT the topic of the paper, however. Although we have added text that makes reference to this body of work. Our topic is on ‘diagnosis’ as defined in OGMS. We added text that elaborates on this point in the paper, especially in the discussion. Directly associated with the issue above, the authors use problematic definitions of lay persons and experts in their representations of lucky guess and hearsay. From current knowledge in medical education, the notion of expert is only applicable in a certain domain. Competencies and expertise are acquired over long periods and change throughout the complete professional lifetime. Levels of expertise - if of ontological importance - must be carefully defined, otherwise this is not useful. Somebody is only an expert in the domain of his expertise (this means expertise is not domain indepenent), this should be clearly ontologically defined. -> Again true, but not the topic of the paper. We didn’t define competence, lay person, etc. because as we discussed in the paper, it is not that a lay person cannot make a diagnosis, even if the diagnostic process is totally nonsensical: if she uses one, and the inputs are a clinical picture and mental representations of relevant disease types and phenotypes, and there is a conclusion, then the conclusion is a diagnosis. However, we do now make reference to the literature, and its findings about how mental representations are used as input into diagnostic processes and evolve as expertise develops. Most physicians are „very careful" when diagnosing in areas where they might be not competent. From empirical work it is also known, that many diagnosis even in the domain of the diagnosing person are based on wrong assumptions, ill defined knowledge or information, or wrong inference. Is this covered in the current model? -> That would fit into an improved model of the diagnostic process. A typology of diagnostic processes may result in a typology of diagnoses. It is future work, and we

	now say so in the manuscript. We think working with experts in this field would be very important and rewarding. From the limited experience of the reviewer, it might not be correct to base the definition of diagnostic process on the existence of a clinical picture. There are very abstract diagnostic processes in action: in imaging, in pathology, and in laboratory work. Diagnosing works on laboratory material and specimen only, in the reviewers view, to arrive at a complete diagnosis a clinical picture is not necessary. Same with images, to diagnose a bone fracture or a lung metastasis - no clinical picture as input is necessary. The result of initial assessment can be the diagnosis. -> Everything the reviewer states here is in the definition of clinical picture, which we clarify significantly in the revised manuscript. We agree that the name is not intuitive, and say so in the manuscript. We added text that elaborates more on each of the definitions. Another issue with the provided theory is the foundational role of intention in the definition of misdiagnosis. To operationally apply this ontological definition a realistic definition of intention should also be provided. In the view of the reviewer, the definition of misdiagnosis in the current representation is problematic. -> We explain intention better in the Methods section with citations. We also refer the reader to Smith and Ceusters 2015 for more information. Page 18 line 27: if -> fixed Page 20 line 57: only one „that" -> fixed Literature: Custers, Eugène J. F. M. 2014. Thirty years of illness scripts: Theoretical origins and practical applications. Medical Teacher (2. September): 1-6. doi:10.3109/0142159X.2014.956052, . Goldszmidt, Mark, John Paul Minda und Georges Bordage. 2013. Developing a Unified List of Physicians' Reasoning Tasks During Clinical Encounters. Academic Medicine 88, Nr. 3 (März): 390-394. doi:10.1097/ACM.0b013e31827fc58d, . Norman, Geoffrey. 2005. Research in clinical reasoning: past history and current trends. Medical Education 39, Nr. 4 (April): 418-427. doi:10.1111/j.1365-2929.2005.02127.x, . Norman, Geoff, Meredith Young und Lee Brooks. 2007. Non-analytical models of clinical reasoning: the role of experience. Medical Education 41, Nr. 12: 1140-1145. doi:10.1111/j.1365-2923.2007.02914.x, . Schmidt, Henk G. und Remy M. J. P. Rikers. 2007. How expertise develops in medicine: knowledge encapsulation and illness script formation. Medical Education 41, Nr. 12: 1133-1139. doi:10.1111/j.1365-2923.2007.02915.x, . Sherbino, Jonathan, Geoffrey R. Norman und Wolfgang Gaissmaier. 2013. Clinical Decision Making. Academic Medicine 88, Nr. 2 (Februar): 150-151. doi:10.1097/ACM.0b013e31827b2941, . -> We thank this reviewer for his kindness in pointing us to this body of work and recommending it to us. We now cite two of the above six papers in the manuscript, as evidence that knowledge structures, or as we refer to them “cognitive representations of types”, are indeed input into the diagnostic process.

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: An ontological analysis.

William R. Hogan (corresponding)
University of Florida
P.O. Box 100219
2004 Mowry Rd
Gainesville, FL 32610-0219
hoganwr@ufl.edu
(352) 294-4197

Werner Ceusters
University at Buffalo
921 Main Street
Buffalo, NY 14203
ceusters@buffalo.edu

Abstract

Background

Disease and diagnosis have been the subject of much ontological inquiry. However, the insights gained therein have not yet been applied to the study, management, and improvement of data error in electronic health records (EHR) and administrative systems, whose data suffer from workarounds due to limitations in the current state-of-the art in system design. These limitations include difficulties accounting for misdiagnosis, the various types of entities that diagnoses as information content entities can be and are about, and the status of coincidentally correct statements about disease.

Methods

We applied recent advances in the ontological understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created six scenarios that we analyzed using the method of Referent Tracking, to identify all the entities and their relationships which must be present for each scenario to hold true. We discovered deficiencies in existing ontological definitions and proposed revisions of them to account for the improved understanding that resulted from our analysis.

Results

Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are typically about a configuration in which there exists a disease that inheres in an organism and instantiates a certain type (e.g., hypertension). Misdiagnoses

are ICEs whose concretizations succeed in aboutness on the level of reference for individual entities and types (the organism and the disease), but fail in aboutness on the level of compound expression (i.e., there is no configuration that corresponds in total with what is asserted). Provenance of diagnoses as concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief.

Conclusions

Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR and other systems that allow diagnostic assertions to be captured with less ambiguity.

Keywords

Biomedical ontology

Disease

Diagnosis

Information content entity

Representation

Ontological realism

Background

As administrative, clinical, and patient-reported data are increasingly shared and reused, especially for patient care [1-4] and research [1, 5-7], several issues with these data—including diagnosis data—are of increasing concern. The issue that appears to be of greatest concern for administrative and clinical data is error in the data and the implications of that error for making decisions and conclusions based on them [8-13]. Although Shapiro et al., in a report for the Office of the National Coordinator for Health Information Technology, do not cite error as a concern for including patient-generated health data into the electronic health record (EHR) [14], there are known errors with patient self reporting especially in research [15-22]. A second issue of concern is data provenance [10, 23], i.e. information about who created the data, in what setting, how, when, for what purpose, and so on. For example, Johnson et al. noted that the provenance of symptom data was essential to using those data correctly to determine whether a colonoscopy was a screening vs. diagnostic procedure [23].

Data error and data provenance are closely related. For example, Hersh et al. note that data recorded in billing workflows for financial purposes are less accurate than clinical data [10]. Thus, timing, method, and purpose of recording data at a minimum—all aspects of provenance—are intertwined with accuracy. Furthermore, a key result of the Johnson et al. study is that "Researchers who do not consider data provenance risk compiling data that are systematically incomplete or incorrect" [23].

An ontological account of data error and data provenance can identify crucial distinctions. For example, there are significant differences among (1) a measured weight that is off because the scale was not properly tared, (2) a 'rough' weight of 70kg entered in an emergency when the patient cannot be weighed, and (3) a weight measurement entered on the wrong patient. Detecting and accounting for these differences and their causes—especially the aspects of provenance that influence them—is necessary to inform strategies to study, cope with, and improve data error when using pre-existing EHR data for research.

In this work, we apply Smith and Ceusters' recent ontological account of incorrect information [24] to diagnosis data in administrative systems, EHRs, and patient-reported information. Their account holds that a statement such as a diagnostic assertion can succeed or fail in aboutness on least two levels: (1) the level of denotating single entities and/or types (i.e., the level of *reference*) and (2) the level of veridical representation of a configuration of multiple entities and/or types (i.e., the level of *compound expression*).

To succeed on the second level (compound expression), the information content entity (ICE) must be correct about *all* particulars, their relationships, and their instantiations of types that it mentions. Failure on a single particular, relation, or instantiation causes the ICE to fail at the second level while still potentially succeeding at the first level. For example, if Mrs. Jones has type 1 diabetes mellitus, then the sentence '*Mrs. Jones suffers from type 2 diabetes mellitus'* fails in aboutness on the level of compound expression because it misstates one thing: her disease does not instantiate type 2 diabetes mellitus. However, despite this failure the sentence is nevertheless still about Mrs. Jones, her

disease, and type 2 diabetes mellitus on the level of reference, because indeed it mentions those three entities. It is therefore, per Smith and Ceusters, an ICE that is about *something* even though it is a misdiagnosis.

We note that prior ontological work on the aboutness of clinical statements like diagnoses has been constrained by the view that an ICE is about nothing (or is perhaps not even an ICE at all) if it fails on the level of compound expression. Martínez Costa and Schulz, for example, use the universal quantifier when defining 'information entity about a clinical situation' *…to avoid asserting the existence of an entity the existence of which cannot be guaranteed* [25]. For an ICE such as 'suspected heart failure' they want to avoid the implication that there is some instance of heart failure that it is about. Because they cannot guarantee the existence of some heart failure, they use universal quantification to say 'if it is about anything, it is about an instance heart failure'. Researchers working in areas other than diagnosis have encountered similar issues. For example, Hastings et al. note that chemical graphs and diagrams are not always about types of molecules that exist [26]. They, too, used the workaround of replacing existential quantification with universal quantification to avoid asserting that every chemical graph/diagram is about some type of molecule that exists (level of compound expression).

In our own, previous ontological analysis of diagnosis, using the methodology of referent tracking, we identified what entities must exist or must have existed for a particular diagnostic statement to hold true [27, 28]. A key result of this work is that a diagnosis is minimally about *both* the patient and the type of disease asserted to exist. In addition,

building on previous work on the Ontology for General Medical Science (OGMS), the foundations of which were laid down in Scheuermann et al. [29], we noted that for a diagnosis to exist (at least in medicine and under the assumption that the diagnosis was made *lege artis*), there must also have existed a diagnostic process, a person who carried out that process, and a clinical picture which was used as input into that process.

The hypothesis for the work described here was that applying Smith and Ceusters' results to disease and diagnosis, in combination with prior work on the ontology of disease and diagnosis (including provenance of the latter), could address limitations encountered in previous ontological work on disease and diagnosis and improve our representations of them in support of studying, coping with, and reducing diagnostic error.

Methods

To test this hypothesis, we analyzed a set of scenarios that we created and that involve correct and incorrect diagnoses, lucky guesses, and justified layperson belief in the existence of a disease of a certain type. The goal was to explore whether, and if so how, a realism-based account of information can deal successfully not only with diagnostic statements asserting the ideal case of a correct diagnosis, but also with deviations from the ideal.

Materials

In our analysis we used as input (1) Smith and Ceusters' work on aboutness and their definitions of representation, mental quality, cognitive representation, and information

quality entity (Table 1), (2) definitions of disease, disorder, and diagnosis from the Ontology for General Medical Science (Table 2), and (3) our prior work on analysis of diagnostic statements [27, 28].

Smith and Ceusters stressed that the relation of aboutness includes any portion of reality, rather than being limited to just a single particular or a single universal. A portion of reality (POR) can be a particular, a universal, a relation, or a configuration. A configuration is a combination of particulars and/or universals and certain relation(s) that hold among them.

A representation, then, that is intended to be about a POR but fails in its aboutness because it misrepresents that POR in some way, is misinformation. The sentence *Bob Dylan was in the Beatles* fails to represent not because Bob Dylan or the Beatles did not exist, but because such a configuration involving Bob Dylan and the Beatles in the way as expressed, never existed. The sentence fails in aboutness on the level of compound expression, but nevertheless is about Bob Dylan and the Beatles individually (on the level of reference) and thus is still an information content entity.

Note that Smith and Ceusters [24] deal more fully with the issue of what it means that a representation is "intended to be about" some entity. Here, we highlight that it follows the doctrine of the "primacy of the intentional" [30], where our written and verbal expressions are to be understood on the basis of the cognitive acts that generated them. That is, a

sentence is about that to which its author was directing his or her thoughts when she wrote it.

In addition to Smith and Ceusters' work, we also founded our ontological analysis on the Ontology for General Medical Science or OGMS [29]. This work distinguishes disease, disorder, and diagnosis, and we used definitions from OGMS as starting points for our analysis (Table 2). Note that in OGMS, a diagnosis refers to the existence of a disease of a given type. In clinical medicine, however, diagnoses also refer to (1) disease courses (e.g., acute hepatitis vs. chronic hepatitis), (2) disorders (e.g., fractures and tumors), and (3) the absence of any disease (i.e., a conclusion that a person is healthy also is a diagnosis). It was not our goal to address this issue in this work, as it was not our goal to refine the OGMS definition of diagnosis.

The scenarios

All the scenarios have in common a particular patient, Mr. Adam Jones, who suffers from type 2 diabetes mellitus. Thus in every scenario, there exists Mr. Jones, his disease, the type *Type 2 diabetes mellitus*, the configuration of these three entities (which includes the "bearer of" and "instance of" relationships), and the placement in space and time of this configuration (Figure 1).

Scenario 1: correct diagnosis by physician (ideal case)

Dr. Anne Smith sees Mr. Jones in the office. She takes a history and physical, performs certain laboratory testing, and based on her analysis of the findings, correctly concludes

that Mr. Jones has type 2 diabetes mellitus. She subsequently writes her diagnosis in the patient's medical record.

Scenario 2: subsequent correct diagnosis by physician using first diagnosis

A second doctor, Dr. John Brown, sees Mr. Jones in the office at some later date. Mr. Jones has released his records from Dr. Smith to Dr. Brown, who subsequently sees Dr. Smith's diagnosis prior to seeing Mr. Jones. He uses that diagnosis plus his own findings to infer a new clinical picture of Mr. Jones, which he subsequently uses to make another correct diagnosis of Mr. Jones' disease. He writes his diagnosis in Mr. Jones' medical record.

Scenario 3: incorrect diagnosis by physician

Mr. Jones is traveling on vacation, when he falls ill. He sees Dr. Jane Miller who does not have any of his past records available, and thus she is not aware of the previous diagnoses of Drs. Smith or Brown. She infers a new clinical picture of Mr. Jones, and based on it incorrectly concludes that Mr. Jones has *type 1 diabetes mellitus* (as opposed to type 2). She records a diagnosis of type 1 diabetes mellitus in her medical record for for Mr. Jones.

Scenario #4: coincidentally correct conclusion by layperson (lucky guess)

A friend of Mr. Jones is a "seer". Mr. Jones asks his friend what is in his future. Having no prior knowledge of Mr. Jones medical conditions, the "seer" concludes based on Mr. Jones' horoscope and the position of the moon that he has type 2 diabetes mellitus. He subsequently predicts that Mr. Jones will be hospitalized for his diabetes and miss his daughter's wedding.

Scenario #5: layperson's justifiable conclusion

Mr. Jones' daughter, upon learning of her father's type 2 diabetes mellitus, adds this information into her letter to her brother, writing "Dad has type 2 diabetes mellitus".

Scenario #6: correct diagnosis by computer-based expert system

A medical student is seeing Mr. Jones in the clinic. He performs a history and physical, and types his findings into a diagnostic expert system. The diagnostic expert system infers based on these findings that Mr. Jones has type 2 diabetes mellitus. The medical student writes this diagnosis in Mr. Jones' medical record.

The analysis

Our analysis follows the method of Referent Tracking, which we have found to be a stringent test of ontologies and their definitions [27]. This approach proceeds in three main steps. First, we systematically identify all the relevant particulars that must exist for the scenario to be true, regardless of whether the scenario explicitly mentions them or only implies their existence. We assign each particular an instance unique identifier (IUI), of the form 'IUI-n', where 'n' is any integer. Second, we identify for each particular the type it instantiates and the temporal interval during which it exists (and assign an identifier of the form *tn* to that interval). Lastly, we identify all the relationships that hold between the particulars as well as all relevant relations particulars have to universals other than instantiation, including situations where a particular lacks a given relation to any instance of a certain type (for example, a statement that a patient has had no cough in the last two

weeks means that the patient does not stand in the agent_of relation to any instance of the type *Coughing event*, indexed temporally to the two-week interval).

This approach identifies problems in ontologies and their definitions in two major ways. First, it identifies problems that occur when the scenario explicitly rules out the existence of a particular whose existence is implied by an ontological definition (and vice versa). Second, it helps identify exceptions to existing definitions and situations that should not fall under a definition but are erroneously captured by it. Definitions in ontologies can subsequently be adjusted to avoid the errors so identified.

Although our approach is to identify particulars implied by sentences in natural language, the ontological analysis of language and the mechanism(s) by which it makes implicit reference to certain entities is not the focus of this work. Therefore, we convert a sentence like "Mr. Jones has type 2 diabetes mellitus" to Referent Tracking Tuples (e.g., as in Tables 3 through 7) and it is these tuples in which inhere representations that are the objects of our analysis.

To simplify our analysis somewhat, we wrote the scenarios such that humans record diagnoses on paper. However, concretization of ICEs also occurs by pixels on monitors, binary switches in memory and processor chips, and magnetic fields on hard disks. But a detailed account of these concretizations and transformations among them is not central to our analysis of what is a diagnosis. Our analysis can be extended to these concretizations without modification.

Results and discussion

In each scenario, Mr. Jones (IUI-1) and his disease (IUI-2) exist, the latter inhering in the former (Table 3). Furthermore, his disease is an instance of the type 'type 2 diabetes mellitus' at any moment in time during which a diagnosis is formulated in any of the scenarios. Mr. Jones (IUI-1) exists through a certain period of time (*t1*) of which we do not know the exact beginning or end. We use temporal identifiers of the form '*t*n' to clearly distinguish such identifiers from IUIs: where IUIs are always intended to be globally and singularly unique, distinct temporal identifiers may denote a unique period of time which is also denoted by another temporal identifier. We also assign an identifier to the time interval during which his disease (IUI-2) exists (*t2*). Diseases usually begin to exist after the organism does, but in the case of congenital genetic diseases, the two intervals might be coextensive. Also, we assume that disease IUI-2 existed at the time of diagnosing, but we recognize that diagnosing a disease thousands of years after it existed is possible, such as in the case of archaeologists' recent diagnosis of Tutankhamun's malaria [31].

Note that the configuration of organism, disease, and disease type is anchored at a particular location in spacetime, as is the diagnosis. But note also that the diagnosis additionally has an implicit or explicit reference to the location of the configuration in spacetime. To be a correct diagnosis, this reference must also be correct (it has to refer to some part, not necessarily the entirety of spacetime, occupied by the configuration). Thus, for example, to say that Tutankhamun had malaria in 1000 C.E. or today is incorrect, as it would be to say that Mr. Jones had type 2 diabetes mellitus before his parents were born.

***Scenario 1: correct diagnosis*.**

In this scenario, numerous PORs in addition to Mr. Jones and his disease must exist and stand in certain relationships to each other (Tables 4-6). Before Dr. Smith (IUI-3) writes (IUI-13) her diagnosis (IUI-8), there is a cognitive representation (IUI-6) that is concretized in some anatomical part (IUI-5) of her cognitive system (IUI-4). Note that we follow Ceusters and Smith [32] in asserting that all anatomical entities in which cognitive representations inhere are part of a person's cognitive system (that is, any entity used in cognition, including the bearing of cognitive representations, are necessarily within a person's cognitive system) at least during the temporal interval that the cognitive representation exists. If, for example, it would be the case that some white blood cell flowing through some brain capillary would through some of its molecules take part in the concretization of a cognitive representation, then that white blood cell would be part of the cognitive system at least during the existence of that concretization. It would not anymore be part of the cognitive system once it continues its journey through the body without participating in thought formation. Additionally, Ceusters and Smith take the position (which we also follow) that the cognitive system is not necessarily strictly limited to the brain or even to the entire neurological system of a person: the current state-of-the-art of neuroscience is yet searching for answers to questions such as "what is it in which cognitive representations inhere?" but until it reaches such answers, we remain in our representations agnostic.

IUI-9 denotes the sentence Dr. Smith wrote, as it exists on the particular piece of paper she used to write it on: 'The patient has type 2 diabetes mellitus'. This written statement on paper (IUI-9) bears an information quality entity (IQE, IUI-10) that concretizes her diagnosis (IUI-8). The cognitive representation (IUI-6) and IQE (IUI-10) that concretize the diagnosis are both about the configuration (IUI-7) in Table 1 (the level of compound expression), as well as about Mr. Jones, Mr. Jones' disease, and the universal *Type 2 diabetes mellitus* individually (the level of reference). The cognitive representation (IUI-6) and the diagnosis (IUI-8) are the output of Dr. Smith's diagnostic process (IUI-11), which had as input Dr. Smith's clinical picture (IUI-12) of Mr. Jones. Because the cognitive representation and IQE concretize the same ICE, the latter is conformant to the former (see Table 1).

A correct diagnosis is thus fundamentally an information content entity that is concretized by a representation that stands in an is_about relation to the configuration of an organism, its disease, the relation of inherence between the disease and the organism, a type that the disease instantiates, and the instantiation relation of the disease to the type, all within a given portion of spacetime (Figure 2). Furthermore, diagnoses are additionally differentiated from other ICEs by the fact that they are generated by a diagnostic process that has a clinical picture as input. We expand further on what constitutes a clinical picture in the next scenario, *Scenario 2*, as well as revisit the diagnostic process briefly in *Scenario 4*, although it was not our objective in this work to develop a fuller account of this process.

Note that it is trivial to state that the particular disease inhering in the organism is an instance of *entity* or even *disease*. Thus, there is an expectation that a diagnosis be as precise (the most specific type) as possible and at a minimal level of granularity that is relevant to treat the patient appropriately and to provide a reasonable prognosis.

Scenario 2*: *second diagnosis.

The second physician, Dr. Brown, makes a second diagnosis at a later point in time, using the first diagnosis in addition to clinical and possibly other findings to infer a new clinical picture of Mr. Jones. With the exception of the configuration of Mr. Jones/his disease/type 2 diabetes mellitus (IUI-7), there is a one-to-one correspondence of PORs as in Scenario 1, numbered IUI-23 through IUI-33 (Additional file 1 : Tables S1-S3). That is, there is no IUI-27 because the configuration is the same POR across scenarios. Similarly, there is no IUI-21 or IUI-22 because Mr. Jones (IUI-1) and his disease (IUI-2) are the same entities.

In this scenario, Dr. Brown (IUI-23) makes a new diagnosis (IUI-28), concretized both by his cognitive representation (IUI-26) in some part (IUI-25) of his cognitive system (IUI-24) and by the IQE (IUI-30) inhering in the sentence in his note (IUI-29). Dr. Smith's previous diagnosis (IUI-8) can be viewed as either (*view1*) being in the aggregate of things that Dr. Brown uses to infer his clinical picture (IUI-32) that serves as input into his diagnostic process (IUI-31), or (*view2*) as something which serves as extra input—alongside his clinical picture—for the diagnostic process. The cognitive representation and the IQE are about the configuration (IUI-7) as well as Mr. Jones (IUI-1), his disease (IUI-2), and type 2 diabetes mellitus (UUI-1).

The current definition of 'clinical picture' in OGMS (see Table 2) seems to conflict with *view1* about this scenario, because the definition seems to exclude using a past diagnosis to infer a clinical picture. Although the current OGMS definition of 'clinical picture' is inclusive of clinical findings, diagnosis as currently defined is not an explicit subtype of clinical finding in OGMS. Furthermore, it is common for clinicians to elicit a previous provider's past diagnosis from the patient or the patient's caregiver during an interview (for example, if Mr. Jones in scenario #2 would have said: 'Dr. Smith says I have type 2 diabetes mellitus'). But the current OGMS definition of 'clinical history' (Table 2) conflicts with this possibility. It refers to health-relevant features of a patient, but features as elucidated by OGMS include only qualities, processes, and physical components of the organism—not dispositions of which disease is a subtype. Therefore, a representation of a disease such as a diagnosis is currently excluded from the OGMS definition of 'clinical history'.

We also note that the OGMS definition of 'clinical picture' is ambiguous in that it is not clear whether it *requires* that laboratory and image findings must always be used to infer a clinical picture, or that they are the only entities that can be used. Regardless, it would be a mistake to do so, because diagnoses can and frequently are made from symptom findings alone. Laboratory and image findings are not necessary components of a clinical picture in reality. Note that a clinical picture can comprise findings of a single type (laboratory alone, pathology image alone, radiology image alone, physical exam finding alone), or even a single finding instance (e.g. Reed-Sternberg cells for a diagnosis of Hodgkin's lymphoma).

All these issues are compounded by the fact that the term 'clinical picture' itself is not intuitive.

Given that clinical history taking elicits past diagnoses routinely in clinical medicine, we propose modifying the definition of 'clinical history' to accommodate this reality (bolded sections represent changes to the definition):

clinical history =def. – *A series of statements representing one or more health-relevant features of a patient**, possibly complemented by representations of diseases and configurations**.*

Note that the definition already allows—under the broader heading of 'feature'—representations of disorders (kinds of physical component) and disease courses (kinds of process). Thus, the definition already accommodates these aspects of clinical histories. We also allow the statements to represent configurations, in line with Smith and Ceusters [2]. These configurations might or might not include various relevant types (for example, "The patient has not participated in any instance of vomiting in the last two weeks."). Finally, note that by using the word 'representing', the definition also accommodates per Smith and Ceusters [2] that some statements might fail in aboutness despite their intention to be about such features. In other words, some statements in the clinical picture might be wrong: for example, a statement that the patient has a disease or pain that she does not in fact have.

To clarify that laboratory and imaging findings are not always required inputs into the diagnostic process, and to capture realistic scenarios compatible with *view2* (for example, Dr. Brown reads Dr. Smith's note in the chart), we also propose a modified definition of 'clinical picture' (changes in bold):

clinical picture =def. – *A representation of a clinical phenotype that is inferred from* ***a*** *combination of,* ***for example, diagnoses and*** *laboratory, image, and clinical findings about a given patient.*

These changes to the definitions of 'clinical history' and 'clinical picture' now properly capture situations where past diagnoses are elicited from the patient and/or her caregiver during a clinical history taking: these diagnoses are now clinical findings in the clinical history that was generated by the clinical history taking (see the definition of 'clinical finding' in Table 2).

Scenario 3: Misdiagnosis.

The third physician, Dr. Miller, misdiagnoses Mr. Jones' type 2 diabetes mellitus as type 1 diabetes mellitus (Figure 3). Per Smith and Ceusters, because the misdiagnosis is still about Mr. Jones, his disease, the relationship between them, and the type 'type 1 diabetes mellitus' on the level of reference, it is an information content entity. However, it fails to be about the configuration IUI-7 as a whole on the level of compound expression.

Again, in this scenario there exist PORs in one-to-one correspondence (except the configuration and its components) numbered IUI-43 through IUI-53 (Additional file 2 : Tables S4-S6). Dr. Miller (IUI-43) writes (IUI-53) his misdiagnosis (IUI-48) in Mr. Jones' chart, and the IQE (IUI-50) inhering in the ink (IUI-49) is conformant to his cognitive representation (IUI-46), and both are about—on the level of reference—Mr. Jones, his disease, and type 1 diabetes mellitus. But neither one is about the configuration (IUI-7). To capture the relation both (1) between the cognitive representation and the configuration and (2) between the IQE and the configuration, we define a new relation:

> **is-misrepresentation-of**: domain: representation, range: portion of reality.
>
> Def: *x* is-misrepresentation of *y* iif *x* is a representation and *x* is intended to be about *y* and it is not the case that *x* is about *y*.

Then we assert that the representations (IUI-46 and IUI-50) are misrepresentations of the configuration (Table 7 and Additional file 2 : Table S6). Note that our definition precludes the cognitive representation (IUI-46) and IQE (IUI-50) being about any configuration other than IUI-7, because they are not intended to be about, for example, the configuration of the sun, earth, and moon at a particular date and time.

Note that asserting the incorrect disease type is not the only way to make a misdiagnosis. There are at least six possibilities where a diagnosis fails to be about a configuration on the level of compound expression (Table 8). If a representation fails on the level of reference, it also fails on the level of compound expressions, because a configuration cannot consist of that which does not exist. These six possibilities could also exist in combination, but if the

2nd, 3rd, and 4th possibilities are all present (for example, "Ron Weasley has spattergroit"), then there is not a diagnosis, or even any information content entity at all, because the representation is not about anything even on the level of reference. Of course, if the organism itself does not exist, then there cannot be a clinical picture inferred, and thus it would not be a diagnosis or misdiagnosis, although it could still be an ICE if it is about a really-existing disease type (for example, "James Bond has influenza"). Also, as medical knowledge evolves, the profession comes to understand that certain types of disease thought to exist in fact do not. Thus past diagnoses of *dropsy* and *consumption* we now understand to be misdiagnoses.

Despite searching the extensive literature on diagnostic error, we could not find any studies that looked at what percentages of misdiagnoses fall into these categories. We conjecture based on our past clinical expertise and experience that asserting the incorrect disease type is the most common mistake among those in Table 8, but confirmation or rejection of this conjecture requires study.

Scenario 4: the lucky guess.

In this scenario, a layperson (the "seer"—IUI 63) correctly concluded coincidentally that Mr. Jones had type 2 diabetes mellitus based on the position of the moon and Mr. Jones' horoscope (Additional file 3 : Tables S7-S9). It would be wrong to say the seer's reasoning (IUI-71) constituted a diagnostic process. To avoid coincidentally correct statements from qualifying as diagnoses, we additionally require as input into the diagnostic process cognitive representations of the disease type and the types instantiated by the sequalae,

signs, symptoms, and any clinical, laboratory, or imaging findings or phenotypes of the instances of this disease type. Note that this is a minimal requirement: clinicians often additionally include in their diagnostic reasoning cognitive representations of other disease types and associated PORs when considering alternative possibilities for the disease type.

This view is based on the extensive literature on clinical reasoning processes, especially diagnosis (for a review, see Norman [33]). This research has established the use of representations, called 'knowledge structures', in the diagnostic process. The nature and form of these representations evolves as clinical expertise develops [34], and we note that the differences in diagnostic processes that result could result in a typology of diagnostic processes in OGMS.

Because the seer had no cognitive representations of type 2 diabetes mellitus, let alone used them as input into his "reasoning", his conclusion (IUI-68), although an ICE, is not a diagnosis. Similarly, if a physician makes a lucky guess based not on his cognitive representations of the stated disease type but instead by flipping a coin or some such, that too would not be a diagnosis.

To Table 3 we add an aggregate of cognitive representations of disease types and associated entities as input into the diagnostic process (Table 9).

We propose to redefine diagnostic process as follows:

Diagnostic process =def. *An interpretive PROCESS that has as input (1) a CLINICAL PICTURE of a given patient* ***AND (2) an aggregate of REPRESENTATIONs of at least one type of disease and at least one type of phenotype whose instances are associated with instances of that disease,*** *and as output an assertion to the effect that the patient has a DISEASE of a certain type.*

Scenario 5: layperson's justifiable conclusion.

Mr. Jones' daughter wrote a sentence in her letter to her brother based on reading Dr. Smith's progress note saying that that her father has type 2 diabetes mellitus (Additional file 4 : Tables S10-S12). Of course, the daughter has not made a diagnosis. She is communicating to her brother what she believes to be the case.

Had she merely written "Dr. Smith says" and then copied Dr. Smith's sentence word for word into her letter, then her writing would concretize Dr. Smith's diagnosis (IUI-8). Similarly, if a person who does not know German copies German texts, then she reproduces the representations in the texts and they concretize the same ICEs as the original texts. This is the case of hearsay ("so-and-so said it was the case that…").

As Smith and Ceusters showed, however, the same sentence written by two different people does not guarantee they concretize the same ICE. ICEs are further differentiated by the provenance of their concretizations, including who created them and when, and to what POR they intend to be about. In their example, two people writing the sentence *Barack Obama has never been President of the United States*—one before and one after

Obama's inauguration as President—generate two different ICEs. The one written after fails on the level of compound expressions but not on the level of reference, whereas the one written before succeeds on both levels (it remains true that at the time when the sentence was written, he had never been President).

We therefore distinguish between a human (1) merely copying a representation, in which case the copy concretizes the same ICE as the original text and (2) creating her own cognitive representation of the POR—which involves forming a belief that the POR really existed as represented—and then subsequently creating an IQE that is conformant to the cognitive representation. In the former case, a new ICE does not come into being. It does not even require in the cognitive system of the copier any representation of the POR that the original representation is about (as in the case of copying German text that one does not understand at all). In the latter case, by contrast, a new ICE does come into being.

In Scenario 5, the daughter did not merely repeat Dr. Smith's diagnosis. She communicated to her brother *her* belief about her father's disease. She deliberately chose not to merely convey Dr. Smith's diagnosis, but rather her belief that her father has type 2 diabetes mellitus. She heard the opinion of an expert, in whom she had trust. Based on (1) her observations of her father, (2) Dr. Smith's diagnosis, and (3) her trust in Dr. Smith, she reached the conclusion herself that her father suffers from type 2 diabetes mellitus. Because she did not begin with a clinical picture and her own cognitive representations of type 2 diabetes mellitus, her conclusion is not a diagnosis.

However, consider the scenario where she is given the clinical picture and has enough knowledge to arrive at a conclusion, which could be the case either if she were a physician or somehow other acquired or were given the necessary knowledge: it is analagous to Scenario #6, where she takes the place of the expert system (see analysis of that scenario below). Thus, here in Scenario #5 it is important to note that she did not reason from a clinical picture to the diagnosis.

In this scenario, therefore, the daughter has created a new ICE (IUI-88) that is not a diagnosis. She has concretized it in the sentence (IUI-89) in her letter.

Scenario 6: diagnosis by non-human.

The diagnostic decision support system has made a diagnosis (or misdiagnosis depending on whether it is correct), because it (1) takes as input a clinical picture and representations of the relevant disease type and one or more types of phenotypes with which it is associated; (2) participates in a process of making a conclusion based on this input; and (3) outputs from this process a statement about a configuration involving an organism, a disease, and a disease type.

In this case, there are no cognitive representations. In their place are digital representations on hard drives, memory chips, and central processing units. If we assume the system generates a sentence and prints it on paper, then we have an analagous IQE to the written diagnosis of the physician and ICE of the sister.

Nothing in our proposed definitions conflicts with this scenario. Replacing Dr. Smith and associated representations and diagnostic process with various components of the computer and its digital representations as well as inferential process (which is an instance of diagnostic process) is straightforward.

Returning briefly to a point made in Scenario #5, Mr. Jones' daughter could follow the exact same algorithm(s) of the diagnostic expert system using the exact same clinical picture as input, and she would arrive at (or make) a diagnosis, in contrast to scenario #5 where her conclusion was an ICE but not a diagnosis.

Conclusions

We applied Smith and Ceusters' results on aboutness [24] to diagnosis in order to develop an account of diagnosis, misdiagnosis, lucky guesses, hearsay, a layperson's justified belief about disease configurations, and a diagnosis made by an expert system. Our key result is that a correct diagnosis, as defined by OGMS, is about a configuration of an organism, its disease, and the type the disease instantiates (level of compound expression) in a specified portion of spacetime. We identified several subtypes of misdiagnosis (e.g., wrong disease subtype, wrong patient, wrong temporal placement) that have not been differentiated in the literature on diagnostic error, to our knowledge. Studying the incidence and causes of these subtypes might advance the study of diagnostic error and strategies to reduce it. Note that as we have defined it, 'misdiagnosis' does not refer to the diagnostic errors of absent diagnosis (failing to diagnose a disease at all, let alone incorrectly) and delayed diagnosis.

The provenance of the ICE and its concretizations are critical: lucky guesses, hearsay, and laypersons' conclusions about disease (when not arrived at through a diagnostic process using a clinical picture and cognitive representations of the associated type(s) of disease as input) do not constitute diagnoses and therefore are different types of ICE than diagnoses. Provenance also includes which findings and other information constituted the clinical picture used in the diagnostic process. Our analysis of the scenarios identified past diagnoses as important input into the diagnostic process, leading to proposed redefinitions of 'clinical history', 'clinical picture', and 'diagnostic process' for OGMS.

Smith and Ceusters' results on aboutness and our extension of them here to diagnosis reduce the need for the workarounds reported by Martínez Costa and Schulz [25] and Hastings et al. [26] It is perfectly legitimate to define 'suspected heart failure finding' with an existential quantifier: if an instance of this type is not about a really-existing configuration of patient–disease–heart failure, it is still an ICE that is individually about the patient, her condition, and the type *heart failure* on the level of reference. In OWL, we could assert:

Suspected heart failure ICE -> ICE and (**is about** SOME *Organism*)

Suspected heart failure ICE -> ICE and (**is about** SOME *Condition*)

In more expressive formalisms including first-order logic, we could also assert that it is about *heart failure*, where 'Type', 'Instance_of', and 'Is_about' are predicates in what follows, where the universal quantification applies to the ICE, not what it is about:

Type(heart_failure)

Type(suspected_heart_failure_ICE)

∀*x* (Instance_of(*x*, suspected_heart_failure_ICE) -> Is_about(*x*, heart_failure))

Similarly, chemical graphs and diagrams are ICEs about individual types of atoms such as carbon, oxygen, hydrogen, and so on, even when they fail to be about any type of configuration (molecule) of such atoms. However, because they are typically not about any instances, proper existential quantification in OWL is not possible. However, we could assert in first-order logic, for the diagram of *octaazacubane* (a hypothetical molecule which would be comprised of eight nitrogen atoms arranged in a cubic structure), again where the universal quantification applies to the ICE and not what it is about:

Type(nitrogen_atom)

Type(octaazacubane _diagram)

∀*x* (Instance_of(*x*, octaazacubane _diagram) -> Is_about(*x*, nitrogen_atom))

It is therefore not required to use universal quantification over the range of things that an ICE is about, when defining ICEs, to avoid failure of aboutness on the level of compound expression. This result is qualified by the constraints of representational formalisms such as OWL that prevent directly asserting aboutness to types. Note however that Schulz et al. describe workarounds in OWL to asserting aboutness to types, that may be of benefit in some use cases [35].

This work is limited by the fact that we did not conduct further ontological analysis of the diagnostic process beyond OGMS and beyond what our scenarios required, as this was not the purpose of the present work. We do note that our requirement for including cognitive representations of disease types as input into the diagnostic process is based on this

literature, however. Engaging experts in the study of clinical reasoning in future work to develop a typology of diagnostic processes has the potential to result in a corresponding typology of diagnoses.

Future work includes (1) an account of differential diagnosis, where a clinician or expert system generates a list of likely types of disease for further investigation to identify the actual type the organism's disease instantiates; (2) proposing to the OGMS community to clarify the definitions of 'clinical history', 'clinical picture', and 'diagnostic process' as suggested here, and to expand the definition of diagnosis to include disorders, disease courses, and absence of disease (i.e., healthy); (3) extending our analysis as reported here to this expanded definition of 'diagnosis'; and (4) conducting deeper ontological analysis of the diagnostic process, in coordination with experts in the study of clinical reasoning.

List of Abbreviations

BFO	Basic Formal Ontology
GDC	Generically dependent continuant
IAO	Information Artifact Ontology
ICE	Information Content Entity
IQE	Information Quality Entity
OGMS	Ontology for General Medical Science
POR	Portion of Reality
RT	Referent Tracking
RTT	Referent Tracking Tuple

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

The authors contributed equally to the ontological analysis and development of results. Author WRH created the first version of the manuscript. Both authors had full access to all materials and analysis and participated in revising the manuscript. Both authors approved the final version of the manuscript.

Acknowledgments

This work was suppoted in part by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1TR001427.

References

1. Bingham CO, 3rd, Bartlett SJ, Merkel PA, Mielenz TJ, Pilkonis PA, Edmundson L, Moore E, Sabharwal RK: **Using patient-reported outcomes and PROMIS in research and clinical applications: experiences from the PCORI pilot projects**. *Qual Life Res* 2016.
2. Scanlon L: **PatientsLikeMe Survey Shows Vast Majority of People With Health Conditions Are Willing To Share Their Health Data**. In. Cambridge, Mass: PatientsLikeMe; 2014.
3. Rudin RS, Motala A, Goldzweig CL, Shekelle PG: **Usage and effect of health information exchange: a systematic review**. *Ann Intern Med* 2014, **161**(11):803-811.
4. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P: **From the Office of the National Coordinator: the strategy for advancing the exchange of health information**. *Health Aff (Millwood)* 2012, **31**(3):527-536.
5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS: **Launching PCORnet, a national patient-centered clinical research network**. *J Am Med Inform Assoc* 2014, **21**(4):578-582.

6. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM *et al*: **The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies**. *BMC Med Genomics* 2011, **4**:13.
7. Owens B: **DATA SHARING. Montreal institute going 'open' to accelerate science**. *Science* 2016, **351**(6271):329.
8. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF: **Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors**. *Med Care* 2005, **43**(5):480-485.
9. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM: **Measuring diagnoses: ICD code accuracy**. *Health Serv Res* 2005, **40**(5 Pt 2):1620-1639.
10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ *et al*: **Caveats for the use of operational electronic health record data in comparative effectiveness research**. *Med Care* 2013, **51**(8 Suppl 3):S30-37.
11. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS: **Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied**. *Med Care* 2013, **51**(8 Suppl 3):S80-86.
12. Botsis T, Hartvigsen G, Chen F, Weng C: **Secondary Use of EHR: Data Quality Issues and Informatics Opportunities**. *AMIA Jt Summits Transl Sci Proc* 2010, **2010**:1-5.
13. Benesch C, Witter DM, Jr., Wilder AL, Duncan PW, Samsa GP, Matchar DB: **Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease**. *Neurology* 1997, **49**(3):660-664.
14. Shapiro M, Johnston D, Wald J, Mon D: **Patient-Generated Health Data White Paper**. In.: RTI International Research Triangle Park, NC 27709; 2012.
15. Gordon NP, Mellor RG: **Accuracy of parent-reported information for estimating prevalence of overweight and obesity in a race-ethnically diverse pediatric clinic population aged 3 to 12**. *BMC pediatrics* 2015, **15**(1):5.
16. Komaroff AL: **The variability and inaccuracy of medical data**. *Proceedings of the IEEE* 1979, **67**(9):1196-1296.
17. Callahan CM, Tu W, Stump TE, Clark DO, Unroe KT, Hendrie HC: **Errors in self-reports of health services use: impact on alzheimer disease clinical trial designs**. *Alzheimer Dis Assoc Disord* 2015, **29**(1):75-81.
18. Monte AA, Heard KJ, Hoppe JA, Vasiliou V, Gonzalez FJ: **The accuracy of self-reported drug ingestion histories in emergency department patients**. *J Clin Pharmacol* 2015, **55**(1):33-38.
19. Gerritsen M, Berndt N, Lechner L, de Vries H, Mudde A, Bolman C: **Self-Reporting of Smoking Cessation in Cardiac Patients: How Reliable Is It and Is Reliability Associated With Patient Characteristics?** *J Addict Med* 2015, **9**(4):308-316.
20. Raphael KG, Janal MN, Sirois DA, Dubrovsky B, Klausner JJ, Krieger AC, Lavigne GJ: **Validity of self-reported sleep bruxism among myofascial temporomandibular disorder patients and controls**. *J Oral Rehabil* 2015, **42**(10):751-758.

21. Patel M, Perrin K, Pritchard A, Williams M, Wijesinghe M, Weatherall M, Beasley R: **Accuracy of patient self-report as a measure of inhaled asthma medication use**. *Respirology* 2013, **18**(3):546-552.
22. Woodfield R, Group UKBSO, Follow-up UKB, Outcomes Working G, Sudlow CL: **Accuracy of Patient Self-Report of Stroke: A Systematic Review from the UK Biobank Stroke Outcomes Group**. *PLoS One* 2015, **10**(9):e0137538.
23. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ: **How the provenance of electronic health record data matters for research: a case example using system mapping**. *EGEMS (Wash DC)* 2014, **2**(1):1058.
24. Smith B, Ceusters W: **Aboutness: Towards Foundations for the Information Artifact Ontology**. In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015.
25. Martínez-Costa C, Schulz S: **Ontology-based reinterpretation of the SNOMED CT context model**. In: *Proceedings of the Fourth International Conference on Biomedical Ontology: July 7th-12th, 2013; Montreal*: Edited by Dumontier M, Hoehndorf R, Baker CJO. 2013: 90-95.
26. Hastings J, Batchelor C, Neuhaus F, Steinbeck C: **What's in an 'is about' link? Chemical diagrams and the information artifact ontology**. In: *Proceedings of the 2nd International Conference on Biomedical Ontology; Buffalo, New York*: Edited by Bodenreider O, Martone ME, Ruttenberg A. 2011: 201-208.
27. Ceusters W, Hogan WR: **An ontological analysis of diagnostic assertions in electronic healthcare records** In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015.
28. Hogan WR: **To what entities does an ICD-9-CM code refer? A realist approach.** In: *Bio-ontologies; Boston, MA*: Edited by Shah N, Sansone S-A, Stephens S, Soldatova L. 2010.
29. Scheuermann RH, Ceusters W, Smith B: **Toward an ontological treatment of disease and diagnosis**. In: *AMIA Summit on Translational Bioinformatics: 2009*. 116-120.
30. Chisholm RM: **The primacy of the intentional**. *Synthese*, **61**(1):89-109.
31. Hawass Z, Gad YZ, Ismail S, Khairat R, Fathalla D, Hasan N, Ahmed A, Elleithy H, Ball M, Gaballah F *et al*: **Ancestry and pathology in King Tutankhamun's family**. *JAMA* 2010, **303**(7):638-647.
32. Ceusters W, Smith B: **Foundations for a realist ontology of mental disease**. *J Biomed Semantics* 2010, **1**(1):10.
33. Norman G: **Research in clinical reasoning: past history and current trends**. *Medical Education* 2005, **39**(4):418-427.
34. Schmidt HG, Rikers RMJP: **How expertise develops in medicine: knowledge encapsulation and illness script formation**. *Medical Education* 2007, **41**(12):1133-1139.
35. Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A: **An Ontological Analysis of Reference in Health Record Statements**. In: *Formal Ontology in Information Systems: Proceedings of the Eighth International Conference (FOIS 2014): 2014*. IOS Press: 289.

Table 1. Definitions based on Smith and Ceusters [24].

Term	**Definition**
INFORMATION CONTENT ENTITY	An ENTITY which is (1) GENERICALLY DEPENDENT on (2) some MATERIAL ENTITY and which is (3) concretized by a QUALITY (a) inhering in the MATERIAL ENTITY and (b) that is_about some PORTION OF REALITY
INFORMATION QUALITY ENTITY	A REPRESENTATION that is the concretization of some INFORMATION CONTENT ENTITY
REPRESENTATION	A QUALITY which is_about or is intended to be about a PORTION OF REALITY
MENTAL QUALITY	A QUALITY which specifically depends on an ANATOMICAL STRUCTURE in the cognitive system of an organism
COGNITIVE REPRESENTATION	A REPRESENTATION which is a MENTAL QUALITY
Relation	**Explanation**
x ***is_about*** *y*	*x refers to or is cognitively directed towards y*. **Domain:** representations; **Range:** portions of reality
x ***concretizes*** *y*	*x* is a QUALITY and *y* is a GENERICALLY DEPENDENT CONTINUANT (GDC) and for some MATERIAL ENTITY *z*, *x* ***specifically_depends_on*** z at *t* and *y* ***generically_depends_on*** *z* at *t*, and if *y* migrates from bearer *z* to another bearer *w* then a copy of *x* will be created in *w*.
x ***is_conformant_to*** *y*	=def. *x* is an INFORMATION QUALITY ENTITY and *y* is a COGNITIVE REPRESENTATION and there is some GDC *g* such that *x* ***concretizes*** *g* and *y* ***concretizes*** *g*.

Table 2. Key definitions from OGMS used in the analysis

Term	Definition
DISEASE	A DISPOSITION (i) to undergo PATHOLOGICAL PROCESSes that (ii) exists in an ORGANISM because of one or more DISORDERs in that ORGANISM.
DISORDER	A causally relatively isolated combination of physical components that is (a) clinically abnormal and (b) maximal, in the sense that it is not a part of some larger such combination.
DIAGNOSIS	A conclusion of an interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion (diagnostic statement) to the effect that the patient has a DISEASE of such and such a type.
DIAGNOSTIC PROCESS	An interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion to the effect that the patient has a DISEASE of a certain type.
PATHOLOGICAL PROCESS	A bodily PROCESS that is a manifestation of a DISORDER.
PHENOTYPE	A bodily feature or combination of bodily features of an organism determined by the interaction of the genetic make-up of the organism and its environment.
CLINICAL PHENOTYPE	A clinically abnormal PHENOTYPE.
CLINICAL PICTURE	A representation of a CLINICAL PHENOTYPE that is inferred from the combination of laboratory, image and clinical findings about a given patient.
CLINICAL FINDING	A REPRESENTATION that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof.
MANIFESTATION OF DISEASE	A QUALITY of a patient that is (a) a deviation from clinical normality that exists in virtue of the realization of a disease and (b) is observable.
CLINICAL HISTORY TAKING	An interview in which a clinician elicits a clinical history from a patient or from a third party who is authorized to make health care decisions on behalf of the patient.
CLINICAL HISTORY	A series of statements representing health-relevant features of a patient.

Table 3. Referent tracking tuples true in every scenario

IUI	**Entity**	**Existence period**	**Type**	**Notes**
IUI-1	Mr. Adam Jones	*t1* – the period during which IUI-1 exists	Material Entity	
IUI-2	IUI-1's disease	*t2*	Disposition	
Relationships among particulars				**Notes**
IUI-2	**inheres in**	IUI-1	at *t2*	
IUI-2	**instance of**	UUI-1	at *t2*	UUI-1 is a universal unique identifier that denotes *type 2 diabetes mellitus.* We assume that if something is at any time of its existence an instance of type 2 DM, it is instance of type 2 DM at all times it exists.

Table 4. The entities in Scenario 1

IUI	Entity	Existence period	Type	Notes
IUI-3	Dr. Anne Smith	t3	Human being	
IUI-4	Cognitive system of IUI-3	t4		
IUI-5	An anatomical entity that is part of IUI-4	t5	Anatomical entity	Which anatomical entity and its lifetime cannot be easily specified given current state of neuroscience.
IUI-6	Quality that inheres in IUI-5 and is about IUI-7	t6	Cognitive representation	
IUI-7	The POR that is truth-maker for IUI-8	t7	Configuration	Mr. Jones, his disease, their relationship, and disease's instantiation
IUI-8	Dr. Smith's diagnosis	t8	Diagnosis	ICE concretized by IUI-6 and IUI-10
IUI-9	That which is written down on paper and forms the sentence.	t9	Material entity	*I conclude therefore that Mr. Jones has type 2 diabetes mellitus.*
IUI-10	IQE that inheres in IUI-9.	t10	Information quality entity	The sentence began to exist as soon as ink was laid down on paper, but the IQE did not begin to exist until the sentence was finished.
IUI-11	Dr. Smith's interpretive process	occupies t11	Diagnostic process	Dr. Smith's diagnostic process that led to her diagnosis IUI-8
IUI-12	The clinical picture input into IUI-11	t12	Clinical picture	Dr. Smith's clinical picture as ascertained prior to t6
IUI-13	Dr. Smith writing her diagnosis in the note	occupies t13	Process	

Table 5. Additional temporal entities in Scenario 1.

Temporal identifier	Description	Notes
t14	The interval during which the anatomical entity (IUI-5) is part of the cognitive system (IUI-4)	This interval is not easily specified given the current state of neuroscience. It could be different than t3 and t4.
t15	The interval during which the clinical picture (IUI-12) is used in the interpretive process (IUI-11)	Could be shorter than t11
t16	The point in time at which the cognitive representation (IUI-6) and diagnosis (IUI-8) begin to exist	t16 ends t11. Because the ICE does not exist until the cognitive representation—its first concretization—exists, this is also the point in time at which the diagnosis begins to exist.
t17	The interval during which the cognitive representation (IUI-6) participates in the writing process (IUI-13)	
t18	The interval during which the diagnosis (IUI-8) participates in the writing process (IUI-13)	It is possible that the original cognitive representation (IUI-6) gets copied elsewhere in the brain for reasoning and thus that the ICE continues to participate after the initial cognitive representation
t19	The interval during which that which is written on paper (IUI-10) begins to exist until it exists in full	The writing process begins earlier than the time at which the sentence begins to exist: the author starts the process with getting a pen and paper, any preparation necessary ("clicking" the pen), etc.

Table 6. Relationships among particulars in Scenario 1.

IUI	Relation	IUI	When relation holds in reality	Notes
IUI-4	**part of**	IUI-3	at t4	
IUI-5	**part of**	IUI-4	at t14	All anatomical components in which the cognitive representation inheres are part of the cognitive system. We do not assume the cognitive system is limited to the brain, as the state of neuroscience does not permit such an assumption.
IUI-6	**inheres in**	IUI-5	at t6	
IUI-6	**is about**	IUI-7	at t6	The cognitive representation stands in aboutness to IUI-7 as long as it exists
IUI-6	**is about**	IUI-1	at t6	It is also about Mr. Jones
IUI-6	**is about**	IUI-2	at t6	And about Mr. Jones' disease
IUI-6	**is about**	UUI-1	at t6	And about Type 2 diabetes mellitus
IUI-6	**concretizes**	IUI-8	at t6	It also concretizes the diagnosis
IUI-10	**inheres in**	IUI-9	at t9	The IQE inheres in the sentence on paper
IUI-10	**is about**	IUI-7	at t10	The IQE stands in aboutness to IUI-7
IUI-10	**is about**	IUI-1	at t10	It is also about Mr. Jones
IUI-10	**is about**	IUI-2	at t10	And about Mr. Jones' disease
IUI-10	**is about**	UUI-1	at t10	And about Type 2 diabetes mellitus
IUI-10	**concretizes**	IUI-8	at t10	
IUI-10	**is conformant to**	IUI-6	at t10	Is conformant to the cognitive representation as long as it exists
IUI-3	**agent in**	IUI-11	at t11	
IUI-12	**input into**	IUI-11	at t15	Clinical picture input into IUI-11
IUI-6	**output of**	IUI-11	at t16	Cognitive representation output from IUI-11
IUI-8	**output of**	IUI-11	at t16	Both the diagnosis and its concretization are outputs of IUI-11
IUI-8	**input into**	IUI-13	at t17	The diagnosis is input into writing
IUI-6	**input into**	IUI-13	at t18	As is its cognitive representation
IUI-10	**output of**	IUI-13	at t19	The sentence is output of writing

Table 7. Relationships of representations to portions of reality in Scenario 3: *Incorrect diagnosis*.

Relationships among particulars				Notes
IUI-46	**is about**	IUI-1	at t46	Dr. Jane Miller's cognitive representation is about Mr. Jones
IUI-46	**is about**	IUI-2	at t46	And Mr. Jones' disease
IUI-46	**is about**	UUI-2	at t46	And Type 1 diabetes mellitus (denoted by UUI-2)
IUI-50	**is about**	IUI-1	at t50	Likewise with the IQE inhering in the ink on paper
IUI-50	**is about**	IUI-2	at t50	
IUI-50	**is about**	UUI-2	at t50	
IUI-46	**is misrepresentation of**	IUI-7	at t46	But the cognitive representation is a misrepresentation of the configuration, i.e., it is intended to be about the configuration but fails on the level of compound expression
IUI-50	**is misrepresentation of**	IUI-7	at t50	The same is true of the IQE

Table 8. Six possibilities for a diagnosis failing in aboutness on the level of compound expressions.

Problem	**Where it fails *first***	**Description**
Noninstantation, asserted type exists	Level of compound expression	Disease instantiates a different type than the stated type, but the stated type exists
Noninstantation, asserted type does not exist	Level of reference	Disease instantiates a different type than stated, while the stated type of disease does not exist
Disease nonexistence	Level of reference	The disease instance does not exist
Organism nonexistence	Level of reference	The organism instance does not exist. In this case, there could not be a clinical picture properly inferred and thus it is not a misdiagnosis although it could still be an ICE.
Disease non-inherence	Level of compound expression	The disease inheres in a different organism than the one stated. For example, the doctor mistakenly ascribes Mr. Johnson's hypertension to his twin.
Configuration is not located in that part of spacetime where the diagnosis says it is located.	Level of compound expression	A diagnosis of type 2 diabetes mellitus 5 years ago is wrong because the patient didn't have the disease at that time, even though the patient has type 2 diabetes today. Also, a diagnosis that the patient has an upper respiratory tract infection today when in reality the infection resolved two weeks ago.

Table 9. Additional tuples required to distinguish diagnosing from a lucky guess.

IUI	Entity	Lifetime	Type	Notes
IUI-14	The aggregate of Dr. Smith's cognitive representations of various disease types and their associated types of phenotypes including type 2 diabetes mellitus that he used in the diagnostic process	t20	Aggregate of cognitive representations	
Relationships among particulars				**Notes**
IUI-14	**input into**	IUI-11	at t21	t21 refers to the temporal interval during which IUI-14 participated in the reasoning process. It could start at the same time as t11 or after t11, and end at the same time as or before t11.

Figure legends

Figure 1. The configuration of Mr. Jones, his disease, and type 2 diabetes mellitus

Figure 2. Diagram of diagnostic process, its inputs, a correct diagnosis, its concretization, and the configuration that that the concretization is about

Figure 3. Misdiagnosis of type of disease. The diagnosis is individually about the patient, the disease, and the incorrectly diagnosed disease type Y, but it is not about the configuration of patient, disease, and disease type X.

Figure 1

Click here to download Figure Fig 1.png

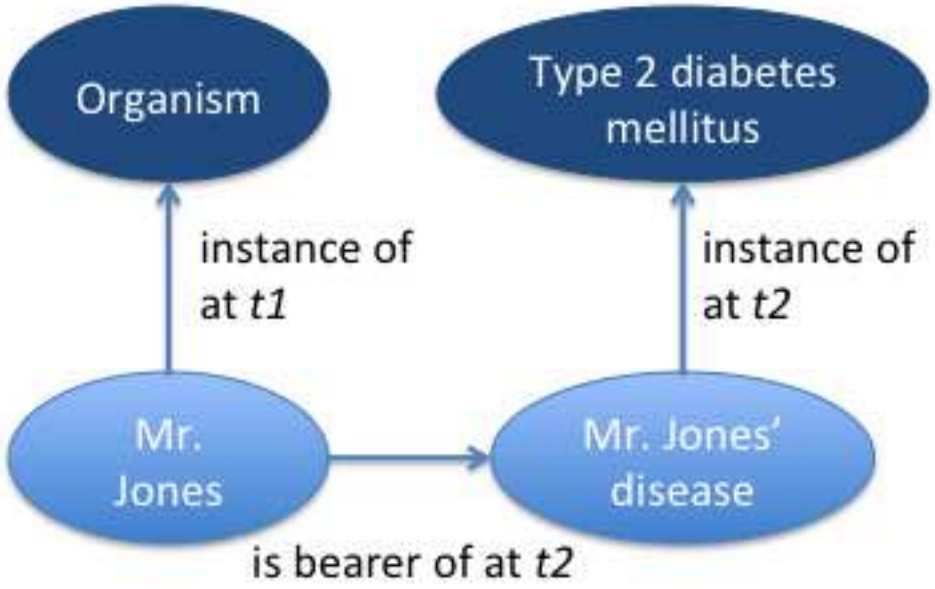


Figure 2

Click here to download Figure Fig 2.png

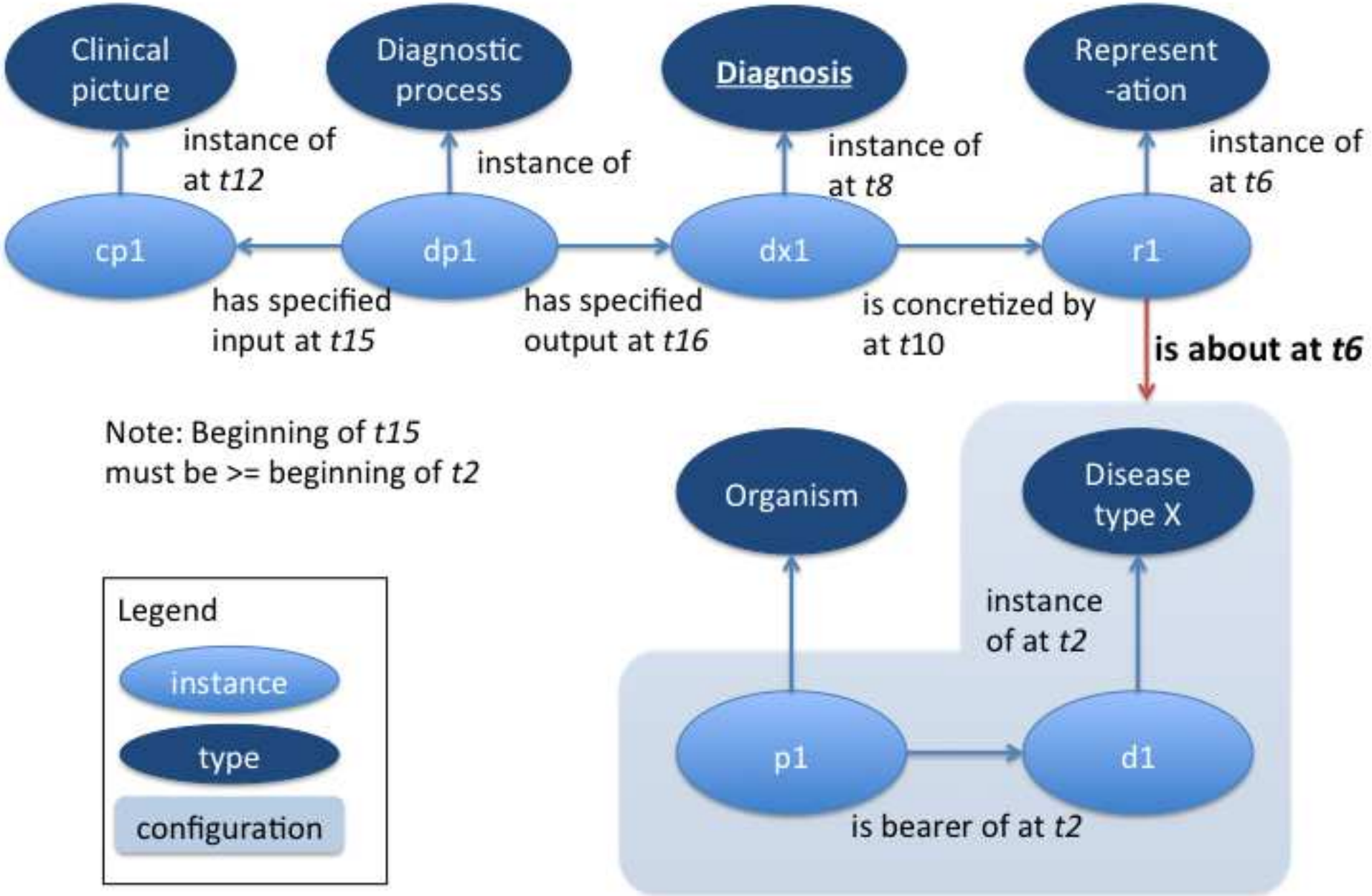
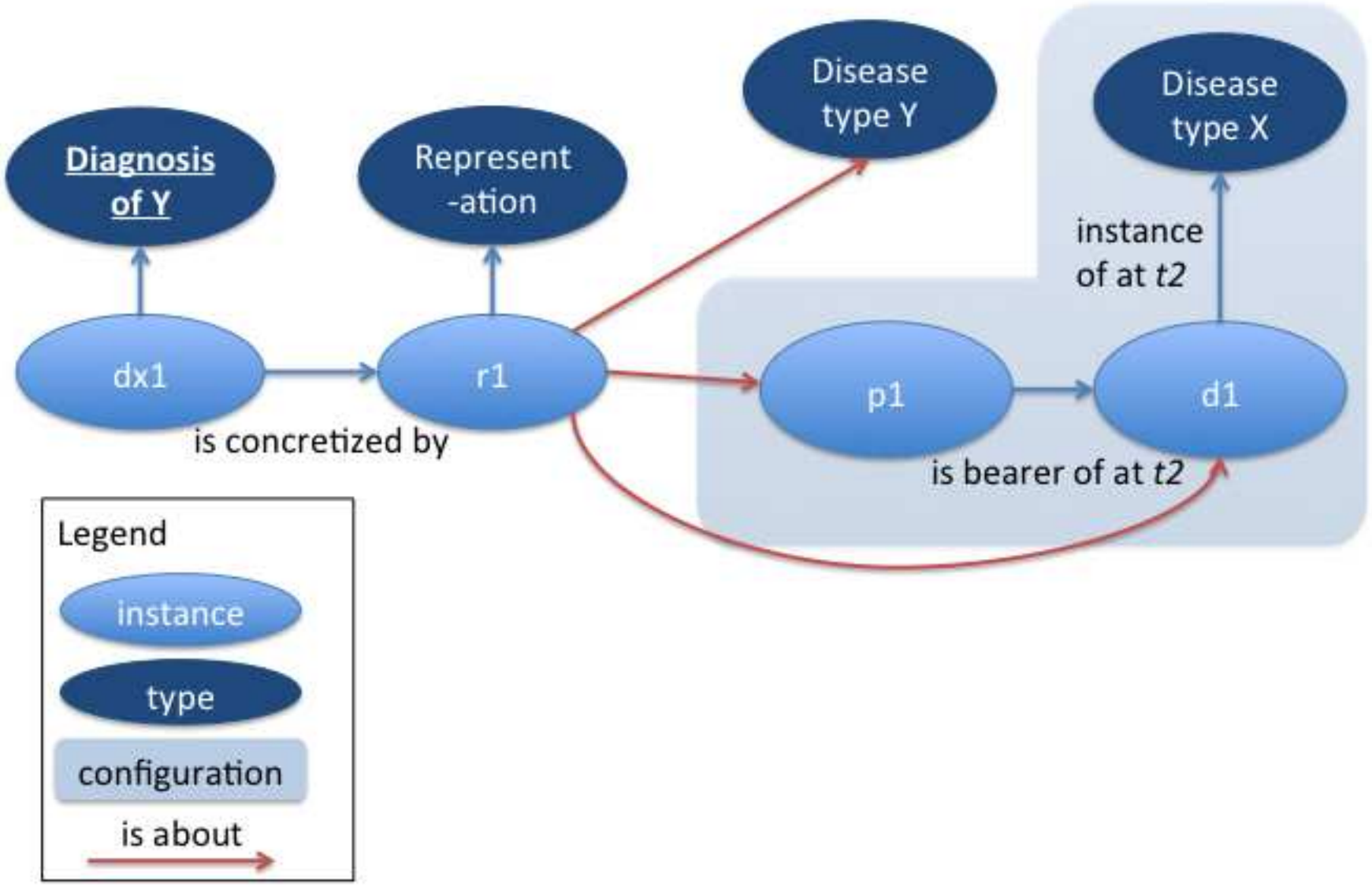


Figure 3

Click here to download Figure Fig 3.png



Additional file 1

Click here to access/download
Supplementary Material
additional file 1 v3.docx

Additional file 2

Click here to access/download
Supplementary Material
additional file 2 v3.docx

Additional file 3

Additional file 4

Journal of Biomedical Semantics

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.

--Manuscript Draft--

Manuscript Number:	JBSM-D-15-00018R2
Full Title:	Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.
Article Type:	Research
Funding Information:	National Center for Advancing Translational Sciences (UL1TR001427) — Not applicable Patient-Centered Outcomes Research Institute (CDRN-1501-26692) — Dr. William R. Hogan
Abstract:	Background Disease and diagnosis have been the subject of much ontological inquiry. However, the insights gained therein have not yet been well enough applied to the study, management, and improvement of data quality in electronic health records (EHR) and administrative systems. Data in these systems suffer from workarounds clinicians are forced to apply due to limitations in the current state-of-the art in system design which ignore the various types of entities that diagnoses as information content entities can be and are about. This leads to difficulties in distinguishing amongst diagnostic assertions misdiagnosis from correct diagnosis, and the former from coincidentally correct statements about disease. Methods We applied recent advances in the ontological understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created six scenarios that we analyzed using the method of Referent Tracking to identify all the entities and their relationships which must be present for each scenario to hold true. We discovered deficiencies in existing ontological definitions and proposed revisions of them to account for the improved understanding that resulted from our analysis. Results Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are typically about a configuration in which there exists a disease that inheres in an organism and instantiates a certain type (e.g., hypertension). Misdiagnoses are ICEs whose concretizations succeed in aboutness on the level of reference for individual entities and types (the organism and the disease), but fail in aboutness on the level of compound expression (i.e., there is no configuration that corresponds in total with what is asserted). Provenance of diagnoses as concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief. Conclusions Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR and other systems to allow diagnostic assertions to be captured with less ambiguity.
Corresponding Author:	William R. Hogan, MD, MS University of Florida Gainesville, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Florida
Corresponding Author's Secondary Institution:	

First Author:	William R. Hogan, MD, MS
First Author Secondary Information:	
Order of Authors:	William R. Hogan, MD, MS
	Werner Ceusters, MD
Order of Authors Secondary Information:	
Response to Reviewers:	>Reviewer reports: > >Reviewer #1: My review is based on the revision and response the authors >provided. I do not believe the authors have sufficiently addressed my >comments, and I do not believe the manuscript has improved >substantially. I respond to the authors' comments in detail below. >1. The manuscripts still lacks appropriate context or any biomedical >applications: This reviewer contradicts him/herself below, as can be judged from the following paragraph in his/her comments: >The authors have now added two paragraphs to the beginning of the >manuscript, where they cite a striking number of 23 articles (in the >first paragraph of the manuscript alone). These articles might be >relevant to the work in the authors' manuscript, but there is no >discussion about what this relation is. The authors provide context, >but do not present their work in this context. With regards to not presenting our work in this context, we disagree. Our work is definitely presented in this context, or it would not be context. Or is this reviewer thinking about context in relation to his/her own work, whatever that might be? Nevertheless, per below comments, we have better related our work to this context (as we perceive it) as requested. Note still, however, that our paper is fundamentally about taking an advance in our ontological understanding of misinformation (i.e., that of Smith and Ceusters) and applying it to the particular case of diagnosis. Regardless of whether the results are an advance or not, they are certainly reportable. >The lack of references does not show the novelty of the approach, it >merely shows that the authors are unable or unwilling to present their >work in the appropriate context. I have not asked for references to >referent tracking, but to related work on the problems that referent >tracking aims to solve. The authors will then have to demonstrate that >their referent tracking based approach indeed solves these problems. > It is not true that this reviewer originally, in his/her first review, asked for references to related work on the problems that referent tracking aims to solve. The reviewer at that time merely complained that six out of 10 references in the original manuscript were to the authors' own work, without asking for anything at all in response, or saying specifically WHY he/she thought this was a problem (or even that he/she thought it was a problem at all). The reviewer in his/her subsequent comments said that 3 references were indeed to related work. Furthermore, the papers we cited as additional context, do indeed talk about problems that are unsatisfactorily solved by other proposals, and do provide evidence that RT is able to solve them. It would be totally out of place, however, to go into these details in this paper since the problem we address here is that of ambiguities arising from how EHRs typically represent diagnoses. So we believe that we have addressed the original issue (lack of context and citations to related work __on misdiagnosis and clinical data error in general__), and that this reviewer is moving the goalposts.

Furthermore, referent tracking is not the major aspect of the novelty of our work—it was merely our methodology. Again, the key issue here is that the manuscript is about applying Smith and Ceusters work on misinformation to diagnosis. We used the methodology of referent tracking for the reasons stated in the manuscript, with appropriate reference to our prior results, which we obtained using referent tracking, to motivate it as our choice of methodology.

Furthermore, the prior literature on referent tracking discusses alternative approaches to referent tracking. Because this paper is not about an advance in, or new aspect of referent tracking, we do not see the need to compare referent tracking with other approaches to the problems that referent tracking aims to solve.

>When a novel theory is presented, it will also have to be presented in
>the context of existing theories. The authors intent is to "develop an
>improved account and explanation of phenomena in reality"; this
>requires an understanding of how this phenomenon is presented now and
>a demonstration of how some aspect of a theory describing this
>phenomenon has been improved. This is not in the manuscript.
>

Here we agree. The key insight that this reviewer's comment has prompted is that we took it for granted that there is almost no prior ontological work on defining misdiagnosis. Furthermore, recent reviews of data quality in general in our context have found very little work (only two papers or so) that is based on formal frameworks. The situation is bad enough that one review paper in fact calls for such frameworks to be developed and used in future work on data quality.

We have revised the manuscript to call out this situation (in the process quoting and citing the review paper in question, which is now reference #24), as well as the novelty of our work in providing the beginnings of a framework for understanding data error with respect to diagnosis. We do not claim that the current work achieves the totality of a framework: there is future work to be done in applying Ceusters' categorization of errors (new references 37 and 38 of the manuscript) to diagnosis to understand the totality of ways in which a misdiagnosis can be wrong.

We agree that it was an omission. We made this omission because we were so close to the work that we could not see that a reader unfamiliar with ontology of information and the quality of clinical and research data in general, would be unable make the connection. Plus, it is always best to make the connection explicitly regardless.

This is indeed a great improvement to the manuscript, even if the state-of-the-art over which we have made an improvement is mostly nothing that adds to our own work (the review paper mentions Hogan's review paper of 1996 as an exemplar that is not ad hoc. Hogan bases whether data are correct or not on whether they accurately reflect the state of the patient at the time they were created.).

Also, the manuscript does present several other advances that this reviewer has chosen to ignore, including several improvements to definitions in the Ontology for General Medical Science. We note that the method of referent tracking has here proved itself again as being a sensitive instrument for detecting problems in ontological definitions. We revised the Discussion to highlight this contribution of the work as well.

>2. The main result, that the existential quantifier suffices for
>representing certain phenomena, is not correct, or, if I am wrong
>about that, it has not been demonstrated that this actually holds true.
>

It does suffice for _representing_ them, if not for defining them. These are two different issues. See below.

>Specifically, the authors claim: "It is perfectly legitimate to define
'>suspected heart failure finding' with an existential quantifier: if
>an instance of this type is not about a really-existing configuration
>of patient-disease-heart failure, it is still an ICE that is

>individually about the patient, her condition, and the type heart
>failure on the level of reference."
>
>If this was indeed a definition, than it would hold in both
>directions, in particular it would also be true that any ICE that is
>individually about a patient, condition, and the "heart failure" type
>would be a "suspected heart failure finding". This would include ICEs
>such as a confirmation that a heart failure does not exist. The
>authors have added the necessary restrictions in the manuscript, but
>these are not sufficient; a definition requires both. And the missing
>parts are exactly what will require the universal quantifier (if this
>is not so, please add the full formalization to the manuscript,
>including an actual definition of an ICE).

Our choice of the word “define” was thoroughly regrettable as it is a complete and entire mischaracterization of what we, Catalina-Martinez, and Hastings were all trying to accomplish, as we discuss below. The existential quantifier is indeed sufficient for representing the entities in question (a looser criterion for which necessary conditions suffice), which as we shall see below, is all that the groups were trying to accomplish (they were not trying to accomplish necessary and sufficient criteria, as we demonstrate below).

Furthermore, using the universal quantifier erroneously precludes saying that “suspected heart failure finding” is individually about a patient, her condition, and the type heart failure on the level of reference. If it can only be about “congestive heart failure”, then it cannot be about an organism without contradiction (unless material entities and dispositions are not declared to be disjoint, and both Catalina-Martinez and Hastings follow upper ontologies where they are disjoint). We mention this issue in the manuscript. So the universal quantifier is not workable in this context.

Catalina-Martinez and Schulz do not themselves state that their task was “defining”. They state that “…we use the universal quantification ('only') for __relating__ an information entity to a clinical situation, to avoid asserting the existence of an entity the existence of which cannot be guaranteed.” (emphasis added). The task is therefore one of mere relation, not definition, per the authors direct quote.

Hastings et al. also do not present the problem as one of definition. They state “To __place__ SDs and therefore CDs as subtypes of IAO’s ICE, we need to change the fundamental aboutness criterion from Equation (1) to value rather than existential restriction” (emphasis added). Here, the problem is one of placement, not definition. They then proceed to specify ONLY NECESSARY conditions for SD (Structural Diagram): SD subClassOf ICE and is_about only StructuredEntity.

So casting the problem as definition (i.e., necessary and sufficient conditions) was regrettable, overly strong, and a mistake. We retract it and cast the problem correctly as mere representation. In this correct reformulation, our point stands. The two papers cited definitely need not have used universal quantification in their axioms (which were NOT definitional or equivalent class axioms) for their relating and their placing.

We regret the error and have fixed it in the revised manuscript.

>
>
>Reviewer #2: # General considerations
>
>In their work the authors provide an ontological analysis of diagnosis and other related things. This re-review only
>mentions important changes to the prior version of the paper (first review from reviewer #2) which are relevant
>for the reviewers decision.
>
>The authors answered to all reviewers comments. They added to and changed their

prior version accordingly. The
>authors improved their manuscript to a large extent by (1) considering a much larger body of literature as
>relevant for their work, and (2) by clarifying some of the definitions they use.
>

We thank this reviewer for his/her time and effort spent on both reviews.

>
>
># Results
>
>There might be only few analysis on the quantity of the described error types, however, the reviewer has the
>impression that it is more a retrieval, terminology or search problem to identify relevant literature.
>
>There is a large body of literature around on patient safety and critical incidence management. I only performed a
>short search on critical incidence analysis and found e.g. this one:
>https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2658432/ which at least provides numbers of the type 1 error
>of your typology (assertion of wrong disease type).
>

We thank the reviewer for pointing out this paper. Of 17 “misdiagnoses” in the paper, 16 are of our type 1, and 1 is a “delayed diagnosis”.

We note that this discussion stems from our initial claim that our type 1 was the most common type of error without citing any evidence (but instead our combined clinical experience). Although this paper, brought to our attention by the reviewer, suggests we were originally correct, it is still only one study in one ED in one country during a four-month time period. So we hesitate to generalize, although we appreciate that there might be additional literature out there. In addition, we were quite surprised that a major, recent IOM report (2016) that reviewed extensively the literature on diagnostic error tried formulating incidences of diagnostic error and its subtypes but could not because “…available research estimates were not adequate to extrapolate a specific estimate or range of the incidence of diagnostic errors in clinical practice today.” (page 97). We mention this problem in the Discussion. This is why we remained silent on the issue. It was not for lack of interest or ability to search the literature: it was just that the most recent, authoritative review could not compute it so we did not feel it worth our time to attempt it either.

>The second most important error is the mistaken identity of the patient (disease non-inherence). This is mostly an
>assignement error due to wrong documentation, and in the literature there should be substantial information on
>this. One main objective of Master Patient Index systems is to prevent these errors. Here a relevant citation:
>Bittle, Mark J., Patricia Charache, und Daniel M. Wassilchalk. 2007. „Performance Improvement: Registration-
>Associated Patient Misidentification in an Academic Medical Center: Causes and Corrections". Joint
>Commission Journal on Quality and Patient Safety 33 (1): 25-33.
>
>
>https://books.google.de/books?id=1l3ncDHMwGoC&pg=PA111&lpg=PA111&dq=Performance+improvement:+registration-
>associated+patient+misidentification+in+an+academic+medical+center:+causes+and+corrections&source=bl&ots>=rAxBJuX5Zy&sig=k_QrGyYgFSiae1tZAp-
>aC98AgSM&hl=de&sa=X&ved=0ahUKEwjn99uopa7MAhVEOMAKHa5ADpcQ6AEIOjAD
>
>All the errors based on non-existence are of minor importance, due to the widespread

	use of electronical systems >in which predefined terminology (or classification) constrains physician or system itself to base documention on >existing diagnosis (e.g. from ICD) or existing patients (at least with a social security number, in Europe). > Depending on how our understanding of disease develops in the future, it might be premature to say that non-existence of the type of disease is of minor importance (non-existence of disease type). Non-existence of the disease instance itself might be an interesting if minor problem. It is a kind of “overdiagnosis” – saying the patient suffers from a condition and diagnosing it when in fact there is no condition on the side of the patient. This could likely occur if the patient is a hypochondriac or is malingering and the clinician is as yet unaware of it. We discuss these issues in the Discussion further. >This might be again a case for a deeper analysis and this might also be more a "processual problem": who and >where is a wrong diagnosis documented. There might be an initial representation of a correct diagnosis, however, >in real life documentation it is not correct reported and code or assigned. We appreciate this comment. We believe that a formal, ontological framework for understanding types of error (which our work represents the beginning of) helps to understand the associated processes and how they go wrong and result in various types of error. Ameliorating a given type of error depends on understanding where the process goes wrong that generates it. Our recognition of provenance speaks strongly to the process.

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.

William R. Hogan (corresponding)
University of Florida
P.O. Box 100219
2004 Mowry Rd
Gainesville, FL 32610-0219
hoganwr@ufl.edu
(352) 294-4197

Werner Ceusters
University at Buffalo
921 Main Street
Buffalo, NY 14203
ceusters@buffalo.edu

Abstract

Background

Disease and diagnosis have been the subject of much ontological inquiry. However, the insights gained therein have not yet been well enough applied to the study, management, and improvement of data quality in electronic health records (EHR) and administrative systems. Data in these systems suffer from workarounds clinicians are forced to apply due to limitations in the current state-of-the art in system design which ignore the various types of entities that diagnoses as information content entities can be and are about. This leads to difficulties in distinguishing amongst diagnostic assertions misdiagnosis from correct diagnosis, and the former from coincidentally correct statements about disease.

Methods

We applied recent advances in the ontological understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created six scenarios that we analyzed using the method of Referent Tracking to identify all the entities and their relationships which must be present for each scenario to hold true. We discovered deficiencies in existing ontological definitions and proposed revisions of them to account for the improved understanding that resulted from our analysis.

Results

Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are typically about a configuration in which there exists a disease that

inheres in an organism and instantiates a certain type (e.g., hypertension). Misdiagnoses are ICEs whose concretizations succeed in aboutness on the level of reference for individual entities and types (the organism and the disease), but fail in aboutness on the level of compound expression (i.e., there is no configuration that corresponds in total with what is asserted). Provenance of diagnoses as concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief.

Conclusions

Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR and other systems to allow diagnostic assertions to be captured with less ambiguity.

Keywords

Biomedical ontology

Referent tracking

Disease

Diagnosis

Information content entity

Representation

Ontological realism

Background

As administrative, clinical, and patient-reported data are increasingly shared and reused, especially for patient care [1-4] and research [1, 5-7], several issues with these data—including diagnosis data—are of increasing concern. The issue that appears to be of greatest concern is data error and the implications thereof for making decisions and conclusions based on them [8-13]. Although Shapiro et al., in a report for the Office of the National Coordinator for Health Information Technology, do not cite error as a concern for including patient-generated health data into the electronic health record (EHR) [14], there are known errors with patient self reporting especially in research [15-22]. A second issue of concern is data provenance [10, 23], i.e. information about who created the data, in what setting, how, when, for what purpose, and so on. For example, Johnson et al. noted that the provenance of symptom data was essential to using those data correctly to determine whether a colonoscopy was a screening vs. diagnostic procedure [23].

Data error and data provenance are closely related. For example, Hersh et al. note that data recorded in billing workflows for financial purposes are less accurate than clinical data [10]. Thus, timing, method, and purpose of recording data at a minimum—all aspects of provenance—are intertwined with accuracy. Furthermore, a key result of the Johnson et al. study is that "Researchers who do not consider data provenance risk compiling data that are systematically incomplete or incorrect" [23].

An ontological account of data error and data provenance can identify crucial distinctions. For example, there are significant differences among (1) a measured weight that is off

because the scale was not properly tared, (2) a 'rough' weight of 70kg entered in an emergency when the patient cannot be weighed, and (3) a weight measurement entered on the wrong patient. Detecting and accounting for these differences and their causes—especially the aspects of provenance that influence them—is necessary to inform strategies to study, cope with, and improve data error when using pre-existing EHR data for research.

Additionally, a recent review article on the methods for assessing quality of EHR data for clinical research found that: *Most of the studies included in this review presented assessment methodologies that were developed with a minimal empirical or theoretical basis* [24]. It concluded with a call for moving away from ad hoc approaches to data quality assessment, to formal, validated approaches. Although error is only one aspect of data quality (fitness for purpose and completeness are two others), a formal ontological understanding of data error could play a role in more formalized methods for data quality assessment.

In this work, we apply Smith and Ceusters' recent ontological account of incorrect information (i.e., error) [25] to diagnosis data in administrative systems, EHRs, and patient-reported information. Their account holds that a statement such as a diagnostic assertion can succeed or fail in aboutness on at least two levels: (1) the level of denoting single entities and/or types (i.e., the level of *reference*) and (2) the level of veridical representation of a configuration of multiple entities and/or types (i.e., the level of *compound expression*).

To succeed on the second level (compound expression), the information content entity (ICE) must be correct about *all* particulars, their relationships, and their instantiations of types that it mentions. Failure on a single particular, relation, or instantiation causes the ICE to fail at the second level while still potentially succeeding at the first level. For example, if Mrs. Jones has type 1 diabetes mellitus, then the sentence '*Mrs. Jones suffers from type 2 diabetes mellitus'* fails in aboutness on the level of compound expression because it misstates one thing: her disease does not instantiate type 2 diabetes mellitus. However, despite this failure the sentence is nevertheless still about Mrs. Jones, about her disease, and about type 2 diabetes mellitus on the level of reference, because indeed it mentions those three entities. It is therefore, per Smith and Ceusters, an ICE that is about *something* even though it is a misdiagnosis.

Prior ontological work on the aboutness of clinical statements like diagnoses has been constrained by the view that an ICE is about nothing (or is perhaps not even an ICE at all) if it fails on the level of compound expression. Martínez Costa and Schulz, for example, use the universal quantifier when relating an information entity to a clinical situation *…to avoid asserting the existence of an entity the existence of which cannot be guaranteed* [26]. For an ICE such as 'suspected heart failure' they want to avoid the implication that there is some instance of heart failure that it is about. Because they cannot guarantee the existence of some heart failure, they use universal quantification to say 'if it is about anything, then it is about an instance of heart failure'. Researchers working in areas other than diagnosis have encountered similar issues. For example, Hastings et al. note that chemical graphs and diagrams are not always about types of molecules that exist [27]. They, too, used the

workaround of replacing existential quantification with universal quantification to avoid asserting that every chemical graph/diagram is about some type of molecule that exists (level of compound expression), while still allowing such graphs and diagrams to be subtypes of information content entity.

In our own, previous ontological analysis of diagnosis, using the methodology of referent tracking, we identified what entities must exist or must have existed for a particular diagnostic statement to hold true [28, 29]. A key result of this work is that a diagnosis is minimally about *both* the patient and the type of disease that is asserted to exist. In addition, building on previous work on the Ontology for General Medical Science (OGMS), the foundations of which were laid down in Scheuermann et al. [30], we noted that for a diagnosis to exist (at least in medicine and under the assumption that the diagnosis was made *lege artis*), there must also have existed a diagnostic process, a person who carried out that process, and a clinical picture which was used as input into that process.

The hypothesis for the work described here was that applying Smith and Ceusters' results to disease and diagnosis, in combination with prior work on the ontology of disease and diagnosis (including provenance of the latter), could address limitations encountered in previous ontological work on disease and diagnosis and improve our representations of them in support of studying, coping with, and reducing ambiguity in the generation of diagnostic statements and error in the interpretation thereof.

Methods

To test this hypothesis, we analyzed a set of scenarios that we created and that involve correct and incorrect diagnoses, lucky guesses, and justified layperson belief in the existence of a disease of a certain type. The goal was to explore whether, and if so how, a realism-based account of information can deal successfully not only with diagnostic statements asserting the ideal case of a correct diagnosis, but also with deviations from the ideal.

Materials

In our analysis we used as input (1) Smith and Ceusters' work on aboutness and their definitions of representation, mental quality, cognitive representation, and information quality entity (Table 1), (2) definitions of disease, disorder, and diagnosis from the Ontology for General Medical Science (Table 2), and (3) our prior work on analysis of diagnostic statements [27, 28].

Smith and Ceusters stressed that the relation of aboutness includes any portion of reality, rather than being limited to just a single particular or a single universal. A portion of reality (POR) can be a particular, a universal, a relation, or a configuration. A configuration is a combination of particulars and/or universals and certain relation(s) that hold among them.

A representation, then, that is intended to be about a POR but fails in its aboutness because it misrepresents that POR in some way, is misinformation. The sentence *Bob Dylan was in the Beatles* fails to represent not because Bob Dylan or the Beatles did not exist, but

because such a configuration involving Bob Dylan and the Beatles in the way as expressed, never existed. The sentence fails in aboutness on the level of compound expression, but nevertheless is about Bob Dylan and the Beatles individually (on the level of reference) and thus is still an information content entity.

Smith and Ceusters [25] deal more fully with the issue of what it means that a representation is "intended to be about" some entity. Here, we highlight that it follows the doctrine of the "primacy of the intentional" [31], where our written and verbal expressions are to be understood on the basis of the cognitive acts that generated them. That is, a sentence is about that to which its author was directing his or her thoughts when she wrote it.

In addition to Smith and Ceusters' work, we also founded our ontological analysis on the Ontology for General Medical Science or OGMS [30]. This work distinguishes disease, disorder, and diagnosis, and we used definitions from OGMS as starting points for our analysis (Table 2). Note that in OGMS, a diagnosis refers to the existence of a disease of a given type. In clinical medicine, however, diagnoses also refer to (1) disease courses (e.g., acute hepatitis vs. chronic hepatitis), (2) disorders (e.g., fractures and tumors), and (3) the absence of any disease (i.e., a conclusion that a person is healthy also is a diagnosis). It was not our goal to address this issue in this work, as it was not our goal to refine the OGMS definition of diagnosis.

<u>The scenarios</u>

All the scenarios have in common a particular patient, Mr. Adam Jones, who suffers from type 2 diabetes mellitus. Thus in every scenario, there exists Mr. Jones, his disease, the type *Type 2 diabetes mellitus*, the configuration of these three entities (which includes the "bearer of" and "instance of" relationships), and the placement in space and time of this configuration (Figure 1).

Scenario 1: correct diagnosis by physician (ideal case)

Dr. Anne Smith sees Mr. Jones in the office. She takes a history and physical, performs certain laboratory testing, and based on her analysis of the findings, correctly concludes that Mr. Jones has type 2 diabetes mellitus. She subsequently writes her diagnosis in the patient's medical record.

Scenario 2: subsequent correct diagnosis by physician using first diagnosis

A second doctor, Dr. John Brown, sees Mr. Jones in the office at some later date. Mr. Jones has released his records from Dr. Smith to Dr. Brown, who subsequently sees Dr. Smith's diagnosis prior to seeing Mr. Jones. He uses that diagnosis plus his own findings to infer a new clinical picture of Mr. Jones, which he subsequently uses to make another correct diagnosis of Mr. Jones' disease. He writes his diagnosis in Mr. Jones' medical record.

Scenario 3: incorrect diagnosis by physician

Mr. Jones is traveling on vacation, when he falls ill. He sees Dr. Jane Miller who does not have any of his past records available, and thus she is not aware of the previous diagnoses of Drs. Smith or Brown. She infers a new clinical picture of Mr. Jones, and based on it

incorrectly concludes that Mr. Jones has *type 1 diabetes mellitus* (as opposed to type 2). She records a diagnosis of type 1 diabetes mellitus in her medical record for for Mr. Jones.

Scenario #4: coincidentally correct conclusion by layperson (lucky guess)

A friend of Mr. Jones is a "seer". Mr. Jones asks his friend what is in his future. Having no prior knowledge of Mr. Jones medical conditions, the "seer" concludes based on Mr. Jones' horoscope and the position of the moon that he has type 2 diabetes mellitus. He subsequently predicts that Mr. Jones will be hospitalized for his diabetes and miss his daughter's wedding.

Scenario #5: layperson's justifiable conclusion

Mr. Jones' daughter, upon learning of her father's type 2 diabetes mellitus, adds this information into her letter to her brother, writing "Dad has type 2 diabetes mellitus".

Scenario #6: correct diagnosis by computer-based expert system

A medical student is seeing Mr. Jones in the clinic. He performs a history and physical, and types his findings into a diagnostic expert system. The diagnostic expert system infers based on these findings that Mr. Jones has type 2 diabetes mellitus. The medical student writes this diagnosis in Mr. Jones' medical record.

<u>The analysis</u>

Our analysis follows the method of Referent Tracking, which we have found to be a stringent test of ontologies and their definitions [28]. This approach proceeds in three

main steps. First, we systematically identify all the relevant particulars that must exist for the scenario to be true, regardless of whether the scenario explicitly mentions them or only implies their existence. We assign each particular an instance unique identifier (IUI), of the form 'IUI-n', where 'n' is any integer. Second, we identify for each particular the type it instantiates and the temporal interval during which it exists (and assign an identifier of the form *tn* to that interval). Lastly, we identify the relationships that hold between the particulars as well as all relevant relations particulars have to universals other than instantiation, including situations where a particular lacks a given relation to any instance of a certain type (for example, a statement that a patient has had no cough in the last two weeks means that the patient does not stand in the agent_of relation to any instance of the type *Coughing event*, indexed temporally to the two-week interval). [32]

This approach identifies problems in ontologies and their definitions in two major ways. First, it identifies problems that occur when the scenario explicitly rules out the existence of a particular whose existence is implied by an ontological definition (and vice versa). Second, it helps identify exceptions to existing definitions and situations that should not fall under a definition but are erroneously captured by it. Definitions in ontologies can subsequently be adjusted to avoid the errors so identified.

Although our approach is to identify particulars implied by sentences in natural language, the ontological analysis of language and the mechanism(s) by which it makes implicit reference to certain entities is not the focus of this work. Therefore, we convert a sentence like "Mr. Jones has type 2 diabetes mellitus" to Referent Tracking Tuples (e.g., as in Tables 3

through 7) and it is these tuples in which inhere representations that are the objects of our analysis.

To simplify our analysis somewhat, we wrote scenarios under which humans record diagnoses on paper. However, concretization of ICEs also occurs by pixels on monitors, binary switches in memory and processor chips, and magnetic fields on hard disks. But a detailed account of these concretizations and transformations among them is not central to our analysis of what is a diagnosis. Our analysis can be extended to these concretizations without modification of the method.

Results and discussion

In each scenario, Mr. Jones (IUI-1) and his disease (IUI-2) exist, the latter inhering in the former (Table 3). Furthermore, his disease is an instance of the type 'type 2 diabetes mellitus' at any moment in time during which a diagnosis is formulated in any of the scenarios. Mr. Jones (IUI-1) exists through a certain period of time (*t1*) of which we do not know the exact beginning or end. We use temporal identifiers of the form '*t*n' to clearly distinguish such identifiers from IUIs: where IUIs are always intended to be globally and singularly unique, distinct temporal identifiers may denote a unique period of time which is also denoted by another temporal identifier. We also assign an identifier to the time interval during which his disease (IUI-2) exists (*t2*). Diseases usually begin to exist after the organism does, but in the case of congenital genetic diseases, the two intervals might be coextensive. Also, we assume that disease IUI-2 existed at the time of diagnosing, but we

recognize that diagnosing a disease thousands of years after it existed is possible, such as in the case of archaeologists' recent diagnosis of Tutankhamun's malaria [33].

Note that the configuration of organism, disease, and disease type is anchored at a particular location in spacetime, as is the diagnosis. But note also that the diagnosis additionally has an implicit or explicit reference to the location of the configuration in spacetime. To be a correct diagnosis, this reference must also be correct (it has to refer to some part, not necessarily the entirety of spacetime, occupied by the configuration). Thus, for example, to say that Tutankhamun had malaria in 1000 C.E. or today is incorrect, as it would be to say that Mr. Jones had type 2 diabetes mellitus before his parents were born.

***Scenario 1: correct diagnosis*.**

In this scenario, numerous PORs in addition to Mr. Jones and his disease must exist and stand in certain relationships to each other (Tables 4-6). Before Dr. Smith (IUI-3) writes (IUI-13) her diagnosis (IUI-8), there is a cognitive representation (IUI-6) that is concretized in some anatomical part (IUI-5) of her cognitive system (IUI-4). Note that we follow Ceusters and Smith [34] in asserting that all anatomical entities in which cognitive representations inhere are part of a person's cognitive system (that is, any entity used in cognition, including the bearing of cognitive representations, are necessarily within a person's cognitive system) at least during the temporal interval that the cognitive representation exists. If, for example, it would be the case that some white blood cell flowing through some brain capillary would through some of its molecules take part in the concretization of a cognitive representation, then that white blood cell would be part of the

cognitive system at least during the existence of that concretization. It would not anymore be part of the cognitive system once it continues its journey through the body without participating in thought formation. Additionally, Ceusters and Smith take the position (which we also follow) that the cognitive system is not necessarily strictly limited to the brain or even to the entire neurological system of a person: the current state-of-the-art of neuroscience is yet searching for answers to questions such as "what is it in which cognitive representations inhere?" but until it reaches such answers, we remain in our representations agnostic.

IUI-9 denotes the sentence Dr. Smith wrote, as it exists on the particular piece of paper she used to write it on: 'The patient has type 2 diabetes mellitus'. This written statement on paper (IUI-9) bears an information quality entity (IQE, IUI-10) that concretizes her diagnosis (IUI-8). The cognitive representation (IUI-6) and IQE (IUI-10) that concretize the diagnosis are both about the configuration (IUI-7) (the level of compound expression), as well as about Mr. Jones, Mr. Jones' disease, and the universal *Type 2 diabetes mellitus* individually (the level of reference). The cognitive representation (IUI-6) and the diagnosis (IUI-8) are the output of Dr. Smith's diagnostic process (IUI-11), which had as input Dr. Smith's clinical picture (IUI-12) of Mr. Jones. Because the cognitive representation and IQE concretize the same ICE, the latter is conformant to the former (see Table 1).

A correct diagnosis is thus fundamentally an information content entity that is concretized by a representation that stands in an is_about relation to the configuration of an organism, its disease, the relation of inherence between the disease and the organism, a type that the

disease instantiates, and the instantiation relation of the disease to that type, all within a given portion of spacetime (Figure 2). Furthermore, diagnoses are additionally differentiated from other ICEs by the fact that they are generated by a diagnostic process that has a clinical picture as input. We expand further on what constitutes a clinical picture in the next scenario, *Scenario 2*, as well as revisit the diagnostic process briefly in *Scenario 4*, although it was not our objective in this work to develop a fuller account of this process.

Note that it is trivial to state that the particular disease inhering in the organism is an instance of *entity* or even *disease*. Thus, there is an expectation that a diagnosis be as precise (the most specific type) as possible and at a minimal level of granularity that is relevant to treat the patient appropriately and to provide a reasonable prognosis.

Scenario 2*: *second diagnosis.

The second physician, Dr. Brown, makes a second diagnosis at a later point in time, using the first diagnosis in addition to clinical and possibly other findings to infer a new clinical picture of Mr. Jones. With the exception of the configuration of Mr. Jones/his disease/type 2 diabetes mellitus (IUI-7), there is a one-to-one correspondence of PORs as in Scenario 1, numbered IUI-23 through IUI-33 (Additional file 1 : Tables S1-S3). That is, there is no IUI-27 because the configuration is the same POR across scenarios. Similarly, there is no IUI-21 or IUI-22 because IUI-1 and IUI-2 already identify Mr. Jones and his disease, respectively, uniquely.

In this scenario, Dr. Brown (IUI-23) makes a new diagnosis (IUI-28), concretized both by his cognitive representation (IUI-26) in some part (IUI-25) of his cognitive system (IUI-24) and by the IQE (IUI-30) inhering in the sentence in his note (IUI-29). Dr. Smith's previous diagnosis (IUI-8) can be viewed as either (*view1*) being in the aggregate of things that Dr. Brown uses to infer his clinical picture (IUI-32) that serves as input into his diagnostic process (IUI-31), or (*view2*) as something which serves as extra input—alongside his clinical picture—for the diagnostic process. The cognitive representation and the IQE are about the configuration (IUI-7) as well as Mr. Jones (IUI-1), his disease (IUI-2), and type 2 diabetes mellitus (UUI-1).

The current definition of 'clinical picture' in OGMS (see Table 2) seems to conflict with *view1* about this scenario, because the definition seems to exclude using a past diagnosis to infer a clinical picture. Although the current OGMS definition of 'clinical picture' is inclusive of clinical findings, diagnosis as currently defined is not an explicit subtype of clinical finding in OGMS. Furthermore, it is common for clinicians to elicit a previous provider's past diagnosis from the patient or the patient's caregiver during an interview (for example, if Mr. Jones in scenario #2 would have said: 'Dr. Smith says I have type 2 diabetes mellitus'). But the current OGMS definition of 'clinical history' (Table 2) conflicts with this possibility. It refers to health-relevant features of a patient, but features as elucidated by OGMS include only qualities, processes, and physical components of the organism—not dispositions of which disease is a subtype. Therefore, a representation of a disease such as a diagnosis is currently excluded from the OGMS definition of 'clinical history'.

We also note that the OGMS definition of 'clinical picture' is ambiguous in that it is not clear whether it *requires* that laboratory and image findings must always be used to infer a clinical picture, or that they are the only entities that can be used. Regardless, it would be a mistake to do so, because diagnoses can and frequently are made from symptom findings alone. Laboratory and image findings are not necessary components of a clinical picture in reality. Note that a clinical picture can comprise findings of a single type (laboratory alone, pathology image alone, radiology image alone, physical exam finding alone), or even a single finding instance (e.g. Reed-Sternberg cells for a diagnosis of Hodgkin's lymphoma). All these issues are compounded by the fact that the term 'clinical picture' itself is not intuitive.

Given that clinical history taking elicits past diagnoses routinely in clinical medicine, we propose modifying the definition of 'clinical history' to accommodate this reality (bolded sections represent changes to the definition):

clinical history =def. – *A series of statements representing one or more health-relevant features of a patient**, possibly complemented by representations of diseases and configurations**.*

Note that the definition already allows—under the broader heading of 'feature'—representations of disorders (kinds of physical component) and disease courses (kinds of process). Thus, the definition already accommodates these aspects of clinical histories. We

also allow the statements to represent configurations, in line with Smith and Ceusters [2]. These configurations might or might not include various relevant types (for example, "The patient has not participated in any instance of vomiting in the last two weeks."). Finally, note that by using the word 'representing', the definition also accommodates per Smith and Ceusters [2] that some statements might fail in aboutness despite their intention to be about such features. In other words, some statements in the clinical picture might be wrong: for example, a statement that the patient has a disease or pain that she does not in fact have.

To clarify that laboratory and imaging findings are not always required inputs into the diagnostic process, and to capture realistic scenarios compatible with *view2* (for example, Dr. Brown reads Dr. Smith's note in the chart), we also propose a modified definition of 'clinical picture' (changes in bold):

clinical picture =def. – *A representation of a clinical phenotype that is inferred from* ***a*** *combination of,* ***for example, diagnoses and*** *laboratory, image, and clinical findings about a given patient.*

These changes to the definitions of 'clinical history' and 'clinical picture' now properly capture situations where past diagnoses are elicited from the patient and/or her caregiver during a clinical history taking: these diagnoses are now clinical findings in the clinical history that was generated by the clinical history taking (see the definition of 'clinical finding' in Table 2).

Scenario 3: Misdiagnosis.

The third physician, Dr. Miller, misdiagnoses Mr. Jones' type 2 diabetes mellitus as type 1 diabetes mellitus (Figure 3). Per Smith and Ceusters, because the misdiagnosis is still about Mr. Jones, his disease, the relationship between them, and the type 'type 1 diabetes mellitus' on the level of reference, it is an information content entity. However, it fails to be about the configuration IUI-7 as a whole on the level of compound expression.

Again, in this scenario there exist PORs in one-to-one correspondence (except the configuration and its components) numbered IUI-43 through IUI-53 (Additional file 2 : Tables S4-S6). Dr. Miller (IUI-43) writes (IUI-53) his misdiagnosis (IUI-48) in Mr. Jones' chart, and the IQE (IUI-50) inhering in the ink (IUI-49) is conformant to his cognitive representation (IUI-46), and both are about—on the level of reference—Mr. Jones, his disease, and type 1 diabetes mellitus. But neither one is about the configuration (IUI-7). To capture the relation both (1) between the cognitive representation and the configuration and (2) between the IQE and the configuration, we define a new relation:

> **is-misrepresentation-of**: domain: representation, range: portion of reality.
>
> Def: *x* is-misrepresentation of *y* iif *x* is a representation and *x* is intended to be about *y* and it is not the case that *x* is about *y*.

Then we assert that the representations (IUI-46 and IUI-50) are misrepresentations of the configuration (Table 7 and Additional file 2 : Table S6). Note that our definition precludes the cognitive representation (IUI-46) and IQE (IUI-50) being about any configuration other

than IUI-7, because they are not intended to be about, for example, the configuration of the sun, earth, and moon at a particular date and time.

Note that asserting the incorrect disease type is not the only way to make a misdiagnosis. There are at least six possibilities where a diagnosis fails to be about a configuration on the level of compound expression (Table 8). If a representation fails on the level of reference, it also fails on the level of compound expressions, because a configuration cannot consist of that which does not exist. These six possibilities could also exist in combination, but if the 2nd, 3rd, and 4th possibilities are all present (for example, "Ron Weasley has spattergroit"), then there is not a diagnosis, or even any information content entity at all, because the representation is not about anything even on the level of reference. Of course, if the organism itself does not exist, then there cannot be a clinical picture inferred, and thus it would not be a diagnosis or misdiagnosis, although it could still be an ICE if it is about a really-existing disease type (for example, "James Bond has influenza"). Also, as medical knowledge evolves, the profession comes to understand that certain types of disease thought to exist in fact do not. Thus past diagnoses of *dropsy* and *consumption* we now understand to be misdiagnoses.

Despite searching the extensive literature on diagnostic error, we could not find any studies that looked at what percentages of misdiagnoses fall into these categories. We conjecture based on our past clinical expertise and experience that asserting the incorrect disease type is the most common mistake among those in Table 8, but confirmation or rejection of this conjecture requires study.

Scenario 4: the lucky guess.

In this scenario, a layperson (the "seer"—IUI 63) correctly concluded coincidentally that Mr. Jones had type 2 diabetes mellitus based on the position of the moon and Mr. Jones' horoscope (Additional file 3 : Tables S7-S9). It would be wrong to say the seer's reasoning (IUI-71) constituted a diagnostic process. To avoid coincidentally correct statements from qualifying as diagnoses, we additionally require as input into the diagnostic process cognitive representations of the disease type and the types instantiated by the sequalae, signs, symptoms, and any clinical, laboratory, or imaging findings or phenotypes of the instances of this disease type. Note that this is a minimal requirement: clinicians often additionally include in their diagnostic reasoning cognitive representations of other disease types and associated PORs when considering alternative possibilities for the disease type.

This view is based on the extensive literature on clinical reasoning processes, especially diagnosis (for a review, see Norman [35]). This research has established the use of representations, called 'knowledge structures', in the diagnostic process. The nature and form of these representations evolves as clinical expertise develops [36], and we note that the differences in diagnostic processes that result could result in a typology of diagnostic processes in OGMS.

Because the seer had no cognitive representations of type 2 diabetes mellitus, let alone used them as input into his "reasoning", his conclusion (IUI-68), although an ICE, is not a

diagnosis. Similarly, if a physician makes a lucky guess based not on his cognitive representations of the stated disease type but instead by flipping a coin or some such, that too would not be a diagnosis.

To Table 3 we add an aggregate of cognitive representations of disease types and associated entities as input into the diagnostic process (Table 9).

We propose to redefine diagnostic process as follows:

Diagnostic process =def. *An interpretive PROCESS that has as input (1) a CLINICAL PICTURE of a given patient* ***AND (2) an aggregate of REPRESENTATIONs of at least one type of disease and at least one type of phenotype whose instances are associated with instances of that disease,*** *and as output an assertion to the effect that the patient has a DISEASE of a certain type.*

Scenario 5: layperson's justifiable conclusion.

Mr. Jones' daughter wrote a sentence in her letter to her brother based on reading Dr. Smith's progress note saying that that her father has type 2 diabetes mellitus (Additional file 4 : Tables S10-S12). Of course, the daughter has not made a diagnosis. She is communicating to her brother what she believes to be the case.

Had she merely written "Dr. Smith says" and then copied Dr. Smith's sentence word for word into her letter, then her writing would concretize Dr. Smith's diagnosis (IUI-8). This is the case of hearsay ("so-and-so said it was the case that…").

As Smith and Ceusters showed, however, the same sentence written by two different people does not guarantee they concretize the same ICE. ICEs are further differentiated by the provenance of their concretizations, including who created them and when, and to what POR they intend to be about. In their example, two people writing the sentence *Barack Obama has never been President of the United States*—one before and one after Obama's inauguration as President—generate two different ICEs. The one written after fails on the level of compound expressions but not on the level of reference, whereas the one written before succeeds on both levels (it remains true that at the time when the sentence was written, he had never been President).

We therefore distinguish between a human (1) merely copying a representation, in which case the copy concretizes the same ICE as the original text and (2) creating her own cognitive representation of the POR—which involves forming a belief that the POR really existed as represented—and then subsequently creating an IQE that is conformant to the cognitive representation. In the former case, a new ICE does not come into being. It does not even require in the cognitive system of the copier any representation of the POR that the original representation is about (as in the case of copying German text that one does not understand at all). In the latter case, by contrast, a new ICE does come into being.

In Scenario 5, the daughter did not merely repeat Dr. Smith's diagnosis. She communicated to her brother *her* belief about her father's disease. She deliberately chose not to merely convey Dr. Smith's diagnosis, but rather her belief that her father has type 2 diabetes

mellitus. She heard the opinion of an expert, in whom she had trust. Based on (1) her observations of her father, (2) Dr. Smith's diagnosis, and (3) her trust in Dr. Smith, she reached the conclusion herself that her father suffers from type 2 diabetes mellitus. Because she did not begin with a clinical picture and her own cognitive representations of type 2 diabetes mellitus, her conclusion is not a diagnosis.

However, consider the scenario where she is given the clinical picture and has enough knowledge to arrive at a conclusion, which could be the case either if she were a physician or somehow other acquired or were given the necessary knowledge: it is analagous to Scenario #6, where she takes the place of the expert system (see analysis of that scenario below). Thus, here in Scenario #5 it is important to note that she did not reason from a clinical picture to the diagnosis.

In this scenario, therefore, the daughter has created a new ICE (IUI-88) that is not a diagnosis. She has concretized it in the sentence (IUI-89) in her letter.

Scenario 6: diagnosis by non-human.

The diagnostic decision support system has made a diagnosis (or misdiagnosis depending on whether it is correct), because it (1) takes as input a clinical picture and representations of the relevant disease type and one or more types of phenotypes with which it is associated; (2) participates in a process of making a conclusion based on this input; and (3) outputs from this process a statement about a configuration involving an organism, a disease, and a disease type.

In this case, there are no cognitive representations. In their place are digital representations on hard drives, memory chips, and central processing units. If we assume the system generates a sentence and prints it on paper, then we have an analagous IQE to the written diagnosis of the physician and ICE of the sister.

Nothing in our proposed definitions conflicts with this scenario. Replacing Dr. Smith and associated representations and diagnostic process with various components of the computer and its digital representations as well as inferential process (which is an instance of diagnostic process) is straightforward.

Returning briefly to a point made in Scenario #5, Mr. Jones' daughter could follow the exact same algorithm(s) of the diagnostic expert system using the exact same clinical picture as input, and she would arrive at (or make) a diagnosis, in contrast to scenario #5 where her conclusion was an ICE but not a diagnosis.

Conclusions

We applied Smith and Ceusters' results on aboutness [25] to diagnosis in order to develop an account of diagnosis, misdiagnosis, lucky guesses, hearsay, a layperson's justified belief about disease configurations, and a diagnosis made by an expert system. Our key result is that a correct diagnosis, as defined by OGMS, is about a configuration of an organism, its disease, and the type the disease instantiates (level of compound expression) in a specified portion of spacetime. We identified several subtypes of misdiagnosis (e.g., wrong disease

subtype, wrong patient, wrong temporal placement) that have not been differentiated in the literature on diagnostic error, to our knowledge. Studying the incidence and causes of these subtypes might advance the study of diagnostic error and strategies to reduce it. Note that as we have defined it, 'misdiagnosis' does not refer to the diagnostic errors of absent diagnosis (failing to diagnose a disease at all, let alone incorrectly) and delayed diagnosis. Lastly, we note that the current literature on diagnostic error, per a 2016 Institute of Medicine report, does not lend itself to generating reliable estimates of incidence of diagnostic error per se, let alone any subtype of such error [37].

Although misdiagnoses involving non-existence of certain entities might at first seem to be of minor importance, we highlight two cases where non-existence is relevant. First, in the case where the type of disease does not exist (consider past diagnoses of "dropsy"), it could well be that our understanding of disease decades from now is much more advanced, and what we think are types of disease today in fact are not. So just as with past diagnoses of "dropsy", it could be that today's diagnoses of "schizophrenia" are misdiagnoses merely by referring to a type that does not exist. Second, in the case where the instance of disease does not exist, we consider two scenarios. The first scenario involves past diagnoses of mental illness where neither the instance nor the type exists. For example, past diagnoses of runaway slaves as having "drapetomania" involved neither a really existing instance nor a really existing type of disease. The second scenario involves patients with hypochondria or who are malingering. They feign a condition for which the unassuming practitioner mistakenly asserts both the existence of an instance and a type.

Our results and typology of misdiagnosis could serve as the beginnings of a formal framework for studying diagnostic error as a component of data quality in EHRs and research data collections, in response to the call by Weiskopf and Weng for more formal, generalizable, and validated methods for assessing data quality [24]. Applying Ceusters' detailed typology of mistakes in ontology (e.g., asserting a type that does not exist) [38] and referent tracking systems (assigning an identifier but there is no corresponding particular that it identifies, assigning one identifier to two particulars, assigning two identifiers to one particular, etc.) [39] to diagnosis could build on our work here to build out such a framework. It remains future work to do so.

The provenance of the ICE and its concretizations are critical: lucky guesses, hearsay, and laypersons' conclusions about disease (when not arrived at through a diagnostic process using a clinical picture and cognitive representations of the associated type(s) of disease as input) do not constitute diagnoses and therefore are different types of ICE than diagnoses. Provenance also includes which findings and other information constituted the clinical picture used in the diagnostic process. Our analysis of the scenarios identified past diagnoses as important input into the diagnostic process, leading to proposed redefinitions of 'clinical history', 'clinical picture', and 'diagnostic process' for OGMS.

Smith and Ceusters' results on aboutness and our extension of them here to diagnosis reduce the need for the workarounds reported by Martínez Costa and Schulz [26] and Hastings et al. [27] It is perfectly legitimate to relate 'suspected heart failure finding' to 'congestive heart failure' with an existential quantifier: if an instance of this type is not

about a really-existing configuration of patient–disease–heart failure, it is still an ICE that is individually about the patient, her condition, and the type *heart failure* on the level of reference. In OWL, we could assert:

Suspected heart failure ICE -> ICE and (**is about** SOME *Organism*)

Suspected heart failure ICE -> ICE and (**is about** SOME *Condition*)

In more expressive formalisms including first-order logic, we could also assert that it is about the type *heart failure*, where 'Type', 'Instance_of', and 'Is_about' are predicates in what follows, where the universal quantification applies to the ICE, not what it is about:

Type(heart_failure)

Type(suspected_heart_failure_ICE)

$\forall x$ (Instance_of(x, suspected_heart_failure_ICE) -> Is_about(x, heart_failure))

Similarly, chemical graphs and diagrams are ICEs about individual types of atoms such as carbon, oxygen, hydrogen, and so on, even when they fail to be about any type of configuration (molecule) of such atoms. However, because they are typically not about any instances, proper existential quantification in OWL is not possible. However, we could relate in first-order logic the diagram of *octaazacubane* (a hypothetical molecule which would be comprised of eight nitrogen atoms arranged in a cubic structure) to the *nitrogen* type of atom using existential quantification (again where the universal quantification in what follows applies to the ICE and not what it is about):

Type(nitrogen_atom)

Type(octaazacubane _diagram)

$\forall x$ (Instance_of(x, octaazacubane_diagram) -> Is_about(x, nitrogen_atom))

It is therefore not required to use universal quantification over the range of things that an ICE is about, when relating ICEs to those entities they are about, to avoid failure of aboutness on the level of compound expression. This result is qualified by the constraints of representational formalisms such as OWL that prevent directly asserting aboutness to types. Schulz et al. describe workarounds in OWL to asserting aboutness to types, that may be of benefit in some use cases [40].

The use of universal quantification actually introduces problems when we account for aboutness on the level of individual reference. For example, if we leave the 'suspected heart failure finding' of Martínez Costa and Schulz as being *only* about 'congestive heart failure', then it would result in a contradiction to say that it is about some organism. Likewise for condition. So use of the universal quantifier precludes aboutness on the level of individual reference, in direct conflict with the results of Smith and Ceusters on misinformation.

Our analysis also identified problems with, and suggested improvements to, the definitions of core terms from the Ontology for General Medical Science including 'diagnostic process' and 'clinical picture'. This result is consistent with our past work, where we have found the method of referent tracking analysis to be a stringent test of definitions in ontologies.

This work is limited by the fact that we did not conduct further ontological analysis of the diagnostic process beyond OGMS and beyond what our scenarios required, as this was not the purpose of the present work. We do note that our requirement for including cognitive

representations of disease types as input into the diagnostic process is based on this literature, however. Engaging experts in the study of clinical reasoning in future work to develop a typology of diagnostic processes has the potential to result in a corresponding typology of diagnoses.

Future work includes (1) an account of differential diagnosis, where a clinician or expert system generates a list of likely types of disease for further investigation to identify the actual type the organism's disease instantiates; (2) proposing to the OGMS community to clarify the definitions of 'clinical history', 'clinical picture', and 'diagnostic process' as suggested here, and to expand the definition of diagnosis to include disorders, disease courses, and absence of disease (i.e., healthy); (3) extending our analysis as reported here to this expanded definition of 'diagnosis'; and (4) conducting deeper ontological analysis of the diagnostic process, in coordination with experts in the study of clinical reasoning.

List of Abbreviations

BFO	Basic Formal Ontology
GDC	Generically dependent continuant
IAO	Information Artifact Ontology
ICE	Information Content Entity
IQE	Information Quality Entity
OGMS	Ontology for General Medical Science
POR	Portion of Reality
RT	Referent Tracking

RTT Referent Tracking Tuple

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

The authors contributed equally to the ontological analysis and development of results. Author WRH created the first version of the manuscript. Both authors had full access to all materials and analysis and participated in revising the manuscript. Both authors approved the final version of the manuscript.

Acknowledgments

This work was suppoted in part by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1TR001427.

References

1. Bingham CO, 3rd, Bartlett SJ, Merkel PA, Mielenz TJ, Pilkonis PA, Edmundson L, Moore E, Sabharwal RK: **Using patient-reported outcomes and PROMIS in research and clinical applications: experiences from the PCORI pilot projects**. *Qual Life Res* 2016.
2. Scanlon L: **PatientsLikeMe Survey Shows Vast Majority of People With Health Conditions Are Willing To Share Their Health Data**. In. Cambridge, Mass: PatientsLikeMe; 2014.
3. Rudin RS, Motala A, Goldzweig CL, Shekelle PG: **Usage and effect of health information exchange: a systematic review**. *Ann Intern Med* 2014, **161**(11):803-811.
4. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P: **From the Office of the National Coordinator: the strategy for advancing the exchange of health information**. *Health Aff (Millwood)* 2012, **31**(3):527-536.

5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS: **Launching PCORnet, a national patient-centered clinical research network**. *J Am Med Inform Assoc* 2014, **21**(4):578-582.
6. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM *et al*: **The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies**. *BMC Med Genomics* 2011, **4**:13.
7. Owens B: **DATA SHARING. Montreal institute going 'open' to accelerate science**. *Science* 2016, **351**(6271):329.
8. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF: **Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors**. *Med Care* 2005, **43**(5):480-485.
9. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM: **Measuring diagnoses: ICD code accuracy**. *Health Serv Res* 2005, **40**(5 Pt 2):1620-1639.
10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ *et al*: **Caveats for the use of operational electronic health record data in comparative effectiveness research**. *Med Care* 2013, **51**(8 Suppl 3):S30-37.
11. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS: **Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied**. *Med Care* 2013, **51**(8 Suppl 3):S80-86.
12. Botsis T, Hartvigsen G, Chen F, Weng C: **Secondary Use of EHR: Data Quality Issues and Informatics Opportunities**. *AMIA Jt Summits Transl Sci Proc* 2010, **2010**:1-5.
13. Benesch C, Witter DM, Jr., Wilder AL, Duncan PW, Samsa GP, Matchar DB: **Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease**. *Neurology* 1997, **49**(3):660-664.
14. Shapiro M, Johnston D, Wald J, Mon D: **Patient-Generated Health Data White Paper**. RTI International, Research Triangle Park, NC 27709; 2012.
15. Gordon NP, Mellor RG: **Accuracy of parent-reported information for estimating prevalence of overweight and obesity in a race-ethnically diverse pediatric clinic population aged 3 to 12**. *BMC pediatrics* 2015, **15**(1):5.
16. Komaroff AL: **The variability and inaccuracy of medical data**. *Proceedings of the IEEE* 1979, **67**(9):1196-1296.
17. Callahan CM, Tu W, Stump TE, Clark DO, Unroe KT, Hendrie HC: **Errors in self-reports of health services use: impact on alzheimer disease clinical trial designs**. *Alzheimer Dis Assoc Disord* 2015, **29**(1):75-81.
18. Monte AA, Heard KJ, Hoppe JA, Vasiliou V, Gonzalez FJ: **The accuracy of self-reported drug ingestion histories in emergency department patients**. *J Clin Pharmacol* 2015, **55**(1):33-38.
19. Gerritsen M, Berndt N, Lechner L, de Vries H, Mudde A, Bolman C: **Self-Reporting of Smoking Cessation in Cardiac Patients: How Reliable Is It and Is Reliability Associated With Patient Characteristics?** *J Addict Med* 2015, **9**(4):308-316.

20. Raphael KG, Janal MN, Sirois DA, Dubrovsky B, Klausner JJ, Krieger AC, Lavigne GJ: **Validity of self-reported sleep bruxism among myofascial temporomandibular disorder patients and controls**. *J Oral Rehabil* 2015, **42**(10):751-758.
21. Patel M, Perrin K, Pritchard A, Williams M, Wijesinghe M, Weatherall M, Beasley R: **Accuracy of patient self-report as a measure of inhaled asthma medication use**. *Respirology* 2013, **18**(3):546-552.
22. Woodfield R, Group UKBSO, Follow-up UKB, Outcomes Working G, Sudlow CL: **Accuracy of Patient Self-Report of Stroke: A Systematic Review from the UK Biobank Stroke Outcomes Group**. *PLoS One* 2015, **10**(9):e0137538.
23. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ: **How the provenance of electronic health record data matters for research: a case example using system mapping**. *EGEMS (Wash DC)* 2014, **2**(1):1058.
24. Weiskopf NG, Weng C: **Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research**. *J Am Med Inform Assoc* 2013, **20**(1):144-151.
25. Smith B, Ceusters W: **Aboutness: Towards Foundations for the Information Artifact Ontology**. In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015.
26. Martínez-Costa C, Schulz S: **Ontology-based reinterpretation of the SNOMED CT context model**. In: *Proceedings of the Fourth International Conference on Biomedical Ontology: July 7th-12th, 2013; Montreal*: Edited by Dumontier M, Hoehndorf R, Baker CJO. 2013: 90-95.
27. Hastings J, Batchelor C, Neuhaus F, Steinbeck C: **What's in an 'is about' link? Chemical diagrams and the information artifact ontology**. In: *Proceedings of the 2nd International Conference on Biomedical Ontology; Buffalo, New York*: Edited by Bodenreider O, Martone ME, Ruttenberg A. 2011: 201-208.
28. Ceusters W, Hogan WR: **An ontological analysis of diagnostic assertions in electronic healthcare records** In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015.
29. Hogan WR: **To what entities does an ICD-9-CM code refer? A realist approach.** In: *Bio-ontologies; Boston, MA*: Edited by Shah N, Sansone S-A, Stephens S, Soldatova L. 2010.
30. Scheuermann RH, Ceusters W, Smith B: **Toward an ontological treatment of disease and diagnosis**. In: *AMIA Summit on Translational Bioinformatics: 2009*. 116-120.
31. Chisholm RM: **The primacy of the intentional**. *Synthese*, **61**(1):89-109.
32. Ceusters W, Elkin P, Smith B: **Negative findings in electronic health records and biomedical ontologies: a realist approach**. *Int J Med Inform* 2007, **76 Suppl 3**:S326-333.
33. Hawass Z, Gad YZ, Ismail S, Khairat R, Fathalla D, Hasan N, Ahmed A, Elleithy H, Ball M, Gaballah F *et al*: **Ancestry and pathology in King Tutankhamun's family**. *JAMA* 2010, **303**(7):638-647.
34. Ceusters W, Smith B: **Foundations for a realist ontology of mental disease**. *J Biomed Semantics* 2010, **1**(1):10.
35. Norman G: **Research in clinical reasoning: past history and current trends**. *Medical Education* 2005, **39**(4):418-427.

36. Schmidt HG, Rikers RMJP: **How expertise develops in medicine: knowledge encapsulation and illness script formation**. *Medical Education* 2007, **41**(12):1133-1139.
37. Erin P. Balogh, Bryan T. Miller, Ball JR: **Improving Diagnosis in Health Care**. In. Edited by Board on Health Care Services, Medicine Io. Washington, DC: The National Academies Press; 2015.
38. Ceusters W: **Towards A realism-based metric for quality assurance in ontology matching**. In: *Proceedings of Formal Ontology in Information Systems 2006 (FOIS 2006)*: Edited by Bennett B, Fellbaum C. IOS Press 2006: 321.
39. Ceusters W: **Dealing with mistakes in a referent tracking system**. In: *Ontology for the Intelligence Community (OIC-2007); Columbia, Maryland*. 2007: 5-8.
40. Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A: **An Ontological Analysis of Reference in Health Record Statements**. In: *Formal Ontology in Information Systems: Proceedings of the Eighth International Conference (FOIS 2014): 2014*. IOS Press: 289.

Table 1. Definitions based on Smith and Ceusters [25].

Term	**Definition**
INFORMATION CONTENT ENTITY	An ENTITY which is (1) GENERICALLY DEPENDENT on (2) some MATERIAL ENTITY and which is (3) concretized by a QUALITY (a) inhering in the MATERIAL ENTITY and (b) that is_about some PORTION OF REALITY
INFORMATION QUALITY ENTITY	A REPRESENTATION that is the concretization of some INFORMATION CONTENT ENTITY
REPRESENTATION	A QUALITY which is_about or is intended to be about a PORTION OF REALITY
MENTAL QUALITY	A QUALITY which specifically depends on an ANATOMICAL STRUCTURE in the cognitive system of an organism
COGNITIVE REPRESENTATION	A REPRESENTATION which is a MENTAL QUALITY
Relation	**Explanation**
x ***is_about*** *y*	*x refers to or is cognitively directed towards y*. **Domain:** representations; **Range:** portions of reality
x ***concretizes*** *y*	*x* is a QUALITY and *y* is a GENERICALLY DEPENDENT CONTINUANT (GDC) and for some MATERIAL ENTITY *z*, *x* ***specifically_depends_on*** z at *t* and *y* ***generically_depends_on*** *z* at *t*, and if *y* migrates from bearer *z* to another bearer *w* then a copy of *x* will be created in *w*.
x ***is_conformant_to*** *y*	=def. *x* is an INFORMATION QUALITY ENTITY and *y* is a COGNITIVE REPRESENTATION and there is some GDC *g* such that *x* ***concretizes*** *g* and *y* ***concretizes*** *g*.

Table 2. Key definitions from OGMS used in the analysis

Term	Definition
DISEASE	A DISPOSITION (i) to undergo PATHOLOGICAL PROCESSes that (ii) exists in an ORGANISM because of one or more DISORDERs in that ORGANISM.
DISORDER	A causally relatively isolated combination of physical components that is (a) clinically abnormal and (b) maximal, in the sense that it is not a part of some larger such combination.
DIAGNOSIS	A conclusion of an interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion (diagnostic statement) to the effect that the patient has a DISEASE of such and such a type.
DIAGNOSTIC PROCESS	An interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion to the effect that the patient has a DISEASE of a certain type.
PATHOLOGICAL PROCESS	A bodily PROCESS that is a manifestation of a DISORDER.
PHENOTYPE	A bodily feature or combination of bodily features of an organism determined by the interaction of the genetic make-up of the organism and its environment.
CLINICAL PHENOTYPE	A clinically abnormal PHENOTYPE.
CLINICAL PICTURE	A representation of a CLINICAL PHENOTYPE that is inferred from the combination of laboratory, image and clinical findings about a given patient.
CLINICAL FINDING	A REPRESENTATION that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof.
MANIFESTATION OF DISEASE	A QUALITY of a patient that is (a) a deviation from clinical normality that exists in virtue of the realization of a disease and (b) is observable.
CLINICAL HISTORY TAKING	An interview in which a clinician elicits a clinical history from a patient or from a third party who is authorized to make health care decisions on behalf of the patient.
CLINICAL HISTORY	A series of statements representing health-relevant features of a patient.

Table 3. Referent tracking tuples true in every scenario

IUI	**Entity**	**Existence period**	**Type**	**Notes**
IUI-1	Mr. Adam Jones	*t1* – the period during which IUI-1 exists	Material Entity	
IUI-2	IUI-1's disease	*t2*	Disposition	
Relationships among particulars				**Notes**
IUI-2	**inheres in**	IUI-1	at *t2*	
IUI-2	**instance of**	UUI-1	at *t2*	UUI-1 is a universal unique identifier that denotes *type 2 diabetes mellitus.* We assume that if something is at any time of its existence an instance of type 2 DM, it is instance of type 2 DM at all times it exists.

Table 4. The entities in Scenario 1

IUI	Entity	Existence period	Type	Notes
IUI-3	Dr. Anne Smith	t3	Human being	
IUI-4	Cognitive system of IUI-3	t4		
IUI-5	An anatomical entity that is part of IUI-4	t5	Anatomical entity	Which anatomical entity and its lifetime cannot be easily specified given current state of neuroscience.
IUI-6	Quality that inheres in IUI-5 and is about IUI-7	t6	Cognitive representation	
IUI-7	The POR that is truth-maker for IUI-8	t7	Configuration	Mr. Jones, his disease, their relationship, and disease's instantiation
IUI-8	Dr. Smith's diagnosis	t8	Diagnosis	ICE concretized by IUI-6 and IUI-10
IUI-9	That which is written down on paper and forms the sentence.	t9	Material entity	*I conclude therefore that Mr. Jones has type 2 diabetes mellitus.*
IUI-10	IQE that inheres in IUI-9.	t10	Information quality entity	The sentence began to exist as soon as ink was laid down on paper, but the IQE did not begin to exist until the sentence was finished.
IUI-11	Dr. Smith's interpretive process	occupies t11	Diagnostic process	Dr. Smith's diagnostic process that led to her diagnosis IUI-8
IUI-12	The clinical picture input into IUI-11	t12	Clinical picture	Dr. Smith's clinical picture as ascertained prior to t6
IUI-13	Dr. Smith writing her diagnosis in the note	occupies t13	Process	

Table 5. Additional temporal entities in Scenario 1.

Temporal identifier	Description	Notes
t14	The interval during which the anatomical entity (IUI-5) is part of the cognitive system (IUI-4)	This interval is not easily specified given the current state of neuroscience. It could be different than t3 and t4.
t15	The interval during which the clinical picture (IUI-12) is used in the interpretive process (IUI-11)	Could be shorter than t11
t16	The point in time at which the cognitive representation (IUI-6) and diagnosis (IUI-8) begin to exist	t16 ends t11. Because the ICE does not exist until the cognitive representation—its first concretization—exists, this is also the point in time at which the diagnosis begins to exist.
t17	The interval during which the cognitive representation (IUI-6) participates in the writing process (IUI-13)	
t18	The interval during which the diagnosis (IUI-8) participates in the writing process (IUI-13)	It is possible that the original cognitive representation (IUI-6) gets copied elsewhere in the brain for reasoning and thus that the ICE continues to participate after the initial cognitive representation
t19	The interval during which that which is written on paper (IUI-10) begins to exist until it exists in full	The writing process begins earlier than the time at which the sentence begins to exist: the author starts the process with getting a pen and paper, any preparation necessary ("clicking" the pen), etc.

Table 6. Relationships among particulars in Scenario 1.

IUI	Relation	IUI	When relation holds in reality	Notes
IUI-4	**part of**	IUI-3	at t4	
IUI-5	**part of**	IUI-4	at t14	All anatomical components in which the cognitive representation inheres are part of the cognitive system. We do not assume the cognitive system is limited to the brain, as the state of neuroscience does not permit such an assumption.
IUI-6	**inheres in**	IUI-5	at t6	
IUI-6	**is about**	IUI-7	at t6	The cognitive representation stands in aboutness to IUI-7 as long as it exists
IUI-6	**is about**	IUI-1	at t6	It is also about Mr. Jones
IUI-6	**is about**	IUI-2	at t6	And about Mr. Jones' disease
IUI-6	**is about**	UUI-1	at t6	And about Type 2 diabetes mellitus
IUI-6	**concretizes**	IUI-8	at t6	It also concretizes the diagnosis
IUI-10	**inheres in**	IUI-9	at t9	The IQE inheres in the sentence on paper
IUI-10	**is about**	IUI-7	at t10	The IQE stands in aboutness to IUI-7
IUI-10	**is about**	IUI-1	at t10	It is also about Mr. Jones
IUI-10	**is about**	IUI-2	at t10	And about Mr. Jones' disease
IUI-10	**is about**	UUI-1	at t10	And about Type 2 diabetes mellitus
IUI-10	**concretizes**	IUI-8	at t10	
IUI-10	**is conformant to**	IUI-6	at t10	Is conformant to the cognitive representation as long as it exists
IUI-3	**agent in**	IUI-11	at t11	
IUI-12	**input into**	IUI-11	at t15	Clinical picture input into IUI-11
IUI-6	**output of**	IUI-11	at t16	Cognitive representation output from IUI-11
IUI-8	**output of**	IUI-11	at t16	Both the diagnosis and its concretization are outputs of IUI-11
IUI-8	**input into**	IUI-13	at t17	The diagnosis is input into writing
IUI-6	**input into**	IUI-13	at t18	As is its cognitive representation
IUI-10	**output of**	IUI-13	at t19	The sentence is output of writing

Table 7. Relationships of representations to portions of reality in Scenario 3: *Incorrect diagnosis*.

Relationships among particulars				Notes
IUI-46	**is about**	IUI-1	at t46	Dr. Jane Miller's cognitive representation is about Mr. Jones
IUI-46	**is about**	IUI-2	at t46	And Mr. Jones' disease
IUI-46	**is about**	UUI-2	at t46	And Type 1 diabetes mellitus (denoted by UUI-2)
IUI-50	**is about**	IUI-1	at t50	Likewise with the IQE inhering in the ink on paper
IUI-50	**is about**	IUI-2	at t50	
IUI-50	**is about**	UUI-2	at t50	
IUI-46	**is misrepresentation of**	IUI-7	at t46	But the cognitive representation is a misrepresentation of the configuration, i.e., it is intended to be about the configuration but fails on the level of compound expression
IUI-50	**is misrepresentation of**	IUI-7	at t50	The same is true of the IQE

Table 8. Six possibilities for a diagnosis failing in aboutness on the level of compound expressions.

Problem	**Where it fails *first***	**Description**
Noninstantation, asserted type exists	Level of compound expression	Disease instantiates a different type than the stated type, but the stated type exists
Noninstantation, asserted type does not exist	Level of reference	Disease instantiates a different type than stated, while the stated type of disease does not exist
Disease nonexistence	Level of reference	The disease instance does not exist
Organism nonexistence	Level of reference	The organism instance does not exist. In this case, there could not be a clinical picture properly inferred and thus it is not a misdiagnosis although it could still be an ICE.
Disease non-inherence	Level of compound expression	The disease inheres in a different organism than the one stated. For example, the doctor mistakenly ascribes Mr. Johnson's hypertension to his twin.
Configuration is not located in that part of spacetime where the diagnosis says it is located.	Level of compound expression	A diagnosis of type 2 diabetes mellitus 5 years ago is wrong because the patient didn't have the disease at that time, even though the patient has type 2 diabetes today. Also, a diagnosis that the patient has an upper respiratory tract infection today when in reality the infection resolved two weeks ago.

Table 9. Additional tuples required to distinguish diagnosing from a lucky guess.

IUI	Entity	Lifetime	Type	Notes
IUI-14	The aggregate of Dr. Smith's cognitive representations of various disease types and their associated types of phenotypes including type 2 diabetes mellitus that he used in the diagnostic process	t20	Aggregate of cognitive representations	
Relationships among particulars				**Notes**
IUI-14	**input into**	IUI-11	at t21	t21 refers to the temporal interval during which IUI-14 participated in the reasoning process. It could start at the same time as t11 or after t11, and end at the same time as or before t11.

Figure legends

Figure 1. The configuration of Mr. Jones, his disease, and type 2 diabetes mellitus

Figure 2. Diagram of diagnostic process, its inputs, a correct diagnosis, its concretization, and the configuration that that the concretization is about

Figure 3. Misdiagnosis of type of disease. The diagnosis is individually about the patient, the disease, and the incorrectly diagnosed disease type Y, but it is not about the configuration of patient, disease, and disease type X.

Figure 1

Click here to download Figure Fig 1.png

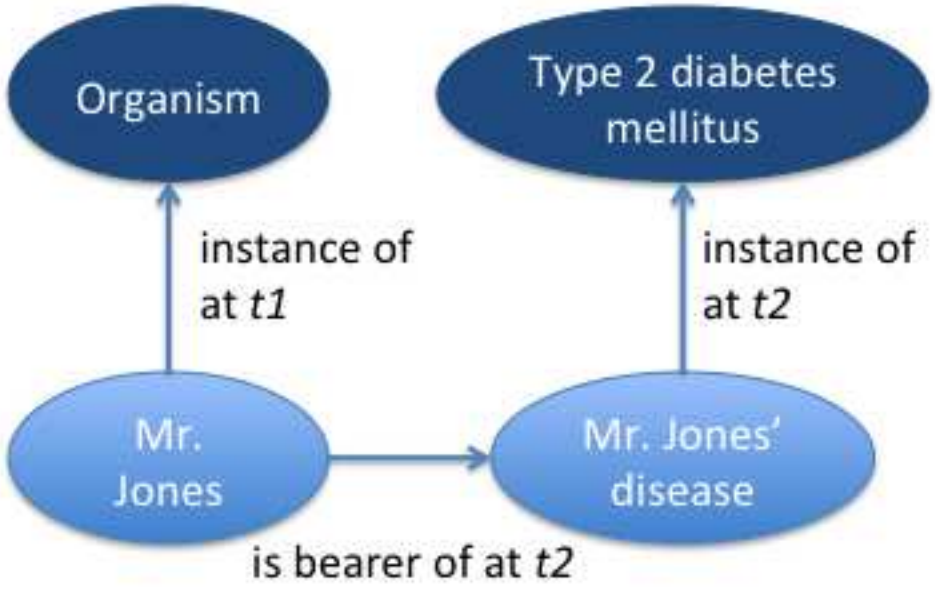


Figure 2

Click here to download Figure Fig 2.png

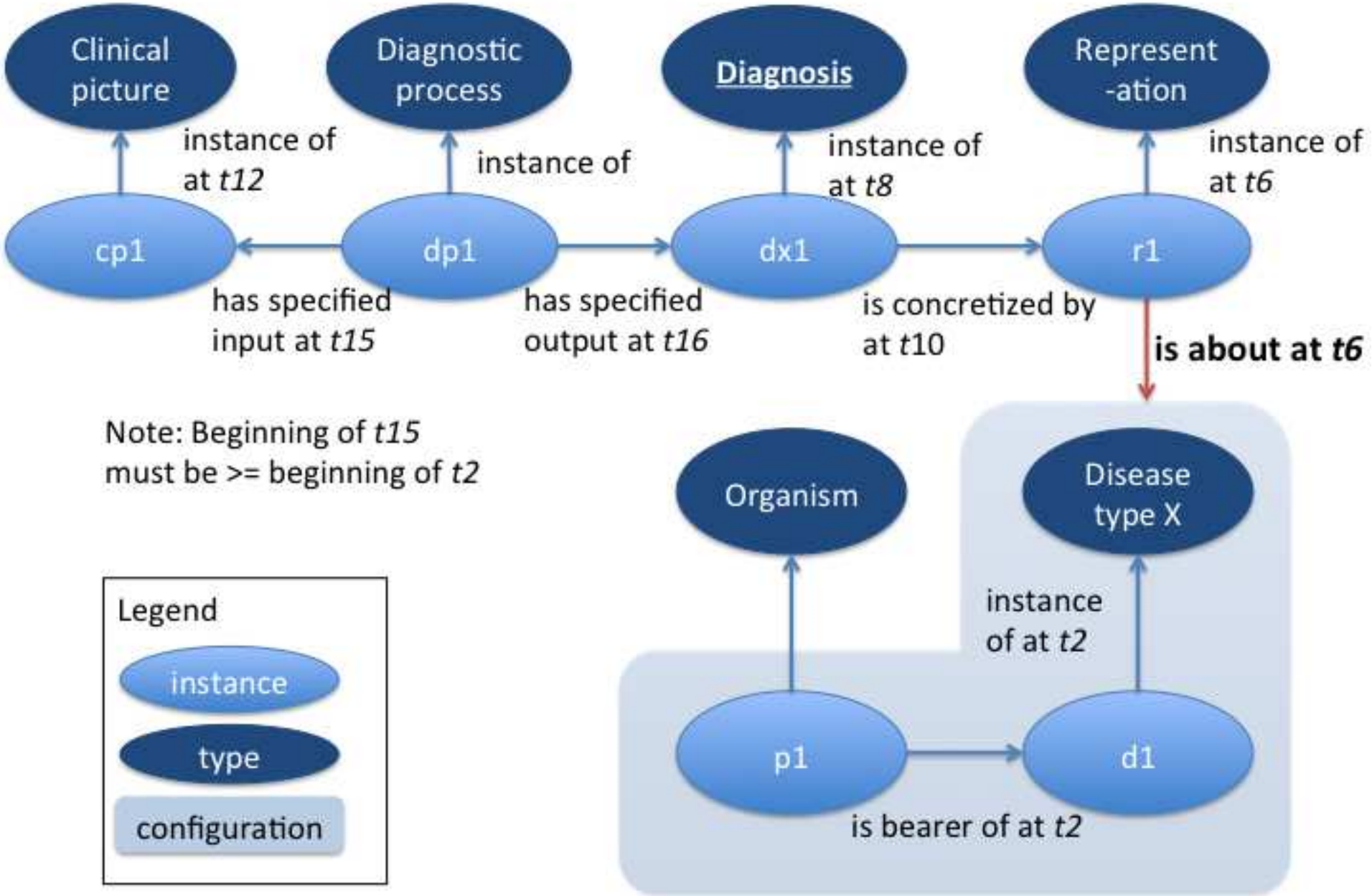
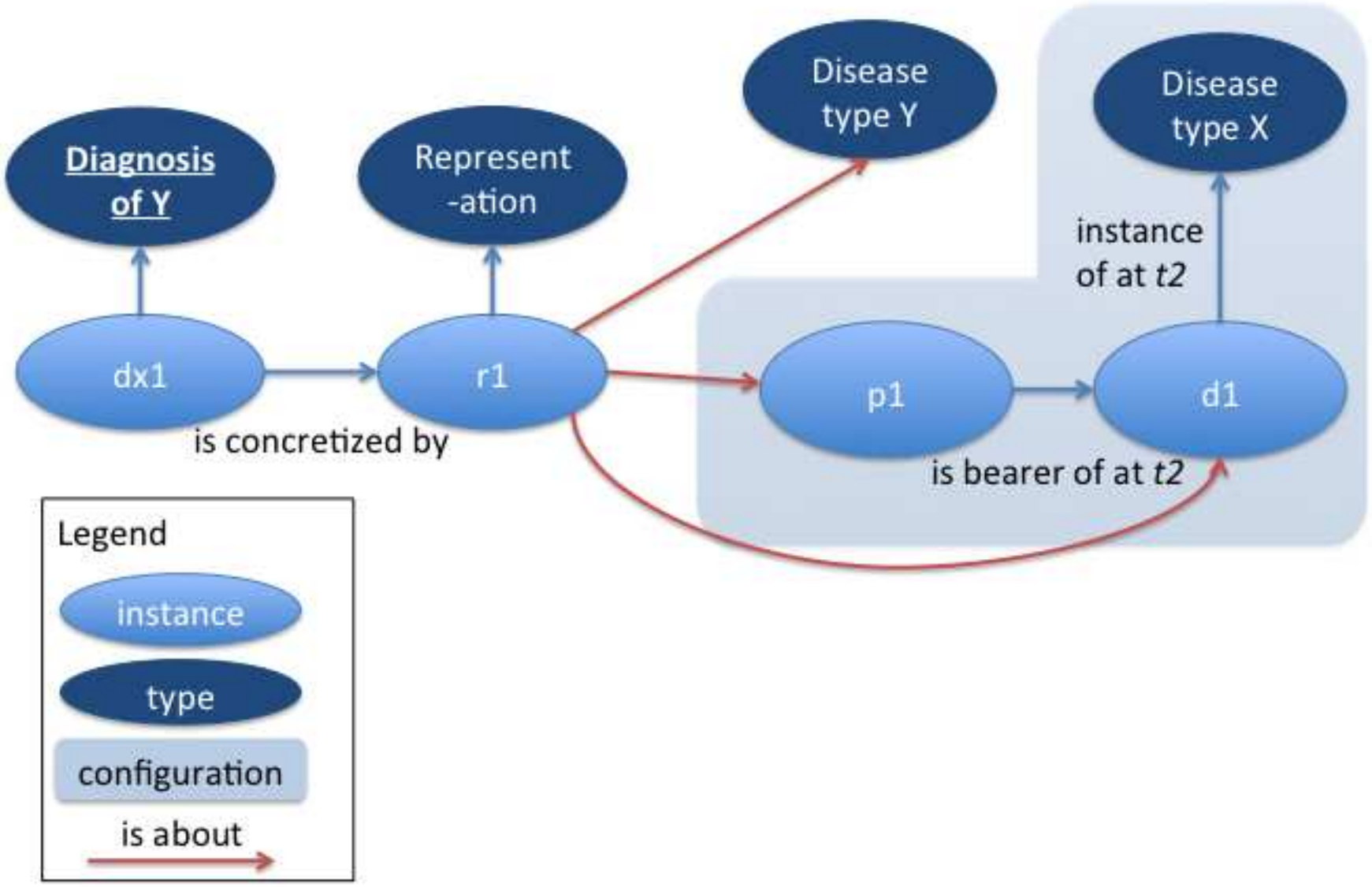


Figure 3

Click here to download Figure Fig 3.png



Additional file 1

Click here to access/download
Supplementary Material
additional file 1 v3.docx

Additional file 2

Click here to access/download
Supplementary Material
additional file 2 v3.docx

Additional file 3

Additional file 4

Journal of Biomedical Semantics

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.

--Manuscript Draft--

Manuscript Number:	JBSM-D-15-00018R3
Full Title:	Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.
Article Type:	Research
Funding Information:	National Center for Advancing Translational Sciences (UL1TR001427) — Not applicable Patient-Centered Outcomes Research Institute (CDRN-1501-26692) — Dr. William R. Hogan
Abstract:	Background Disease and diagnosis have been the subject of much ontological inquiry. However, the insights gained therein have not yet been well enough applied to the study, management, and improvement of data quality in electronic health records (EHR) and administrative systems. Data in these systems suffer from workarounds clinicians are forced to apply due to limitations in the current state-of-the art in system design which ignore the various types of entities that diagnoses as information content entities can be and are about. This leads to difficulties in distinguishing amongst diagnostic assertions misdiagnosis from correct diagnosis, and the former from coincidentally correct statements about disease. Methods We applied recent advances in the ontological understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created six scenarios that we analyzed using the method of Referent Tracking to identify all the entities and their relationships which must be present for each scenario to hold true. We discovered deficiencies in existing ontological definitions and proposed revisions of them to account for the improved understanding that resulted from our analysis. Results Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are typically about a configuration in which there exists a disease that inheres in an organism and instantiates a certain type (e.g., hypertension). Misdiagnoses are ICEs whose concretizations succeed in aboutness on the level of reference for individual entities and types (the organism and the disease), but fail in aboutness on the level of compound expression (i.e., there is no configuration that corresponds in total with what is asserted). Provenance of diagnoses as concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief. Conclusions Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR and other systems to allow diagnostic assertions to be captured with less ambiguity.
Corresponding Author:	William R. Hogan, MD, MS University of Florida Gainesville, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Florida
Corresponding Author's Secondary Institution:	

First Author:	William R. Hogan, MD, MS
First Author Secondary Information:	
Order of Authors:	William R. Hogan, MD, MS
	Werner Ceusters, MD
Order of Authors Secondary Information:	
Response to Reviewers:	We would like to thank the editors and reviewers for their diligent efforts to improve this manuscript. We believe the result is a vastly improved and much clearer exposition of our work (as well as the advancement of our work that occurred during the process). With just one remaining comment left to address, we are gratified that the remainder of the manuscript is publication ready. We endeavor to improve yet further in response to this one last comment. >Reviewer #1: My review is based on revision 2 of the manuscript and the response of the >authors to my previous comments. > >I will leave the decision of whether sufficient context has been included up to the editor and >focus on the second part of my concerns, the kind of representation that is chosen. > We agree and are also happy to let the editor decide. >Essential revision: > >The authors state in their response that the existential quantifier suffices "for _representing_ >them" (phenomena such as misdiagnosis), but not to define them, and have changed their >manuscript accordingly. I agree, but if the aim is to "represent" misdiagnosis, then a single >symbol will suffice; it is perfectly fine to "represent" a misdiagnosis by "A" in any symbolic >formalism. The authors do exactly that by introducing a symbol for misdiagnosis, "is- >misrepresentation-of"; they could also introduce simply a predicate/class "Misdiagnosis(x)", >and they would achieve the goal to "represent" the phenomenon. However, and I believe the >authors would agree with me, this is rarely a useful contribution. Indeed, the authors go much >further and add additional conditions on misrepresentations of reality. They do not explain >why the representation they chose (using additional axioms to further constrain these mis- >representations of reality) are better than merely using a single >symbol. The questions that should be answered are what can be done with that >representation that cannot be done using a single symbol. The contribution (the part involving >the existential quantifier instead of universal one) should also be reformulated and explain >what can be done with, or inferred from, the representation that has been chosen. >Additionally, since the authors add different constraints on mis-diagnosis than Schulz et al., >something will be lost; there needs to be a comparison about what is lost, or what is different >(in terms of the inferences). We have four responses to this comment: 1. First, the reviewer’s numerous mentions of a “single symbol”, or a representation that can be reduced to a single relation or predicate, concerned us (even though we

recognize that the reviewer is not saying this and indeed acknowledges that we go far beyond). We wanted to be absolutely sure that no reader came away equating our representation of misdiagnosis to the mere use of a single symbol 'is-misrepresentation-of'.

Accordingly, we have revised the first paragraph of the conclusions section to summarize our representations of the various entities that are the study of our referent-tracking-based, ontological analysis. In this way, the reader will be left with details of our representation just before finishing the manuscript (assuming the reader reads it sequentially).

2. Second, the reviewer states explicitly that we addressed his previous comment by changing our claim from definition to representation. We quote: The authors state in their response that the existential quantifier suffices "for _representing_ them" (phenomena such as misdiagnosis), but not to define them, and have changed their manuscript accordingly. I agree…

Having explicitly satisfied his past comment, a reasonable person might have expected the issue to end there. Especially because our representation is very much in line with, and as detailed if not more than, that of, say, Martínez Costa and Schulz.

3. However, the reviewer then goes on to raise additional issues not previously raised. This brings us to the last two sentences of the comment. The first of these two sentences moves the goalposts yet again. The reviewer now wants us to compare the relative advantages--for inference--of our representation to (1) that of a single symbol and (2) that of Martínez Costa and Schulz (reference [26]).

#1 is trivial and not worth addressing. We think the reviewer agrees when s/he says that a single symbol alone is "…is rarely a useful contribution". Therefore, demonstrating the utility of something beyond a single symbol is a trivial task and also not a contribution.

With respect to #2, we have already demonstrated several advantages over using the universal quantifier, and in this respect, the comment is misinformed. These advantages include the following. We can represent the aboutness of a misdiagnosis to the organism it is about, whereas Martínez Costa and Schulz cannot. Similarly, we can represent the aboutness of a misdiagnosis to the disease and/or disorder instance that it is about, whereas Martínez Costa and Schulz cannot. We can represent that a misdiagnosis of heart failure is about the type heart failure (in first order logic, and in OWL with workarounds), whereas Martínez Costa and Schulz cannot (except using the workaround of representing the type heart failure with an OWL individual AND declaring that individual to be a member of the separate heart failure class (i.e., heart failure (individual) rdf:type heart failure (class)). Lastly, we can say additionally that a misdiagnosis is simultaneously about all three things (organism, disease, disease type) whereas Martínez Costa and Schulz cannot (without creating a contradiction, as discussed already in the manuscript).

This means that, for example, given suitable data, we can infer all the instances of misdiagnosis of Type 2 diabetes mellitus, in first-order logic minimally and possibly in OWL with workarounds. We can also retrieve all the instances of Type 2 diabetes mellitus that were misdiagnosed as Type 1 diabetes mellitus at any point in time.

We now state these facts in the revised manuscript.

Unlike Hastings et al., we can infer all chemical graphs that depict nitrogen or any other type of atom or chemical group, when the graph is not about any type of existing molecule. In Hastings et al., presuming that nothing is at the same time both a carbon and a nitrogen atom, or both a hydroxyl and a nitrosyl group, or both a benzene ring and a cyclohexane ring, etc. this inference is impossible, because a graph is either about an entire molecule or nothing. In a query for all chemical graphs that depict nitrogen atoms, we retrieve the octaazacubane diagram and they would not.

We include this example, too, in the revised manuscript.

Furthermore, in Martínez Costa and Schulz, their use of the universal quantifier means that any individual specified as (1) being an ICE, (2) being about heart failure, and (3) having a “suspected information object attribute” cannot be inferred as being a “suspected heart failure finding” because of the open world assumption (OWA). Specifically, it is the case that, if this ICE is also about something that is not an instance of heart failure, then it is not a “suspected heart failure finding”. Because of the OWA, the reasoner cannot rule out the existence of such an individual (without additional restrictions on the individual) and therefore does not classify the said individual as a “suspected heart failure finding”.

It also means that any instance declared to be an instance of “suspected heart failure finding” cannot be inferred to be about anything, nor can one infer the existence of any other instance at all—besides an instance of “suspected information object attribute”—including the organism that the finding is about.

4. We could go on. However, NOTHING in the paper deals with reasoning or inference. 'Reasoning' is mentioned in our paper in the context of 'clinical reasoning' as an entity that needs to be represented as an element in the coming into existence of a diagnosis. This is not to say that inferencing with representations of whatever sort is not important, but this is not what this paper is about, nor is it something we wish to discuss further in this paper, and for sure not on the basis of a view of representations as mere symbols tied together in a logical formalism, where the 'quality' of the representation is then determined by the sorts of inferences the logical formalism allows one to make.

It is precisely this reductionist view which we oppose and it is because of this opposition that we use the methodology described, based on representations following the principles of ontological realism and referent tracking. This methodology tries to avoid situations where valid, yet unsound, conclusions are reached by logical formalisms by insisting that every symbol denotes an existing entity.

It is also this methodology that Schulz et al uses in the (other Schulz) paper we refer to (reference [40]); we quote from the discussion section of that paper: 'Much of the work in this paper is motivated by a desire to adhere to Ontological Realism (OR). It assumes that entities in an ontology refer only to individuals in the real world or classes of such individuals [40]'. However, the goal of that paper is, we quote again from the discussion, “…to find a usable solution to incorporate notions of reference in medical discourse into a computable and ontologically sound framework”. Schulz et al [40] clearly indicate that their goal is a framework is BOTH computable and ontologically sound. They created 5 models none of which was able to produce valid inferences using the limited expressiveness of OWL-DL, OWL-Full, and SPARQL without violating good ontological principles, as clearly stated in their conclusion.

Of course we agree with the reviewer that an exploration into logical frameworks that would allow inferencing on the basis of our representations is a worthwhile enterprise, but it is not the enterprise we embarked on here which is about ontologically accurate and precise representation. It is best suited to future work. Also, surely, the reviewer would agree that there is no point in inferencing with ontologically inadequate representations?

We also do not see the need to make any further changes than what we’ve made already in response to this reviewer's last suggestion which is: 'Additionally, since the authors add different constraints on misdiagnosis than Schulz et al., something will be lost; there needs to be a comparison about what is lost, or what is different (in terms of the inferences).' This is because (1) we have already shown a difference (as requested), namely that of aboutness at the level of individual reference, and (2) the Schulz et al. paper is not about misdiagnosis at all, but rather about the degree of certainty diagnoses are asserted as being confirmed or excluded. Note that neither Martínez Costa and Schulz [26], nor Schulz et al.[40], discuss misrepresentation or misdiagnosis as an entity.

We also hold that reviewers should not reject, or insist on revisions of a paper, merely because the topic of the paper is not what s/he is most interested in. In this case, the reviewer seems more interested in inference than in representation.

	A last point is that our refusal to change the manuscript further is not because we don't accept the critique. On the contrary, we now mention that exploring the full spectrum of effects on inference is future work.

Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.

William R. Hogan (corresponding)
University of Florida
P.O. Box 100219
2004 Mowry Rd
Gainesville, FL 32610-0219
hoganwr@ufl.edu
(352) 294-4197

Werner Ceusters
University at Buffalo
921 Main Street
Buffalo, NY 14203
ceusters@buffalo.edu

Abstract

Background

Disease and diagnosis have been the subject of much ontological inquiry. However, the insights gained therein have not yet been well enough applied to the study, management, and improvement of data quality in electronic health records (EHR) and administrative systems. Data in these systems suffer from workarounds clinicians are forced to apply due to limitations in the current state-of-the art in system design which ignore the various types of entities that diagnoses as information content entities can be and are about. This leads to difficulties in distinguishing amongst diagnostic assertions misdiagnosis from correct diagnosis, and the former from coincidentally correct statements about disease.

Methods

We applied recent advances in the ontological understanding of the aboutness relation to the problem of diagnosis and disease as defined by the Ontology for General Medical Science. We created six scenarios that we analyzed using the method of Referent Tracking to identify all the entities and their relationships which must be present for each scenario to hold true. We discovered deficiencies in existing ontological definitions and proposed revisions of them to account for the improved understanding that resulted from our analysis.

Results

Our key result is that a diagnosis is an information content entity (ICE) whose concretization(s) are typically about a configuration in which there exists a disease that

inheres in an organism and instantiates a certain type (e.g., hypertension). Misdiagnoses are ICEs whose concretizations succeed in aboutness on the level of reference for individual entities and types (the organism and the disease), but fail in aboutness on the level of compound expression (i.e., there is no configuration that corresponds in total with what is asserted). Provenance of diagnoses as concretizations is critical to distinguishing them from lucky guesses, hearsay, and justified layperson belief.

Conclusions

Recent improvements in our understanding of aboutness significantly improved our understanding of the ontology of diagnosis and related information content entities, which in turn opens new perspectives for the implementation of data capture methods in EHR and other systems to allow diagnostic assertions to be captured with less ambiguity.

Keywords

Biomedical ontology

Referent tracking

Disease

Diagnosis

Information content entity

Representation

Ontological realism

Background

As administrative, clinical, and patient-reported data are increasingly shared and reused, especially for patient care [1-4] and research [1, 5-7], several issues with these data—including diagnosis data—are of increasing concern. The issue that appears to be of greatest concern is data error and the implications thereof for making decisions and conclusions based on them [8-13]. Although Shapiro et al., in a report for the Office of the National Coordinator for Health Information Technology, do not cite error as a concern for including patient-generated health data into the electronic health record (EHR) [14], there are known errors with patient self reporting especially in research [15-22]. A second issue of concern is data provenance [10, 23], i.e. information about who created the data, in what setting, how, when, for what purpose, and so on. For example, Johnson et al. noted that the provenance of symptom data was essential to using those data correctly to determine whether a colonoscopy was a screening vs. diagnostic procedure [23].

Data error and data provenance are closely related. For example, Hersh et al. note that data recorded in billing workflows for financial purposes are less accurate than clinical data [10]. Thus, timing, method, and purpose of recording data at a minimum—all aspects of provenance—are intertwined with accuracy. Furthermore, a key result of the Johnson et al. study is that "Researchers who do not consider data provenance risk compiling data that are systematically incomplete or incorrect" [23].

An ontological account of data error and data provenance can identify crucial distinctions. For example, there are significant differences among (1) a measured weight that is off

because the scale was not properly tared, (2) a 'rough' weight of 70kg entered in an emergency when the patient cannot be weighed, and (3) a weight measurement entered on the wrong patient. Detecting and accounting for these differences and their causes—especially the aspects of provenance that influence them—is necessary to inform strategies to study, cope with, and improve data error when using pre-existing EHR data for research.

Additionally, a recent review article on the methods for assessing quality of EHR data for clinical research found that: *Most of the studies included in this review presented assessment methodologies that were developed with a minimal empirical or theoretical basis* [24]. It concluded with a call for moving away from ad hoc approaches to data quality assessment, to formal, validated approaches. Although error is only one aspect of data quality (fitness for purpose and completeness are two others), a formal ontological understanding of data error could play a role in more formalized methods for data quality assessment.

In this work, we apply Smith and Ceusters' recent ontological account of incorrect information (i.e., error) [25] to diagnosis data in administrative systems, EHRs, and patient-reported information. Their account holds that a statement such as a diagnostic assertion can succeed or fail in aboutness on at least two levels: (1) the level of denoting single entities and/or types (i.e., the level of *reference*) and (2) the level of veridical representation of a configuration of multiple entities and/or types (i.e., the level of *compound expression*).

To succeed on the second level (compound expression), the information content entity (ICE) must be correct about *all* particulars, their relationships, and their instantiations of types that it mentions. Failure on a single particular, relation, or instantiation causes the ICE to fail at the second level while still potentially succeeding at the first level. For example, if Mrs. Jones has type 1 diabetes mellitus, then the sentence '*Mrs. Jones suffers from type 2 diabetes mellitus'* fails in aboutness on the level of compound expression because it misstates one thing: her disease does not instantiate type 2 diabetes mellitus. However, despite this failure the sentence is nevertheless still about Mrs. Jones, about her disease, and about type 2 diabetes mellitus on the level of reference, because indeed it mentions those three entities. It is therefore, per Smith and Ceusters, an ICE that is about *something* even though it is a misdiagnosis.

Prior ontological work on the aboutness of clinical statements like diagnoses has been constrained by the view that an ICE is about nothing (or is perhaps not even an ICE at all) if it fails on the level of compound expression. Martínez Costa and Schulz, for example, use the universal quantifier when relating an information entity to a clinical situation *…to avoid asserting the existence of an entity the existence of which cannot be guaranteed* [26]. For an ICE such as 'suspected heart failure' they want to avoid the implication that there is some instance of heart failure that it is about. Because they cannot guarantee the existence of some heart failure, they use universal quantification to say 'if it is about anything, then it is about an instance of heart failure'. Researchers working in areas other than diagnosis have encountered similar issues. For example, Hastings et al. note that chemical graphs and diagrams are not always about types of molecules that exist [27]. They, too, used the

workaround of replacing existential quantification with universal quantification to avoid asserting that every chemical graph/diagram is about some type of molecule that exists (level of compound expression), while still allowing such graphs and diagrams to be subtypes of information content entity.

In our own, previous ontological analysis of diagnosis, using the methodology of referent tracking, we identified what entities must exist or must have existed for a particular diagnostic statement to hold true [28, 29]. A key result of this work is that a diagnosis is minimally about *both* the patient and the type of disease that is asserted to exist. In addition, building on previous work on the Ontology for General Medical Science (OGMS), the foundations of which were laid down in Scheuermann et al. [30], we noted that for a diagnosis to exist (at least in medicine and under the assumption that the diagnosis was made *lege artis*), there must also have existed a diagnostic process, a person who carried out that process, and a clinical picture which was used as input into that process.

The hypothesis for the work described here was that applying Smith and Ceusters' results to disease and diagnosis, in combination with prior work on the ontology of disease and diagnosis (including provenance of the latter), could address limitations encountered in previous ontological work on disease and diagnosis and improve our representations of them in support of studying, coping with, and reducing ambiguity in the generation of diagnostic statements and error in the interpretation thereof.

Methods

To test this hypothesis, we analyzed a set of scenarios that we created and that involve correct and incorrect diagnoses, lucky guesses, and justified layperson belief in the existence of a disease of a certain type. The goal was to explore whether, and if so how, a realism-based account of information can deal successfully not only with diagnostic statements asserting the ideal case of a correct diagnosis, but also with deviations from the ideal.

Materials

In our analysis we used as input (1) Smith and Ceusters' work on aboutness and their definitions of representation, mental quality, cognitive representation, and information quality entity (Table 1), (2) definitions of disease, disorder, and diagnosis from the Ontology for General Medical Science (Table 2), and (3) our prior work on analysis of diagnostic statements [27, 28].

Smith and Ceusters stressed that the relation of aboutness includes any portion of reality, rather than being limited to just a single particular or a single universal. A portion of reality (POR) can be a particular, a universal, a relation, or a configuration. A configuration is a combination of particulars and/or universals and certain relation(s) that hold among them.

A representation, then, that is intended to be about a POR but fails in its aboutness because it misrepresents that POR in some way, is misinformation. The sentence *Bob Dylan was in the Beatles* fails to represent not because Bob Dylan or the Beatles did not exist, but

because such a configuration involving Bob Dylan and the Beatles in the way as expressed, never existed. The sentence fails in aboutness on the level of compound expression, but nevertheless is about Bob Dylan and the Beatles individually (on the level of reference) and thus is still an information content entity.

Smith and Ceusters [25] deal more fully with the issue of what it means that a representation is "intended to be about" some entity. Here, we highlight that it follows the doctrine of the "primacy of the intentional" [31], where our written and verbal expressions are to be understood on the basis of the cognitive acts that generated them. That is, a sentence is about that to which its author was directing his or her thoughts when she wrote it.

In addition to Smith and Ceusters' work, we also founded our ontological analysis on the Ontology for General Medical Science or OGMS [30]. This work distinguishes disease, disorder, and diagnosis, and we used definitions from OGMS as starting points for our analysis (Table 2). Note that in OGMS, a diagnosis refers to the existence of a disease of a given type. In clinical medicine, however, diagnoses also refer to (1) disease courses (e.g., acute hepatitis vs. chronic hepatitis), (2) disorders (e.g., fractures and tumors), and (3) the absence of any disease (i.e., a conclusion that a person is healthy also is a diagnosis). It was not our goal to address this issue in this work, as it was not our goal to refine the OGMS definition of diagnosis.

<u>The scenarios</u>

All the scenarios have in common a particular patient, Mr. Adam Jones, who suffers from type 2 diabetes mellitus. Thus in every scenario, there exists Mr. Jones, his disease, the type *Type 2 diabetes mellitus*, the configuration of these three entities (which includes the "bearer of" and "instance of" relationships), and the placement in space and time of this configuration (Figure 1).

Scenario 1: correct diagnosis by physician (ideal case)

Dr. Anne Smith sees Mr. Jones in the office. She takes a history and physical, performs certain laboratory testing, and based on her analysis of the findings, correctly concludes that Mr. Jones has type 2 diabetes mellitus. She subsequently writes her diagnosis in the patient's medical record.

Scenario 2: subsequent correct diagnosis by physician using first diagnosis

A second doctor, Dr. John Brown, sees Mr. Jones in the office at some later date. Mr. Jones has released his records from Dr. Smith to Dr. Brown, who subsequently sees Dr. Smith's diagnosis prior to seeing Mr. Jones. He uses that diagnosis plus his own findings to infer a new clinical picture of Mr. Jones, which he subsequently uses to make another correct diagnosis of Mr. Jones' disease. He writes his diagnosis in Mr. Jones' medical record.

Scenario 3: incorrect diagnosis by physician

Mr. Jones is traveling on vacation, when he falls ill. He sees Dr. Jane Miller who does not have any of his past records available, and thus she is not aware of the previous diagnoses of Drs. Smith or Brown. She infers a new clinical picture of Mr. Jones, and based on it

incorrectly concludes that Mr. Jones has *type 1 diabetes mellitus* (as opposed to type 2). She records a diagnosis of type 1 diabetes mellitus in her medical record for for Mr. Jones.

Scenario #4: coincidentally correct conclusion by layperson (lucky guess)

A friend of Mr. Jones is a "seer". Mr. Jones asks his friend what is in his future. Having no prior knowledge of Mr. Jones medical conditions, the "seer" concludes based on Mr. Jones' horoscope and the position of the moon that he has type 2 diabetes mellitus. He subsequently predicts that Mr. Jones will be hospitalized for his diabetes and miss his daughter's wedding.

Scenario #5: layperson's justifiable conclusion

Mr. Jones' daughter, upon learning of her father's type 2 diabetes mellitus, adds this information into her letter to her brother, writing "Dad has type 2 diabetes mellitus".

Scenario #6: correct diagnosis by computer-based expert system

A medical student is seeing Mr. Jones in the clinic. He performs a history and physical, and types his findings into a diagnostic expert system. The diagnostic expert system infers based on these findings that Mr. Jones has type 2 diabetes mellitus. The medical student writes this diagnosis in Mr. Jones' medical record.

<u>The analysis</u>

Our analysis follows the method of Referent Tracking, which we have found to be a stringent test of ontologies and their definitions [28]. This approach proceeds in three

main steps. First, we systematically identify all the relevant particulars that must exist for the scenario to be true, regardless of whether the scenario explicitly mentions them or only implies their existence. We assign each particular an instance unique identifier (IUI), of the form 'IUI-n', where 'n' is any integer. Second, we identify for each particular the type it instantiates and the temporal interval during which it exists (and assign an identifier of the form *tn* to that interval). Lastly, we identify the relationships that hold between the particulars as well as all relevant relations particulars have to universals other than instantiation, including situations where a particular lacks a given relation to any instance of a certain type (for example, a statement that a patient has had no cough in the last two weeks means that the patient does not stand in the agent_of relation to any instance of the type *Coughing event*, indexed temporally to the two-week interval). [32]

This approach identifies problems in ontologies and their definitions in two major ways. First, it identifies problems that occur when the scenario explicitly rules out the existence of a particular whose existence is implied by an ontological definition (and vice versa). Second, it helps identify exceptions to existing definitions and situations that should not fall under a definition but are erroneously captured by it. Definitions in ontologies can subsequently be adjusted to avoid the errors so identified.

Although our approach is to identify particulars implied by sentences in natural language, the ontological analysis of language and the mechanism(s) by which it makes implicit reference to certain entities is not the focus of this work. Therefore, we convert a sentence like "Mr. Jones has type 2 diabetes mellitus" to Referent Tracking Tuples (e.g., as in Tables 3

through 7) and it is these tuples in which inhere representations that are the objects of our analysis.

To simplify our analysis somewhat, we wrote scenarios under which humans record diagnoses on paper. However, concretization of ICEs also occurs by pixels on monitors, binary switches in memory and processor chips, and magnetic fields on hard disks. But a detailed account of these concretizations and transformations among them is not central to our analysis of what is a diagnosis. Our analysis can be extended to these concretizations without modification of the method.

Results and discussion

In each scenario, Mr. Jones (IUI-1) and his disease (IUI-2) exist, the latter inhering in the former (Table 3). Furthermore, his disease is an instance of the type 'type 2 diabetes mellitus' at any moment in time during which a diagnosis is formulated in any of the scenarios. Mr. Jones (IUI-1) exists through a certain period of time (*t1*) of which we do not know the exact beginning or end. We use temporal identifiers of the form '*t*n' to clearly distinguish such identifiers from IUIs: where IUIs are always intended to be globally and singularly unique, distinct temporal identifiers may denote a unique period of time which is also denoted by another temporal identifier. We also assign an identifier to the time interval during which his disease (IUI-2) exists (*t2*). Diseases usually begin to exist after the organism does, but in the case of congenital genetic diseases, the two intervals might be coextensive. Also, we assume that disease IUI-2 existed at the time of diagnosing, but we

recognize that diagnosing a disease thousands of years after it existed is possible, such as in the case of archaeologists' recent diagnosis of Tutankhamun's malaria [33].

Note that the configuration of organism, disease, and disease type is anchored at a particular location in spacetime, as is the diagnosis. But note also that the diagnosis additionally has an implicit or explicit reference to the location of the configuration in spacetime. To be a correct diagnosis, this reference must also be correct (it has to refer to some part, not necessarily the entirety of spacetime, occupied by the configuration). Thus, for example, to say that Tutankhamun had malaria in 1000 C.E. or today is incorrect, as it would be to say that Mr. Jones had type 2 diabetes mellitus before his parents were born.

***Scenario 1: correct diagnosis*.**

In this scenario, numerous PORs in addition to Mr. Jones and his disease must exist and stand in certain relationships to each other (Tables 4-6). Before Dr. Smith (IUI-3) writes (IUI-13) her diagnosis (IUI-8), there is a cognitive representation (IUI-6) that is concretized in some anatomical part (IUI-5) of her cognitive system (IUI-4). Note that we follow Ceusters and Smith [34] in asserting that all anatomical entities in which cognitive representations inhere are part of a person's cognitive system (that is, any entity used in cognition, including the bearing of cognitive representations, are necessarily within a person's cognitive system) at least during the temporal interval that the cognitive representation exists. If, for example, it would be the case that some white blood cell flowing through some brain capillary would through some of its molecules take part in the concretization of a cognitive representation, then that white blood cell would be part of the

cognitive system at least during the existence of that concretization. It would not anymore be part of the cognitive system once it continues its journey through the body without participating in thought formation. Additionally, Ceusters and Smith take the position (which we also follow) that the cognitive system is not necessarily strictly limited to the brain or even to the entire neurological system of a person: the current state-of-the-art of neuroscience is yet searching for answers to questions such as "what is it in which cognitive representations inhere?" but until it reaches such answers, we remain in our representations agnostic.

IUI-9 denotes the sentence Dr. Smith wrote, as it exists on the particular piece of paper she used to write it on: 'The patient has type 2 diabetes mellitus'. This written statement on paper (IUI-9) bears an information quality entity (IQE, IUI-10) that concretizes her diagnosis (IUI-8). The cognitive representation (IUI-6) and IQE (IUI-10) that concretize the diagnosis are both about the configuration (IUI-7) (the level of compound expression), as well as about Mr. Jones, Mr. Jones' disease, and the universal *Type 2 diabetes mellitus* individually (the level of reference). The cognitive representation (IUI-6) and the diagnosis (IUI-8) are the output of Dr. Smith's diagnostic process (IUI-11), which had as input Dr. Smith's clinical picture (IUI-12) of Mr. Jones. Because the cognitive representation and IQE concretize the same ICE, the latter is conformant to the former (see Table 1).

A correct diagnosis is thus fundamentally an information content entity that is concretized by a representation that stands in an is_about relation to the configuration of an organism, its disease, the relation of inherence between the disease and the organism, a type that the

disease instantiates, and the instantiation relation of the disease to that type, all within a given portion of spacetime (Figure 2). Furthermore, diagnoses are additionally differentiated from other ICEs by the fact that they are generated by a diagnostic process that has a clinical picture as input. We expand further on what constitutes a clinical picture in the next scenario, *Scenario 2*, as well as revisit the diagnostic process briefly in *Scenario 4*, although it was not our objective in this work to develop a fuller account of this process.

Note that it is trivial to state that the particular disease inhering in the organism is an instance of *entity* or even *disease*. Thus, there is an expectation that a diagnosis be as precise (the most specific type) as possible and at a minimal level of granularity that is relevant to treat the patient appropriately and to provide a reasonable prognosis.

Scenario 2*: *second diagnosis.

The second physician, Dr. Brown, makes a second diagnosis at a later point in time, using the first diagnosis in addition to clinical and possibly other findings to infer a new clinical picture of Mr. Jones. With the exception of the configuration of Mr. Jones/his disease/type 2 diabetes mellitus (IUI-7), there is a one-to-one correspondence of PORs as in Scenario 1, numbered IUI-23 through IUI-33 (Additional file 1 : Tables S1-S3). That is, there is no IUI-27 because the configuration is the same POR across scenarios. Similarly, there is no IUI-21 or IUI-22 because IUI-1 and IUI-2 already identify Mr. Jones and his disease, respectively, uniquely.

In this scenario, Dr. Brown (IUI-23) makes a new diagnosis (IUI-28), concretized both by his cognitive representation (IUI-26) in some part (IUI-25) of his cognitive system (IUI-24) and by the IQE (IUI-30) inhering in the sentence in his note (IUI-29). Dr. Smith's previous diagnosis (IUI-8) can be viewed as either (*view1*) being in the aggregate of things that Dr. Brown uses to infer his clinical picture (IUI-32) that serves as input into his diagnostic process (IUI-31), or (*view2*) as something which serves as extra input—alongside his clinical picture—for the diagnostic process. The cognitive representation and the IQE are about the configuration (IUI-7) as well as Mr. Jones (IUI-1), his disease (IUI-2), and type 2 diabetes mellitus (UUI-1).

The current definition of 'clinical picture' in OGMS (see Table 2) seems to conflict with *view1* about this scenario, because the definition seems to exclude using a past diagnosis to infer a clinical picture. Although the current OGMS definition of 'clinical picture' is inclusive of clinical findings, diagnosis as currently defined is not an explicit subtype of clinical finding in OGMS. Furthermore, it is common for clinicians to elicit a previous provider's past diagnosis from the patient or the patient's caregiver during an interview (for example, if Mr. Jones in scenario #2 would have said: 'Dr. Smith says I have type 2 diabetes mellitus'). But the current OGMS definition of 'clinical history' (Table 2) conflicts with this possibility. It refers to health-relevant features of a patient, but features as elucidated by OGMS include only qualities, processes, and physical components of the organism—not dispositions of which disease is a subtype. Therefore, a representation of a disease such as a diagnosis is currently excluded from the OGMS definition of 'clinical history'.

We also note that the OGMS definition of 'clinical picture' is ambiguous in that it is not clear whether it *requires* that laboratory and image findings must always be used to infer a clinical picture, or that they are the only entities that can be used. Regardless, it would be a mistake to do so, because diagnoses can and frequently are made from symptom findings alone. Laboratory and image findings are not necessary components of a clinical picture in reality. Note that a clinical picture can comprise findings of a single type (laboratory alone, pathology image alone, radiology image alone, physical exam finding alone), or even a single finding instance (e.g. Reed-Sternberg cells for a diagnosis of Hodgkin's lymphoma). All these issues are compounded by the fact that the term 'clinical picture' itself is not intuitive.

Given that clinical history taking elicits past diagnoses routinely in clinical medicine, we propose modifying the definition of 'clinical history' to accommodate this reality (bolded sections represent changes to the definition):

clinical history =def. – *A series of statements representing one or more health-relevant features of a patient**, possibly complemented by representations of diseases and configurations**.*

Note that the definition already allows—under the broader heading of 'feature'—representations of disorders (kinds of physical component) and disease courses (kinds of process). Thus, the definition already accommodates these aspects of clinical histories. We

also allow the statements to represent configurations, in line with Smith and Ceusters [2]. These configurations might or might not include various relevant types (for example, "The patient has not participated in any instance of vomiting in the last two weeks."). Finally, note that by using the word 'representing', the definition also accommodates per Smith and Ceusters [2] that some statements might fail in aboutness despite their intention to be about such features. In other words, some statements in the clinical picture might be wrong: for example, a statement that the patient has a disease or pain that she does not in fact have.

To clarify that laboratory and imaging findings are not always required inputs into the diagnostic process, and to capture realistic scenarios compatible with *view2* (for example, Dr. Brown reads Dr. Smith's note in the chart), we also propose a modified definition of 'clinical picture' (changes in bold):

clinical picture =def. – *A representation of a clinical phenotype that is inferred from* ***a*** *combination of,* ***for example, diagnoses and*** *laboratory, image, and clinical findings about a given patient.*

These changes to the definitions of 'clinical history' and 'clinical picture' now properly capture situations where past diagnoses are elicited from the patient and/or her caregiver during a clinical history taking: these diagnoses are now clinical findings in the clinical history that was generated by the clinical history taking (see the definition of 'clinical finding' in Table 2).

Scenario 3: Misdiagnosis.

The third physician, Dr. Miller, misdiagnoses Mr. Jones' type 2 diabetes mellitus as type 1 diabetes mellitus (Figure 3). Per Smith and Ceusters, because the misdiagnosis is still about Mr. Jones, his disease, the relationship between them, and the type 'type 1 diabetes mellitus' on the level of reference, it is an information content entity. However, it fails to be about the configuration IUI-7 as a whole on the level of compound expression.

Again, in this scenario there exist PORs in one-to-one correspondence (except the configuration and its components) numbered IUI-43 through IUI-53 (Additional file 2 : Tables S4-S6). Dr. Miller (IUI-43) writes (IUI-53) his misdiagnosis (IUI-48) in Mr. Jones' chart, and the IQE (IUI-50) inhering in the ink (IUI-49) is conformant to his cognitive representation (IUI-46), and both are about—on the level of reference—Mr. Jones, his disease, and type 1 diabetes mellitus. But neither one is about the configuration (IUI-7). To capture the relation both (1) between the cognitive representation and the configuration and (2) between the IQE and the configuration, we define a new relation:

> **is-misrepresentation-of**: domain: representation, range: portion of reality.
>
> Def: *x* is-misrepresentation of *y* iif *x* is a representation and *x* is intended to be about *y* and it is not the case that *x* is about *y*.

Then we assert that the representations (IUI-46 and IUI-50) are misrepresentations of the configuration (Table 7 and Additional file 2 : Table S6). Note that our definition precludes the cognitive representation (IUI-46) and IQE (IUI-50) being about any configuration other

than IUI-7, because they are not intended to be about, for example, the configuration of the sun, earth, and moon at a particular date and time.

Note that asserting the incorrect disease type is not the only way to make a misdiagnosis. There are at least six possibilities where a diagnosis fails to be about a configuration on the level of compound expression (Table 8). If a representation fails on the level of reference, it also fails on the level of compound expressions, because a configuration cannot consist of that which does not exist. These six possibilities could also exist in combination, but if the 2nd, 3rd, and 4th possibilities are all present (for example, "Ron Weasley has spattergroit"), then there is not a diagnosis, or even any information content entity at all, because the representation is not about anything even on the level of reference. Of course, if the organism itself does not exist, then there cannot be a clinical picture inferred, and thus it would not be a diagnosis or misdiagnosis, although it could still be an ICE if it is about a really-existing disease type (for example, "James Bond has influenza"). Also, as medical knowledge evolves, the profession comes to understand that certain types of disease thought to exist in fact do not. Thus past diagnoses of *dropsy* and *consumption* we now understand to be misdiagnoses.

Despite searching the extensive literature on diagnostic error, we could not find any studies that looked at what percentages of misdiagnoses fall into these categories. We conjecture based on our past clinical expertise and experience that asserting the incorrect disease type is the most common mistake among those in Table 8, but confirmation or rejection of this conjecture requires study.

Scenario 4: the lucky guess.

In this scenario, a layperson (the "seer"—IUI 63) correctly concluded coincidentally that Mr. Jones had type 2 diabetes mellitus based on the position of the moon and Mr. Jones' horoscope (Additional file 3 : Tables S7-S9). It would be wrong to say the seer's reasoning (IUI-71) constituted a diagnostic process. To avoid coincidentally correct statements from qualifying as diagnoses, we additionally require as input into the diagnostic process cognitive representations of the disease type and the types instantiated by the sequalae, signs, symptoms, and any clinical, laboratory, or imaging findings or phenotypes of the instances of this disease type. Note that this is a minimal requirement: clinicians often additionally include in their diagnostic reasoning cognitive representations of other disease types and associated PORs when considering alternative possibilities for the disease type.

This view is based on the extensive literature on clinical reasoning processes, especially diagnosis (for a review, see Norman [35]). This research has established the use of representations, called 'knowledge structures', in the diagnostic process. The nature and form of these representations evolves as clinical expertise develops [36], and we note that the differences in diagnostic processes that result could result in a typology of diagnostic processes in OGMS.

Because the seer had no cognitive representations of type 2 diabetes mellitus, let alone used them as input into his "reasoning", his conclusion (IUI-68), although an ICE, is not a

diagnosis. Similarly, if a physician makes a lucky guess based not on his cognitive representations of the stated disease type but instead by flipping a coin or some such, that too would not be a diagnosis.

To Table 3 we add an aggregate of cognitive representations of disease types and associated entities as input into the diagnostic process (Table 9).

We propose to redefine diagnostic process as follows:

Diagnostic process =def. *An interpretive PROCESS that has as input (1) a CLINICAL PICTURE of a given patient* ***AND (2) an aggregate of REPRESENTATIONs of at least one type of disease and at least one type of phenotype whose instances are associated with instances of that disease,*** *and as output an assertion to the effect that the patient has a DISEASE of a certain type.*

Scenario 5: layperson's justifiable conclusion.

Mr. Jones' daughter wrote a sentence in her letter to her brother based on reading Dr. Smith's progress note saying that that her father has type 2 diabetes mellitus (Additional file 4 : Tables S10-S12). Of course, the daughter has not made a diagnosis. She is communicating to her brother what she believes to be the case.

Had she merely written "Dr. Smith says" and then copied Dr. Smith's sentence word for word into her letter, then her writing would concretize Dr. Smith's diagnosis (IUI-8). This is the case of hearsay ("so-and-so said it was the case that…").

As Smith and Ceusters showed, however, the same sentence written by two different people does not guarantee they concretize the same ICE. ICEs are further differentiated by the provenance of their concretizations, including who created them and when, and to what POR they intend to be about. In their example, two people writing the sentence *Barack Obama has never been President of the United States*—one before and one after Obama's inauguration as President—generate two different ICEs. The one written after fails on the level of compound expressions but not on the level of reference, whereas the one written before succeeds on both levels (it remains true that at the time when the sentence was written, he had never been President).

We therefore distinguish between a human (1) merely copying a representation, in which case the copy concretizes the same ICE as the original text and (2) creating her own cognitive representation of the POR—which involves forming a belief that the POR really existed as represented—and then subsequently creating an IQE that is conformant to the cognitive representation. In the former case, a new ICE does not come into being. It does not even require in the cognitive system of the copier any representation of the POR that the original representation is about (as in the case of copying German text that one does not understand at all). In the latter case, by contrast, a new ICE does come into being.

In Scenario 5, the daughter did not merely repeat Dr. Smith's diagnosis. She communicated to her brother *her* belief about her father's disease. She deliberately chose not to merely convey Dr. Smith's diagnosis, but rather her belief that her father has type 2 diabetes

mellitus. She heard the opinion of an expert, in whom she had trust. Based on (1) her observations of her father, (2) Dr. Smith's diagnosis, and (3) her trust in Dr. Smith, she reached the conclusion herself that her father suffers from type 2 diabetes mellitus. Because she did not begin with a clinical picture and her own cognitive representations of type 2 diabetes mellitus, her conclusion is not a diagnosis.

However, consider the scenario where she is given the clinical picture and has enough knowledge to arrive at a conclusion, which could be the case either if she were a physician or somehow other acquired or were given the necessary knowledge: it is analagous to Scenario #6, where she takes the place of the expert system (see analysis of that scenario below). Thus, here in Scenario #5 it is important to note that she did not reason from a clinical picture to the diagnosis.

In this scenario, therefore, the daughter has created a new ICE (IUI-88) that is not a diagnosis. She has concretized it in the sentence (IUI-89) in her letter.

Scenario 6: diagnosis by non-human.

The diagnostic decision support system has made a diagnosis (or misdiagnosis depending on whether it is correct), because it (1) takes as input a clinical picture and representations of the relevant disease type and one or more types of phenotypes with which it is associated; (2) participates in a process of making a conclusion based on this input; and (3) outputs from this process a statement about a configuration involving an organism, a disease, and a disease type.

In this case, there are no cognitive representations. In their place are digital representations on hard drives, memory chips, and central processing units. If we assume the system generates a sentence and prints it on paper, then we have an analagous IQE to the written diagnosis of the physician and ICE of the sister.

Nothing in our proposed definitions conflicts with this scenario. Replacing Dr. Smith and associated representations and diagnostic process with various components of the computer and its digital representations as well as inferential process (which is an instance of diagnostic process) is straightforward.

Returning briefly to a point made in Scenario #5, Mr. Jones' daughter could follow the exact same algorithm(s) of the diagnostic expert system using the exact same clinical picture as input, and she would arrive at (or make) a diagnosis, in contrast to scenario #5 where her conclusion was an ICE but not a diagnosis.

Conclusions

We applied Smith and Ceusters' results on aboutness [25] to diagnosis in order to develop an account of diagnosis, misdiagnosis, lucky guesses, hearsay, a layperson's justified belief about disease configurations, and a diagnosis made by an expert system. Our key result is that a correct diagnosis, as defined by OGMS, is about a configuration of an organism, its disease, and the type the disease instantiates (level of compound expression) in a specified portion of spacetime. A misdiagnosis by contrast is a misrepresentation of this

configuration. Nevertheless, both diagnosis and misdiagnosis are still about—at the level of individual reference—the organism and (when they exist) a disease instance and a disease type. Also, they are both the output of a diagnostic process, which differentiates them from lucky guess and hearsay as well as the misinformation-based counterparts to lucky guess and hearsay. We also carefully represented the inputs and outputs of this process.

We identified several subtypes of misdiagnosis (e.g., wrong disease subtype, wrong patient, wrong temporal placement) that have not been differentiated in the literature on diagnostic error, to our knowledge. Studying the incidence and causes of these subtypes might advance the study of diagnostic error and strategies to reduce it. Note that as we have defined it, 'misdiagnosis' does not refer to the diagnostic errors of absent diagnosis (failing to diagnose a disease at all, let alone incorrectly) and delayed diagnosis. Lastly, we note that the current literature on diagnostic error, per a 2016 Institute of Medicine report, does not lend itself to generating reliable estimates of incidence of diagnostic error per se, let alone any subtype of such error [37].

Although misdiagnoses involving non-existence of certain entities might at first seem to be of minor importance, we highlight two cases where non-existence is relevant. First, in the case where the type of disease does not exist (consider past diagnoses of "dropsy"), it could well be that our understanding of disease decades from now is much more advanced, and what we think are types of disease today in fact are not. So just as with past diagnoses of "dropsy", it could be that today's diagnoses of "schizophrenia" are misdiagnoses merely by

referring to a type that does not exist. Second, in the case where the instance of disease does not exist, we consider two scenarios. The first scenario involves past diagnoses of mental illness where neither the instance nor the type exists. For example, past diagnoses of runaway slaves as having "drapetomania" involved neither a really existing instance nor a really existing type of disease. The second scenario involves patients with hypochondria or who are malingering. They feign a condition for which the unassuming practitioner mistakenly asserts the existence of an instance and the instantiation of a type.

Our results and typology of misdiagnosis could serve as the beginnings of a formal framework for studying diagnostic error as a component of data quality in EHRs and research data collections, in response to the call by Weiskopf and Weng for more formal, generalizable, and validated methods for assessing data quality [24]. Applying Ceusters' detailed typology of mistakes in ontology (e.g., asserting a type that does not exist) [38] and referent tracking systems (e.g., assigning an identifier but there is no corresponding particular that it identifies, assigning one identifier to two particulars, assigning two identifiers to one particular, etc.) [39] to diagnosis could build on our work here to build out such a framework. It remains future work to do so.

The provenance of the ICE and its concretizations are critical: lucky guesses, hearsay, and laypersons' conclusions about disease (when not arrived at through a diagnostic process using a clinical picture and cognitive representations of the associated type(s) of disease as input) do not constitute diagnoses and therefore are different types of ICE than diagnoses. Provenance also includes which findings and other information constituted the clinical

picture used in the diagnostic process. Our analysis of the scenarios identified past diagnoses as important input into the diagnostic process, leading to proposed redefinitions of 'clinical history', 'clinical picture', and 'diagnostic process' for OGMS.

Smith and Ceusters' results on aboutness and our extension of them here to diagnosis reduce the need for the workarounds reported by Martínez Costa and Schulz [26] and Hastings et al. [27] It is perfectly legitimate to relate 'suspected heart failure finding' to 'congestive heart failure' with an existential quantifier: if an instance of this type is not about a really-existing configuration of patient–disease–heart failure, it is still an ICE that is individually about the patient, her condition, and the type *heart failure* on the level of reference. In OWL, we could assert:

Suspected heart failure ICE -> ICE and (**is about** SOME *Organism*)

Suspected heart failure ICE -> ICE and (**is about** SOME *Condition*)

In more expressive formalisms including first-order logic, we could also assert that it is about the type *heart failure*, where 'Type', 'Instance_of', and 'Is_about' are predicates in what follows, where the universal quantification applies to the ICE, not what it is about:

Type(heart_failure)

Type(suspected_heart_failure_ICE)

$\forall x$ (Instance_of(x, suspected_heart_failure_ICE) -> Is_about(x, heart_failure))

Similarly, chemical graphs and diagrams are ICEs about individual types of atoms such as carbon, oxygen, hydrogen, and so on, even when they fail to be about any type of configuration (e.g., molecule) of such atoms. However, because they are typically not about any instances, proper existential quantification in OWL is not possible. However, we could

relate in first-order logic the diagram of *octaazacubane* (a hypothetical molecule which would be comprised of eight nitrogen atoms arranged in a cubic structure) to the *nitrogen* type of atom using existential quantification (again where the universal quantification in what follows applies to the ICE and not what it is about):

Type(nitrogen_atom)

Type(octaazacubane _diagram)

$\forall x$ (Instance_of(x, octaazacubane_diagram) -> Is_about(x, nitrogen_atom))

It is therefore not required to use universal quantification over the range of things that an ICE is about, when relating ICEs to those entities they are about, to avoid failure of aboutness on the level of compound expression. This result is qualified by the constraints of representational formalisms such as OWL that prevent directly asserting aboutness to types. Schulz et al. describe workarounds in OWL to asserting aboutness to types, that may be of benefit in some use cases [40].

The use of universal quantification actually introduces problems when we account for aboutness on the level of individual reference. For example, if we leave the 'suspected heart failure finding' of Martínez Costa and Schulz as being *only* about 'congestive heart failure', then it would result in a contradiction to say that it is about some organism. Likewise for condition. So use of the universal quantifier precludes aboutness on the level of individual reference, in direct conflict with the results of Smith and Ceusters on misinformation.

Although it was not the primary or even secondary goal of the present work, other advantages of our approach with respect to inference are easy to derive. First, in our approach with explicit representation of the disease in addition to the diagnosis, we can infer all instances of Type 1 diabetes mellitus that have been misdiagnosed as Type 2 diabetes mellitus at some point in time, in first order logic minimally and possibly in OWL with workarounds. Generalizing slightly, we can query for all conditions that have been misdiagnosed as Type 2 diabetes mellitus. Using a typology of organisms, we can find in the veterinary domain all diagnoses and/or misdiagnoses of a certain type of disease in organisms of a certain type: for example, misdiagnoses of foot and mouth disease in cattle. Having no ability to create an aboutness relation from a misdiagnosis, or more generally an incorrect clinical statement, to the organism it is about (due to the contradictions that will result as pointed out above) or even to anything in reality at all, the universal quantifier approach of Martínez Costa and Schulz would require substantial revision to make these inferences.

In the realm of chemical diagrams, our approach enables one to query for all chemical diagrams that depict nitrogen atoms or certain chemical groups (e.g., hydroxyl group and benzene rings), *including the diagrams that are not about any existing type of molecule*. The universal quantifier approach in Hastings et al., by contrast, would require significant revision to return diagrams that depict nitrogen, hydroxyl groups, benzene rings, and so on, but are not about any existing type of molecule. In depth exploration of the effects of our representation on inference remains future work, as it is not our primary interest here.

Our analysis also identified problems with, and suggested improvements to, the definitions of core terms from the Ontology for General Medical Science including 'diagnostic process' and 'clinical picture'. This result is consistent with our past work, where we have found the method of referent tracking analysis to be a stringent test of definitions in ontologies.

This work is limited by the fact that we did not conduct further ontological analysis of the diagnostic process beyond OGMS and beyond what our scenarios required, as this was not the purpose of the present work. We do note that our requirement for including cognitive representations of disease types as input into the diagnostic process is based on this literature, however. Engaging experts in the study of clinical reasoning in future work to develop a typology of diagnostic processes has the potential to result in a corresponding typology of diagnoses.

Future work includes (1) an account of differential diagnosis, where a clinician or expert system generates a list of likely types of disease for further investigation to identify the actual type the organism's disease instantiates; (2) proposing to the OGMS community to clarify the definitions of 'clinical history', 'clinical picture', and 'diagnostic process' as suggested here, and to expand the definition of diagnosis to include disorders, disease courses, and absence of disease (i.e., healthy); (3) extending our analysis as reported here to this expanded definition of 'diagnosis'; (4) conducting deeper ontological analysis of the diagnostic process, in coordination with experts in the study of clinical reasoning; and (5) more fully exploring the effects of our representations on logical inference beyond some readily evident advantages discussed here.

List of Abbreviations

BFO	Basic Formal Ontology
GDC	Generically dependent continuant
IAO	Information Artifact Ontology
ICE	Information Content Entity
IQE	Information Quality Entity
OGMS	Ontology for General Medical Science
POR	Portion of Reality
RT	Referent Tracking
RTT	Referent Tracking Tuple

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

The authors contributed equally to the ontological analysis and development of results. Author WRH created the first version of the manuscript. Both authors had full access to all materials and analysis and participated in revising the manuscript. Both authors approved the final version of the manuscript.

Acknowledgments

This work was suppoted in part by the NIH/NCATS Clinical and Translational Science Award to the University of Florida UL1TR001427.

References

1. Bingham CO, 3rd, Bartlett SJ, Merkel PA, Mielenz TJ, Pilkonis PA, Edmundson L, Moore E, Sabharwal RK: **Using patient-reported outcomes and PROMIS in research and clinical applications: experiences from the PCORI pilot projects**. *Qual Life Res* 2016.
2. Scanlon L: **PatientsLikeMe Survey Shows Vast Majority of People With Health Conditions Are Willing To Share Their Health Data**. http://news.patientslikeme.com/press-release/patientslikeme-survey-shows-vast-majority-people-health-conditions-are-willing-share-t; 2014. Accessed August 24, 2016.
3. Rudin RS, Motala A, Goldzweig CL, Shekelle PG: **Usage and effect of health information exchange: a systematic review**. *Ann Intern Med* 2014, **161**(11):803-811.
4. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P: **From the Office of the National Coordinator: the strategy for advancing the exchange of health information**. *Health Aff (Millwood)* 2012, **31**(3):527-536.
5. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS: **Launching PCORnet, a national patient-centered clinical research network**. *J Am Med Inform Assoc* 2014, **21**(4):578-582.
6. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM *et al*: **The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies**. *BMC Med Genomics* 2011, **4**:13.
7. Owens B: **DATA SHARING. Montreal institute going 'open' to accelerate science**. *Science* 2016, **351**(6271):329.
8. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF: **Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors**. *Med Care* 2005, **43**(5):480-485.
9. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM: **Measuring diagnoses: ICD code accuracy**. *Health Serv Res* 2005, **40**(5 Pt 2):1620-1639.
10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ *et al*: **Caveats for the use of operational electronic health record data in comparative effectiveness research**. *Med Care* 2013, **51**(8 Suppl 3):S30-37.
11. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS: **Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied**. *Med Care* 2013, **51**(8 Suppl 3):S80-86.

12. Botsis T, Hartvigsen G, Chen F, Weng C: **Secondary Use of EHR: Data Quality Issues and Informatics Opportunities**. *AMIA Jt Summits Transl Sci Proc* 2010, **2010**:1-5.
13. Benesch C, Witter DM, Jr., Wilder AL, Duncan PW, Samsa GP, Matchar DB: **Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease**. *Neurology* 1997, **49**(3):660-664.
14. Shapiro M, Johnston D, Wald J, Mon D: **Patient-Generated Health Data White Paper**. RTI International, Research Triangle Park, NC 27709; 2012. Available at: https://www.healthit.gov/sites/default/files/rti_pghd_whitepaper_april_2012.pdf. Accessed August 24, 2016.
15. Gordon NP, Mellor RG: **Accuracy of parent-reported information for estimating prevalence of overweight and obesity in a race-ethnically diverse pediatric clinic population aged 3 to 12**. *BMC pediatrics* 2015, **15**(1):5.
16. Komaroff AL: **The variability and inaccuracy of medical data**. *Proceedings of the IEEE* 1979, **67**(9):1196-1296.
17. Callahan CM, Tu W, Stump TE, Clark DO, Unroe KT, Hendrie HC: **Errors in self-reports of health services use: impact on alzheimer disease clinical trial designs**. *Alzheimer Dis Assoc Disord* 2015, **29**(1):75-81.
18. Monte AA, Heard KJ, Hoppe JA, Vasiliou V, Gonzalez FJ: **The accuracy of self-reported drug ingestion histories in emergency department patients**. *J Clin Pharmacol* 2015, **55**(1):33-38.
19. Gerritsen M, Berndt N, Lechner L, de Vries H, Mudde A, Bolman C: **Self-Reporting of Smoking Cessation in Cardiac Patients: How Reliable Is It and Is Reliability Associated With Patient Characteristics?** *J Addict Med* 2015, **9**(4):308-316.
20. Raphael KG, Janal MN, Sirois DA, Dubrovsky B, Klausner JJ, Krieger AC, Lavigne GJ: **Validity of self-reported sleep bruxism among myofascial temporomandibular disorder patients and controls**. *J Oral Rehabil* 2015, **42**(10):751-758.
21. Patel M, Perrin K, Pritchard A, Williams M, Wijesinghe M, Weatherall M, Beasley R: **Accuracy of patient self-report as a measure of inhaled asthma medication use**. *Respirology* 2013, **18**(3):546-552.
22. Woodfield R, Group UKBSO, Follow-up UKB, Outcomes Working G, Sudlow CL: **Accuracy of Patient Self-Report of Stroke: A Systematic Review from the UK Biobank Stroke Outcomes Group**. *PLoS One* 2015, **10**(9):e0137538.
23. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ: **How the provenance of electronic health record data matters for research: a case example using system mapping**. *EGEMS (Wash DC)* 2014, **2**(1):1058. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4371416/pdf/egems1058.pdf. Accessed August 24, 2016.
24. Weiskopf NG, Weng C: **Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research**. *J Am Med Inform Assoc* 2013, **20**(1):144-151.
25. Smith B, Ceusters W: **Aboutness: Towards Foundations for the Information Artifact Ontology**. In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015. Available at: http://ceur-ws.org/Vol-1515/regular10.pdf. Accessed August 24, 2016.

26. Martínez-Costa C, Schulz S: **Ontology-based reinterpretation of the SNOMED CT context model**. In: *Proceedings of the Fourth International Conference on Biomedical Ontology: July 7th-12th, 2013; Montreal*: Edited by Dumontier M, Hoehndorf R, Baker CJO. 2013: 90-95. Available at: http://ceur-ws.org/Vol-1060/icbo2013_submission_38.pdf. Accessed August 24, 2016.
27. Hastings J, Batchelor C, Neuhaus F, Steinbeck C: **What's in an 'is about' link? Chemical diagrams and the information artifact ontology**. In: *Proceedings of the 2nd International Conference on Biomedical Ontology; Buffalo, New York*: Edited by Bodenreider O, Martone ME, Ruttenberg A. 2011: 201-208. Available at: http://ceur-ws.org/Vol-833/paper26.pdf. Accessed August 24, 2016.
28. Ceusters W, Hogan WR: **An ontological analysis of diagnostic assertions in electronic healthcare records** In: *Proceedings of the Sixth International Conference on Biomedical Ontology: July 27-30, 2015; Lisboa, Portugal*. 2015. Available at: http://ceur-ws.org/Vol-1515/regular2.pdf. Accessed August 24, 2016.
29. Hogan WR: **To what entities does an ICD-9-CM code refer? A realist approach.** In: *Bio-ontologies; Boston, MA*: Edited by Shah N, Sansone S-A, Stephens S, Soldatova L. 2010.
30. Scheuermann RH, Ceusters W, Smith B: **Toward an ontological treatment of disease and diagnosis**. In: *AMIA Summit on Translational Bioinformatics: 2009*. 116-120.
31. Chisholm RM: **The primacy of the intentional**. *Synthese*, **61**(1):89-109.
32. Ceusters W, Elkin P, Smith B: **Negative findings in electronic health records and biomedical ontologies: a realist approach**. *Int J Med Inform* 2007, **76 Suppl 3**:S326-333.
33. Hawass Z, Gad YZ, Ismail S, Khairat R, Fathalla D, Hasan N, Ahmed A, Elleithy H, Ball M, Gaballah F *et al*: **Ancestry and pathology in King Tutankhamun's family**. *JAMA* 2010, **303**(7):638-647.
34. Ceusters W, Smith B: **Foundations for a realist ontology of mental disease**. *J Biomed Semantics* 2010, **1**(1):10.
35. Norman G: **Research in clinical reasoning: past history and current trends**. *Medical Education* 2005, **39**(4):418-427.
36. Schmidt HG, Rikers RMJP: **How expertise develops in medicine: knowledge encapsulation and illness script formation**. *Medical Education* 2007, **41**(12):1133-1139.
37. Erin P. Balogh, Bryan T. Miller, Ball JR: **Improving Diagnosis in Health Care**. In Edited by Board on Health Care Services, Institute of Medicine. Washington, DC: The National Academies Press; 2015.
38. Ceusters W: **Towards A realism-based metric for quality assurance in ontology matching**. In: *Proceedings of Formal Ontology in Information Systems 2006 (FOIS 2006)*: Edited by Bennett B, Fellbaum C. IOS Press 2006: 321.
39. Ceusters W: **Dealing with mistakes in a referent tracking system**. In: *Ontology for the Intelligence Community (OIC-2007); Columbia, Maryland*. 2007: 5-8.
40. Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A: **An Ontological Analysis of Reference in Health Record Statements**. In: *Formal Ontology in Information Systems: Proceedings of the Eighth International Conference (FOIS 2014): 2014*. IOS Press: 289.

Table 1. Definitions based on Smith and Ceusters [25].

Term	**Definition**
INFORMATION CONTENT ENTITY	An ENTITY which is (1) GENERICALLY DEPENDENT on (2) some MATERIAL ENTITY and which is (3) concretized by a QUALITY (a) inhering in the MATERIAL ENTITY and (b) that is_about some PORTION OF REALITY
INFORMATION QUALITY ENTITY	A REPRESENTATION that is the concretization of some INFORMATION CONTENT ENTITY
REPRESENTATION	A QUALITY which is_about or is intended to be about a PORTION OF REALITY
MENTAL QUALITY	A QUALITY which specifically depends on an ANATOMICAL STRUCTURE in the cognitive system of an organism
COGNITIVE REPRESENTATION	A REPRESENTATION which is a MENTAL QUALITY
Relation	**Explanation**
x ***is_about*** *y*	*x refers to or is cognitively directed towards y*. **Domain:** representations; **Range:** portions of reality
x ***concretizes*** *y*	*x* is a QUALITY and *y* is a GENERICALLY DEPENDENT CONTINUANT (GDC) and for some MATERIAL ENTITY *z, x* ***specifically_depends_on*** z at *t* and *y* ***generically_depends_on*** *z* at *t*, and if *y* migrates from bearer *z* to another bearer *w* then a copy of *x* will be created in *w*.
x ***is_conformant_to*** *y*	=def. *x* is an INFORMATION QUALITY ENTITY and *y* is a COGNITIVE REPRESENTATION and there is some GDC *g* such that *x* ***concretizes*** *g* and *y* ***concretizes*** *g*.

Table 2. Key definitions from OGMS used in the analysis

Term	Definition
DISEASE	A DISPOSITION (i) to undergo PATHOLOGICAL PROCESSes that (ii) exists in an ORGANISM because of one or more DISORDERs in that ORGANISM.
DISORDER	A causally relatively isolated combination of physical components that is (a) clinically abnormal and (b) maximal, in the sense that it is not a part of some larger such combination.
DIAGNOSIS	A conclusion of an interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion (diagnostic statement) to the effect that the patient has a DISEASE of such and such a type.
DIAGNOSTIC PROCESS	An interpretive PROCESS that has as input a CLINICAL PICTURE of a given patient and as output an assertion to the effect that the patient has a DISEASE of a certain type.
PATHOLOGICAL PROCESS	A bodily PROCESS that is a manifestation of a DISORDER.
PHENOTYPE	A bodily feature or combination of bodily features of an organism determined by the interaction of the genetic make-up of the organism and its environment.
CLINICAL PHENOTYPE	A clinically abnormal PHENOTYPE.
CLINICAL PICTURE	A representation of a CLINICAL PHENOTYPE that is inferred from the combination of laboratory, image and clinical findings about a given patient.
CLINICAL FINDING	A REPRESENTATION that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof.
MANIFESTATION OF DISEASE	A QUALITY of a patient that is (a) a deviation from clinical normality that exists in virtue of the realization of a disease and (b) is observable.
CLINICAL HISTORY TAKING	An interview in which a clinician elicits a clinical history from a patient or from a third party who is authorized to make health care decisions on behalf of the patient.
CLINICAL HISTORY	A series of statements representing health-relevant features of a patient.

Table 3. Referent tracking tuples true in every scenario

IUI	**Entity**	**Existence period**	**Type**	**Notes**
IUI-1	Mr. Adam Jones	*t1* – the period during which IUI-1 exists	Material Entity	
IUI-2	IUI-1's disease	*t2*	Disposition	
Relationships among particulars				**Notes**
IUI-2	**inheres in**	IUI-1	at *t2*	
IUI-2	**instance of**	UUI-1	at *t2*	UUI-1 is a universal unique identifier that denotes *type 2 diabetes mellitus.* We assume that if something is at any time of its existence an instance of type 2 DM, it is instance of type 2 DM at all times it exists.

Table 4. The entities in Scenario 1

IUI	Entity	Existence period	Type	Notes
IUI-3	Dr. Anne Smith	t3	Human being	
IUI-4	Cognitive system of IUI-3	t4		
IUI-5	An anatomical entity that is part of IUI-4	t5	Anatomical entity	Which anatomical entity and its lifetime cannot be easily specified given current state of neuroscience.
IUI-6	Quality that inheres in IUI-5 and is about IUI-7	t6	Cognitive representation	
IUI-7	The POR that is truth-maker for IUI-8	t7	Configuration	Mr. Jones, his disease, their relationship, and disease's instantiation
IUI-8	Dr. Smith's diagnosis	t8	Diagnosis	ICE concretized by IUI-6 and IUI-10
IUI-9	That which is written down on paper and forms the sentence.	t9	Material entity	*I conclude therefore that Mr. Jones has type 2 diabetes mellitus.*
IUI-10	IQE that inheres in IUI-9.	t10	Information quality entity	The sentence began to exist as soon as ink was laid down on paper, but the IQE did not begin to exist until the sentence was finished.
IUI-11	Dr. Smith's interpretive process	occupies t11	Diagnostic process	Dr. Smith's diagnostic process that led to her diagnosis IUI-8
IUI-12	The clinical picture input into IUI-11	t12	Clinical picture	Dr. Smith's clinical picture as ascertained prior to t6
IUI-13	Dr. Smith writing her diagnosis in the note	occupies t13	Process	

Table 5. Additional temporal entities in Scenario 1.

Temporal identifier	Description	Notes
t14	The interval during which the anatomical entity (IUI-5) is part of the cognitive system (IUI-4)	This interval is not easily specified given the current state of neuroscience. It could be different than t3 and t4.
t15	The interval during which the clinical picture (IUI-12) is used in the interpretive process (IUI-11)	Could be shorter than t11
t16	The point in time at which the cognitive representation (IUI-6) and diagnosis (IUI-8) begin to exist	t16 ends t11. Because the ICE does not exist until the cognitive representation—its first concretization—exists, this is also the point in time at which the diagnosis begins to exist.
t17	The interval during which the cognitive representation (IUI-6) participates in the writing process (IUI-13)	
t18	The interval during which the diagnosis (IUI-8) participates in the writing process (IUI-13)	It is possible that the original cognitive representation (IUI-6) gets copied elsewhere in the brain for reasoning and thus that the ICE continues to participate after the initial cognitive representation
t19	The interval during which that which is written on paper (IUI-10) begins to exist until it exists in full	The writing process begins earlier than the time at which the sentence begins to exist: the author starts the process with getting a pen and paper, any preparation necessary ("clicking" the pen), etc.

Table 6. Relationships among particulars in Scenario 1.

IUI	Relation	IUI	When relation holds in reality	Notes
IUI-4	**part of**	IUI-3	at t4	
IUI-5	**part of**	IUI-4	at t14	All anatomical components in which the cognitive representation inheres are part of the cognitive system. We do not assume the cognitive system is limited to the brain, as the state of neuroscience does not permit such an assumption.
IUI-6	**inheres in**	IUI-5	at t6	
IUI-6	**is about**	IUI-7	at t6	The cognitive representation stands in aboutness to IUI-7 as long as it exists
IUI-6	**is about**	IUI-1	at t6	It is also about Mr. Jones
IUI-6	**is about**	IUI-2	at t6	And about Mr. Jones' disease
IUI-6	**is about**	UUI-1	at t6	And about Type 2 diabetes mellitus
IUI-6	**concretizes**	IUI-8	at t6	It also concretizes the diagnosis
IUI-10	**inheres in**	IUI-9	at t9	The IQE inheres in the sentence on paper
IUI-10	**is about**	IUI-7	at t10	The IQE stands in aboutness to IUI-7
IUI-10	**is about**	IUI-1	at t10	It is also about Mr. Jones
IUI-10	**is about**	IUI-2	at t10	And about Mr. Jones' disease
IUI-10	**is about**	UUI-1	at t10	And about Type 2 diabetes mellitus
IUI-10	**concretizes**	IUI-8	at t10	
IUI-10	**is conformant to**	IUI-6	at t10	Is conformant to the cognitive representation as long as it exists
IUI-3	**agent in**	IUI-11	at t11	
IUI-12	**input into**	IUI-11	at t15	Clinical picture input into IUI-11
IUI-6	**output of**	IUI-11	at t16	Cognitive representation output from IUI-11
IUI-8	**output of**	IUI-11	at t16	Both the diagnosis and its concretization are outputs of IUI-11
IUI-8	**input into**	IUI-13	at t17	The diagnosis is input into writing
IUI-6	**input into**	IUI-13	at t18	As is its cognitive representation
IUI-10	**output of**	IUI-13	at t19	The sentence is output of writing

Table 7. Relationships of representations to portions of reality in Scenario 3: *Incorrect diagnosis*.

Relationships among particulars				Notes
IUI-46	**is about**	IUI-1	at t46	Dr. Jane Miller's cognitive representation is about Mr. Jones
IUI-46	**is about**	IUI-2	at t46	And Mr. Jones' disease
IUI-46	**is about**	UUI-2	at t46	And Type 1 diabetes mellitus (denoted by UUI-2)
IUI-50	**is about**	IUI-1	at t50	Likewise with the IQE inhering in the ink on paper
IUI-50	**is about**	IUI-2	at t50	
IUI-50	**is about**	UUI-2	at t50	
IUI-46	**is misrepresentation of**	IUI-7	at t46	But the cognitive representation is a misrepresentation of the configuration, i.e., it is intended to be about the configuration but fails on the level of compound expression
IUI-50	**is misrepresentation of**	IUI-7	at t50	The same is true of the IQE

Table 8. Six possibilities for a diagnosis failing in aboutness on the level of compound expressions.

Problem	**Where it fails *first***	**Description**
Noninstantation, asserted type exists	Level of compound expression	Disease instantiates a different type than the stated type, but the stated type exists
Noninstantation, asserted type does not exist	Level of reference	Disease instantiates a different type than stated, while the stated type of disease does not exist
Disease nonexistence	Level of reference	The disease instance does not exist
Organism nonexistence	Level of reference	The organism instance does not exist. In this case, there could not be a clinical picture properly inferred and thus it is not a misdiagnosis although it could still be an ICE.
Disease non-inherence	Level of compound expression	The disease inheres in a different organism than the one stated. For example, the doctor mistakenly ascribes Mr. Johnson's hypertension to his twin.
Configuration is not located in that part of spacetime where the diagnosis says it is located.	Level of compound expression	A diagnosis of type 2 diabetes mellitus 5 years ago is wrong because the patient didn't have the disease at that time, even though the patient has type 2 diabetes today. Also, a diagnosis that the patient has an upper respiratory tract infection today when in reality the infection resolved two weeks ago.

Table 9. Additional tuples required to distinguish diagnosing from a lucky guess.

IUI	Entity	Lifetime	Type	Notes
IUI-14	The aggregate of Dr. Smith's cognitive representations of various disease types and their associated types of phenotypes including type 2 diabetes mellitus that he used in the diagnostic process	t20	Aggregate of cognitive representations	
Relationships among particulars				**Notes**
IUI-14	**input into**	IUI-11	at t21	t21 refers to the temporal interval during which IUI-14 participated in the reasoning process. It could start at the same time as t11 or after t11, and end at the same time as or before t11.

Figure legends

Figure 1. The configuration of Mr. Jones, his disease, and type 2 diabetes mellitus

Figure 2. Diagram of diagnostic process, its inputs, a correct diagnosis, its concretization, and the configuration that that the concretization is about

Figure 3. Misdiagnosis of type of disease. The diagnosis is individually about the patient, the disease, and the incorrectly diagnosed disease type Y, but it is not about the configuration of patient, disease, and disease type X.

Figure 1

Click here to download Figure Fig 1.png

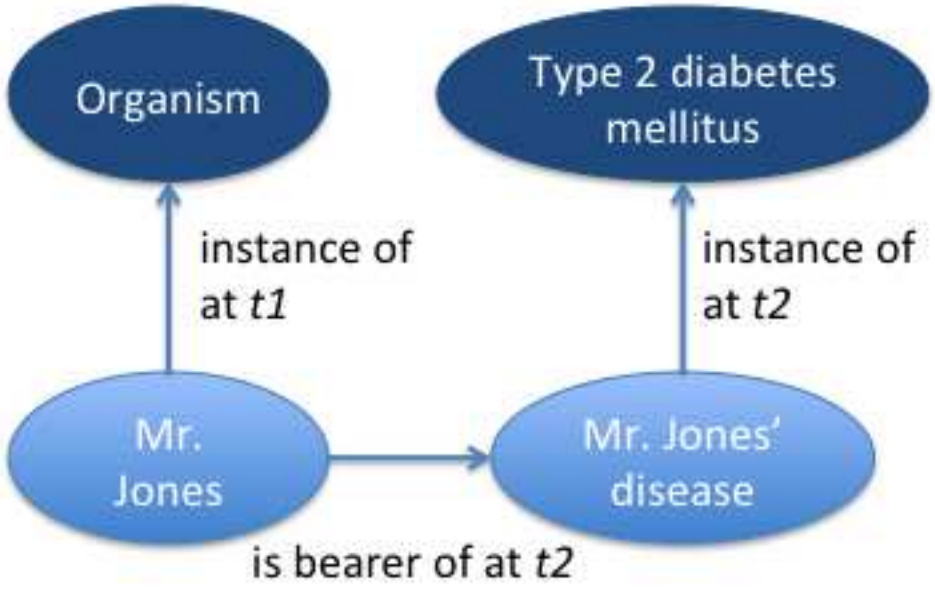


Figure 2

Click here to download Figure Fig 2.png

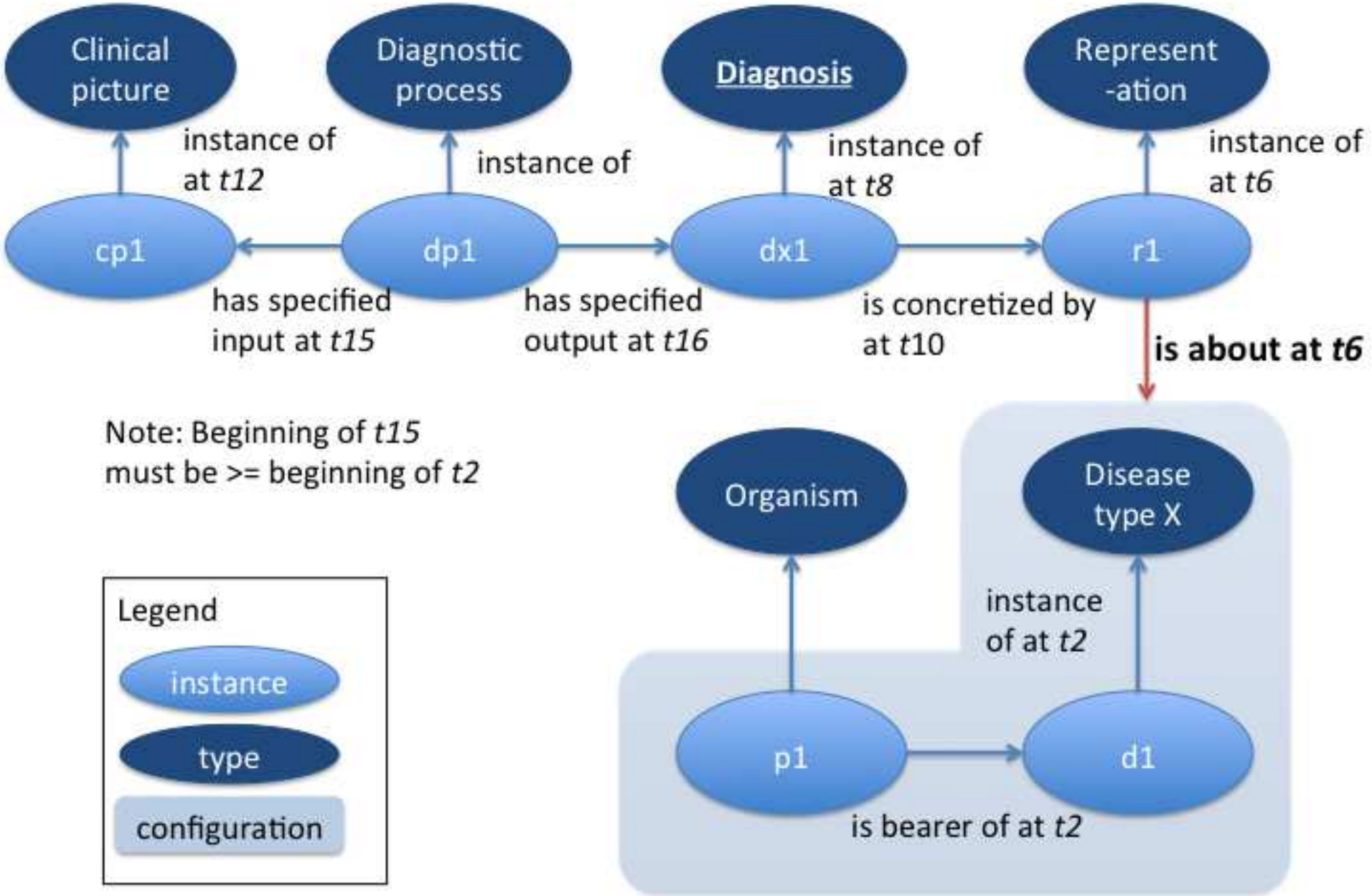
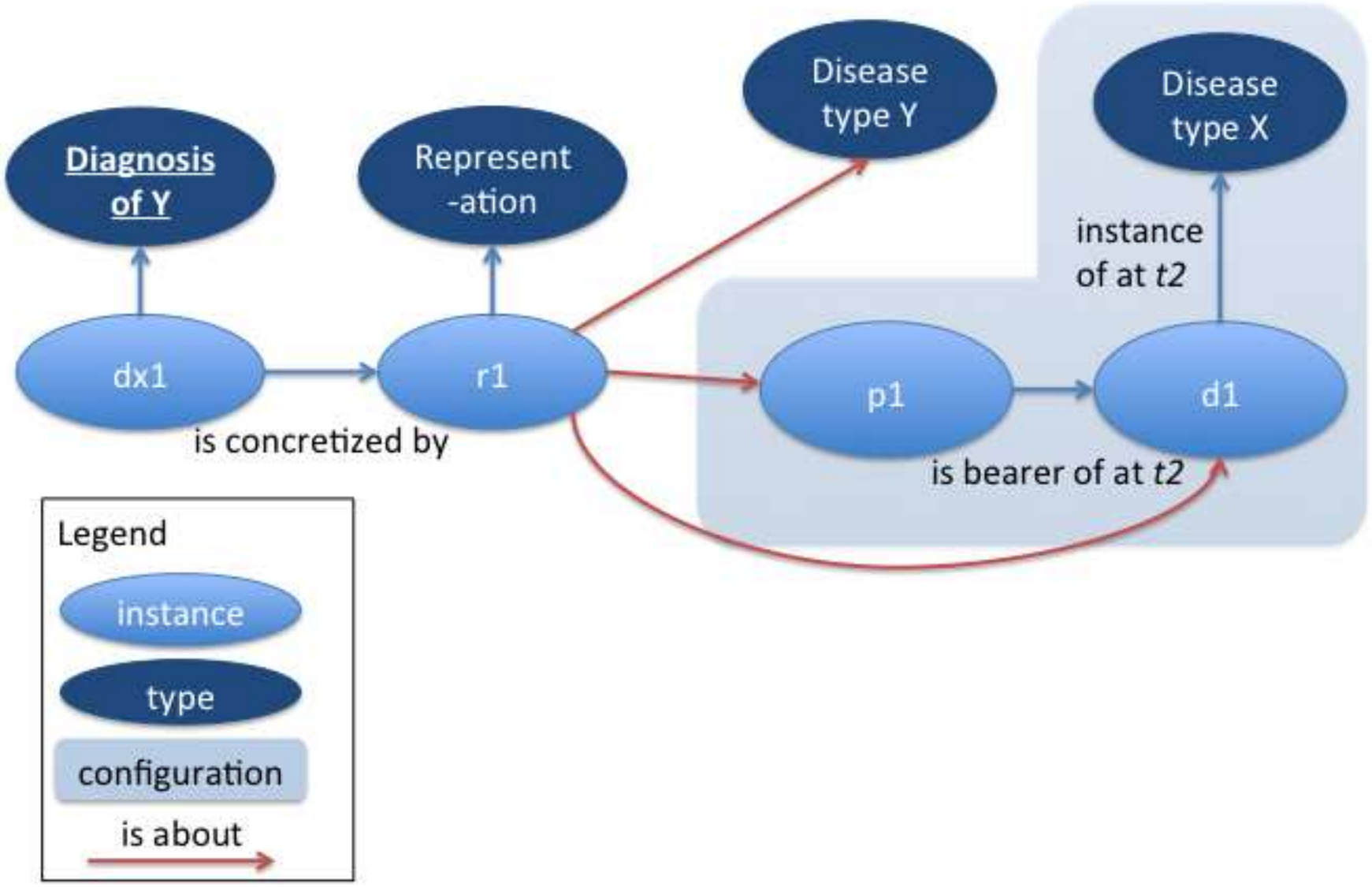


Figure 3

Click here to download Figure Fig 3.png



Additional file 1

Click here to access/download
Supplementary Material
additional file 1 v3.docx

Additional file 2

Click here to access/download
Supplementary Material
additional file 2 v3.docx

Additional file 3

Additional file 4

From: em.jbsm.0.4da259.89e13da5@editorialmanager.com
<em.jbsm.0.4da259.89e13da5@editorialmanager.com> on behalf of Journal of Biomedical Semantics Editorial Office <em@editorialmanager.com>
Sent: Saturday, September 3, 2016 2:42 PM
To: Hogan,William
Subject: Decision has been reached on your submission to Journal of Biomedical Semantics - JBSM-D-15-00018R3

JBSM-D-15-00018R3
Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis.
William R. Hogan, MD, MS; Werner Ceusters, MD
Journal of Biomedical Semantics

Dear Dr. Hogan,

I am pleased to inform you that your manuscript "Diagnosis, misdiagnosis, lucky guess, hearsay, and more: an ontological analysis." (JBSM-D-15-00018R3) has been accepted for publication in Journal of Biomedical Semantics.

I did have some considerations on the discussion between the reviewer and the authors. I don't want to open this discussion again and concluded in favour of the authors; the future may convey expected advantages or further controversy.

Before publication, our production team will check the format of your manuscript to ensure that it conforms to the standards of the journal. They will be in touch shortly to request any necessary changes, or to confirm that none are needed.

Any final comments from our reviewers or editors can be found, below. Please quote your manuscript number, JBSM-D-15-00018R3, when inquiring about this submission.

We look forward to publishing your manuscript and I do hope you will consider Journal of Biomedical Semantics again in the future.

Best wishes,

Dietrich Rebholz-Schuhmann
Journal of Biomedical Semantics
http://www.jbiomedsem.com/