Barton et al. Journal of Biomedical Semantics (2017) 8:1
DOI 10.1186/s13326-016-0099-4

Journal of Biomedical Semantics

RESEARCH Open Access

An ontological analysis of medical Bayesian indicators of performance

Adrien Barton[1,5*†], Jean-François Ethier[1,4,5†], Régis Duvauferrier[2,3] and Anita Burgun[4]

Abstract

Background: Biomedical ontologies aim at providing the most exhaustive and rigorous representation of reality as described by biomedical sciences. A large part of medical reasoning deals with diagnosis and is essentially probabilistic. It would be an asset for biomedical ontologies to be able to support such a probabilistic reasoning and formalize Bayesian indicators of performance: sensitivity, specificity, positive predictive value and negative predictive value. In doing so, one has to consider that not only the positive and negative predictive values, but also sensitivity and specificity depend upon the group under consideration: this is the "spectrum effect".

Methods: The sensitivity value of an index test *IT* for a disease *M* in a group **g** is identified with the proportion of people in **g** who have *M* who would get a positive result to *IT* if the test *IT* was realized on them. This value can be estimated by selecting a reference test *RT* for *M* and a sample **s** of **g**, and measuring the proportion, among members of **s** having a positive result to *RT*, of those who got a positive result to *IT*. Similar approximation strategies hold for prevalence, specificity, PPV and NPV. Indicators of diagnostic performances and their estimations are formalized in the context of the OBO Foundry, built on the realist upper ontology Basic Formal Ontology (BFO).

Results: Entities and relations from the Ontology for Biomedical investigations (OBI) and the Information Artifact Ontology (IAO) are used and complemented to represent reference tests and index tests, tests executions, tests results and the relations involving those entities, as well as the values of indicators of performance and their estimates. The computations taking as input several estimates of an indicator of performance to produce a finer estimate are also represented. The value of e.g. sensitivity estimates should be dissociated from the real sensitivity value – which involves possible, non-actual conditions, namely the result a person would get if a medical test would be performed on her. Such conditions could not be directly represented in a realist ontology, but a representation is proposed that introduces only actual entities by considering a disposition whose probability value is the real sensitivity value. A sensitivity estimate is a data item which is about such a disposition.

Conclusions: This model provides theoretical basis for the representation of entities supporting Bayesian reasoning in ontologies.

Keywords: Sensitivity, Specificity, Medical test, Spectrum effect, Disposition, Realist ontology, Informational entity

Background

Definition of indicators of performance

Biomedical ontologies aim at providing the most exhaustive and rigorous representation of reality as described by biomedical sciences. A large part of medical reasoning deals with diagnosis and is essentially probabilistic. It would be an asset for biomedical ontologies to be able to support such a probabilistic reasoning.

Ledley and Lusted's seminal article [1] on Bayesian reasoning in medicine defines different kinds of probabilistic entities. Consider for example the simple case of an instance of test of type *IT* (for "index test" – a test whose accuracy is being measured) aiming at detecting if a patient in a group **g** has an instance of disease of type *M*.[1] The performance of test *IT* in diagnosing *M* can be quantified by the positive predictive value of this test, hereafter abbreviated PPV, defined by the *Oxford Handbook of Medical*

* Correspondence: adrien.barton@gmail.com
†Equal contributors
[1]Département de médecine, Université de Sherbrooke, Sherbrooke, Québec, Canada
[5]Centre de recherche du CHUS, CIUSSS de l'Estrie-CHUS, Sherbrooke, Québec, Canada
Full list of author information is available at the end of the article

Statistics [2] as the "proportion of tested positives who are true positives" and by the negative predictive value, hereafter abbreviated NPV, defined as the "proportion of tested negatives who are true negatives". These values provide the probability that a patient has or not the disease, depending upon the result (positive or negative) to the test.

However, such values depend on some characteristics of the patient. If a patient received a positive test, the probability that he has the disease can for example depend upon his sex, his status of smoker or non-smoker, and other biological or environmental parameters. In particular, it depends on the prevalence of the disease among the group of persons with those characteristics.

Therefore, the statistical data communicated in the medical literature for a test are generally not the positive and negative predictive values, but the so-called "sensitivity" and "specificity". The *Oxford Handbook of Medical Statistics* defines sensitivity as "the proportion of those who have the disease who are correctly identified by the test as positive" ([2], p. 340) and specificity as "the proportion of those who do not have the disease who are correctly identified by the test as negative". The PPV and NPV can be computed on the basis of the prevalence Prev, sensitivity Se and specificity Sp thanks to the following Bayesian equations:

$$PPV = \frac{Prev.Se}{Prev.Se + (1-Prev)(1-Sp)}$$

$$NPV = \frac{(1-Prev).Sp}{Prev.(1-Se) + (1-Prev).Sp}$$

In the remainder of the article, sensitivity, specificity, PPV and NPV will be called "(Bayesian) indicators of performance" and abbreviated "IPs".

In the wake of Ledley and Lusted [1] the sensitivity and specificity values have often been considered as depending only on the pathophysiological characteristics of the disease and of the test, and thus as being independent of the group of people under consideration. However, sensitivity and specificity values do in fact depend upon the group under consideration: this is the "spectrum effect" [3].

The spectrum effect

If *IT* is an index test and *M* is a disease, let's introduce $f_1(IT,M)$ as "the proportion of individuals who get a positive result to *IT*, among individuals who have *M*", which fits with the usual definition of sensitivity (as provided by [2]). The main problem with this definition is that it does not specify the reference population. "The individuals who have *M*" are part of which population: the population in a given sample? The population of a specific country? The whole human population? Ledley and Lusted [1] considered that sensitivity and specificity depend upon pathophysiological characteristics of the disease, but not upon the population in consideration. If this was the case, the proportion of people tested positive among the diseased would be the same in any group under consideration – abstracting from statistical fluctuations due to randomness. However, as has been recognized by the medical literature, but regularly omitted, this hypothesis is false for at least two reasons. First, most tests are not inherently dichotomous but rely on a categorization of individuals based on continuous traits [3]. Second, various populations can express various disease characteristics (such as various degrees of severity [4]) that will influence the chance to get a positive result to a test.

The latter can be illustrated with the following example. Suppose that around 80 % of people having rheumatoid arthritis have a rheumatoid factor (RF), and would with certainty receive a positive result to a test that would perfectly[2] detect this factor; and that the remaining 20 % do not have a rheumatoid factor, and would receive a negative result (yet do have the disease). The diseased population is then composed of two subgroups: a subgroup $\mathbf{sg_1}$ whose members would all get for sure a positive result to *IT*, and a subgroup $\mathbf{sg_2}$ whose members would all get for sure a negative result (see Fig. 1). The sensitivity calculated in this example would be 80 %.

Nevertheless, in reality, those proportions vary based upon various characteristics of the patients. For example, RF presence increases with age at onset of disease in juvenile arthritis [5]. As a result, the sensitivity of a test for RF will increase according to the age of the individuals of the population being tested. Its sensitivity will be lower in younger patients and higher in older patients.

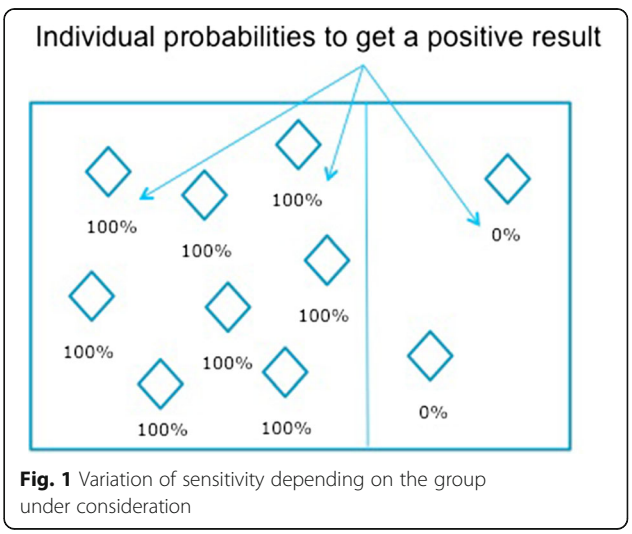

Fig. 1 Variation of sensitivity depending on the group under consideration

Therefore, f_1 is not a well-defined function: the value of the proportion does not depend only upon *IT* and *M*, but also upon the population **g** under consideration (which could be, for example, the whole human population, the Canadian smoker population, the female population, etc.). This is the "spectrum effect", which can also be manifested, for example, as a dependence of sensitivity and specificity on the degree of severity of the disease in the group under consideration [4].

The sensitivity can therefore depend on the group **g** under consideration. A better candidate than f_1(*IT,M*) to the definition of the sensitivity value would be the function f_2(**g**,*IT,M*) defined as "the proportion[3] among people in **g** who have *M* of those who would get a positive result to *IT if the test IT was realized on them*" – the mention in italic is necessary, as a test *IT* will not be realized on all individuals who have *M*, but on a sample only. The next part will distinguish three related entities: the real sensitivity[4] value, its estimates, and the measurements of proportion in samples. It will also explain how such entities should be distinguished in an ontology of IPs.

Methods

Proportion measurement in a sample

It is impossible to know f_2(**g**,*IT,M*) with certainty in practice, for two reasons. The first reason is that it is often not possible to determine with certainty, through reasonable means, whether a given person has the disease *M* or not; in some cases, the only way to be certain would be to perform an autopsy on the deceased patient. Therefore, one needs to use a "reference test", which is the best diagnostic test that is reasonable to perform in the present context (for more on the distinction between a reference test and the associated disease, see section "The challenge of representing indicators of performance in an ontology" below).

If the patient receives a positive result to this reference test, it will be concluded that he has the disease; if he receives a negative result, it will be concluded that he does not have it. But those inferences can be wrong: the reference test might lead to a positive result for a non-diseased person, or a negative result for a diseased person. If *RT* is a reference test for *M* and *IT* is an index test (of unknown accuracy) for *M*, then one can define the function f_3(**g**,*IT,RT*) as "the proportion, among individuals of **g** who would get a positive result to *RT if the test RT had been performed on them*, of people who would get a positive result to *IT if the test IT was realized on them*". Since *RT* is a reference test for *M*, f_3(**g**,*IT,RT*) approximates f_2(**g**,*IT,M*). Both values can differ though: this is a first epistemic limit to the knowledge of f_2(**g**,*IT,M*).

On top of this, f_3(**g**,*IT,RT*) is not directly measurable. As a matter of fact, a test *IT* is never realized on a population as large as e.g., the whole population of smokers, or the whole male population. It is however possible to approximate f_3(**g**,*IT,RT*) by performing both tests *IT* and *RT* on individuals in a sample **s** judged as being representative of the population **g**. Let's define f_4(**s**,*IT,RT*) as "the proportion, among members of **s** who got a positive result to *RT*, of those who got a positive result to *IT*". If **s** is a representative sample of **g**, then f_4(**s**,*IT,RT*) does approximate f_3(**g**,*IT,RT*) – and thus, by transitivity, does approximate f_2(**g**,*IT,M*). Note that as long as the sample **s** is not perfectly representative of **g**, f_4(**s**,*IT,RT*) will differ at least slightly from f_3(**g**,*IT,RT*) (which also differs from f_2(**g**,*IT,M*)): this is a second limit to the knowledge of f_2(**g**,*IT,M*).

Let's illustrate those two limits of estimations with a study [4] which analyzes the quality of the Neer test (here written *IT'*) for diagnosing the shoulder impingement syndrome (written *M'*), a syndrome that is characterized by rotator cuff muscles inflammation near the sub-acromial space. In this study, the Neer test *IT'* is realized on a sample (written **s'**) of 552 patients, judged as representative of the target population (**g'**). Park et al. [4] take as reference test (*RT'*) the surgical observation. Here, f_4(**s'**,*IT',RT'*) is the proportion of people in the sample who have received a positive result to the Neer test, among those diagnosed as positive by surgical operation. f_4(**s'**,*IT',RT'*) approximates f_3(**g'**,*IT',RT'*), namely the proportion of individuals *in the target population* **g'** who would get a positive result to the Neer test among those who would get a positive result by surgical observation, if those tests were performed on them. Finally, f_3(**g'**,*IT',RT'*) itself approximates f_2(**g'**,*IT',M'*), which is the proportion of individuals in **g'** who would receive a positive Neer test result among those *who have an impingement syndrome*. Thus, f_4(**s'**,*IT',RT'*) approximates f_2(**g'**,*IT',M'*).

Note that similar approximation strategies hold for prevalence, specificity, PPV and NPV. Concerning e.g. specificity, one could thus define f'_2(**g**,*IT,M*) as "the proportion[5] among people in **g** who don't have *M* of those who would get a negative result to *IT if the test IT was performed on them*"; and f'_4(**s**,*IT,RT*) as "the proportion, among members of **s** who got a negative result to *RT*, of those who got a negative result to *IT*". Thus, f'_4(**s**,*IT,RT*) approximates f'_2(**g**,*IT,M*).

Sensitivity value and sensitivity estimates

Now that those definitions have been given, we can determine which entity the word 'sensitivity' refers to in the medical literature. At first sight, this term might appear polysemic. To illustrate this, let's consider a study which evaluates the quality of an exercise test in the diagnosis of coronary artery disease, and claims: "The sensitivity varied substantially according to sex (women

30 % and men 64 %)" [6]. On one hand, the statement "sensitivity varies substantially according to the sex" suggests that sensitivity depends on the target population **g** in consideration, and that there is a sensitivity value for the female population, and another one for the male population. This formulation thus suggests that sensitivity value is given by the function f_2(**g**,*IT,M*). However, the value 30 % assigned to the sensitivity of the test for women refers to a proportion which has been measured by the authors in a sample of 37 women, using coronary angiography as a reference test. This might thus suggest that the sensitivity value is in fact given by the function f_4(**s**,*IT,RT*)

However, two arguments suggest that the sensitivity value should be interpreted as f_2(**g**,*IT,M*) rather than f_4(**s**,*IT,RT*). First, the value which is ultimately relevant for medical practice is f_2(**g**,*IT,M*): if **s** is a sample of **g** and *RT* is a reference test for *M*, f_4(**s**,*IT,RT*) is of interest for the medical practitioner only insofar as it provides an information on the disease *M* and the target population **g** from which the sample is taken – that is, insofar as it provides an estimate of f_2(**g**,*IT,M*). Indeed, the fact that a few people who got a positive result to *RT* in a given sample have got a positive or negative result to a test *IT* has medical relevance only insofar as it teaches us something about how *diseased* people *in the target population* (not only in the sample) will react to this test *IT*.

Second, the sensitivity value is usually given with a 95 % confidence interval (see e.g., [7] or [8]), which estimates the likely range of error in determining the sensitivity value. But f_4(**s**,*IT,RT*) can be measured with certainty,[6] and thus the confidence interval cannot characterize the uncertainty on our knowledge of f_4. On the other hand, there is some uncertainty on the knowledge of f_2(**g**,*IT,M*) and f_3(**g**,*IT,RT*), as they are estimated on the basis of f_4(**s**,*IT,RT*). Therefore, the 95 % confidence interval would characterize the uncertainty on the knowledge of f_3(**g**,*IT,RT*), which is taken as a proxy for f_2(**g**,*IT,M*).[7]

Thus, those two arguments suggest that the term "sensitivity" should refer to f_2(**g**,*IT,M*) – which is relative to a disease and a target population – rather than to f_4(**s**,*IT,RT*) – which is relative to a reference test and a sample.[8] As for f_4(**s**,*IT,RT*), it can be interpreted as the value of a measurement of proportion in a sample, which provides an estimate of the sensitivity value.

Therefore, a sentence such as "The sensitivity varied substantially according to sex (women 30 % and men 64 %)" should, more rigorously, be formulated as: "The sensitivity varies substantially depending on the sex: through measurement of proportions in samples, its value was estimated to be 30 % for the women, and 64 % for the men". We could prefer the first formulation, more compact, for practical reasons; but it is important to remember that it is only a shortcut for the second formulation.

Accordingly, we will need to dissociate three different kinds of entities. First, tests execution on a sample **s**, referring more precisely to the process of performing tests *IT* and *RT* and measuring the numbers of true positive, false positive, true negative and false negative as operationalized by *IT* and *RT* - for example, the false positive are people who are tested positive by the index test *IT* but negative by the reference test *RT* in the sample **s**. Second, the proportion of true positives among positives (as given by the reference test) is relative to the index test, the reference test and the sample, and its value is given by the function f_4(**s**,*IT,RT*); as such, it provides an estimate of the sensitivity value. Third, the "real sensitivity", which is relative to an index test, a disease and a population **g**, and whose value f_2(**g**,*IT,M*) is given by the proportion of people in the group who would have a positive result to the test *IT* among those who are diseased. The real sensitivity would provide a better information than a sensitivity estimate on the probability that a random member of the group **g** would get a positive test result, in case he has the disease. However, its value f_2(**g**,*IT,M*) cannot be known with certainty, contrarily to the value of the sensitivity estimate f_4(**s**,*IT,RT*).

More generally, those considerations can be adapted to other indicators of performance (specificity, PPV and NPV), as well as the prevalence. In particular, f'_2(**g**,*IT,M*) should refer to the real specificity value, whereas f'_4(**s**,*IT,RT*) can be interpreted as the value of a measured proportion in a sample that provides an estimate of the real specificity value. In particular, *real* sensitivity, specificity, PPV and NPV, as we have defined them above, depend neither on the sample nor on the reference test. However, they are estimated on the basis of proportion *measurements* which depend both on the sample and the reference test. Accordingly, when a study [9] mentions "cadaveric prevalence" of the rotator cuff tears, this expression should be understood as a linguistical shortcut denoting a proportion measurement in a sample when the cadaverical analysis is adopted as reference test; and the "radiological prevalence" should be understood as a proportion measurement when the radiological analysis is adopted as reference test. The real prevalence, however, does not depend on the reference test.

Aggregation of sensitivity estimates

Finally, we need to add a last layer to this model. Approximations of sensitivity taken in different samples, with different index tests, can be combined in order to build a finer estimate of sensitivity for a more encompassing category of index tests. Consider for example the meta-analysis [7] which assess the quality of peripheral thermometers in detecting fever. They use as reference test a pulmonary artery catheter, and consider 29 studies

assessing the sensitivity and specificity of those devices. Combining those values, they come up with an estimate of 0.64 for the sensitivity and of 0.96 for the specificity.

The challenge of representing indicators of performance in an ontology

To the extent that they aim at representing biomedical knowledge and enabling medical reasoning, biomedical ontologies should provide a formalization of IPs as well as the prevalence, by dissociating e.g. the real sensitivity from the sensitivity estimates, and the process leading to those estimates. This article will introduce such a formalization in the context of the OBO Foundry [10], one of the most massive set of interoperable ontologies in the biomedical domain, built on the upper ontology Basic Formal Ontology (BFO) 1.1 [11].

BFO endorses a realist methodology, which carefully dissociates material entities (such as disorders) from informational entities (such as diagnosis). In common medical practice, a disease may be diagnosed in ideal circumstances by a given gold standard test, which can be defined as the most accurate reference test; but the disease, the diagnosis, and the result to a gold standard test are three different entities that should be distinguished. As a matter of fact, many human diseases already existed a few thousands of years ago, much before they could be diagnosed. Moreover, a diagnosis can be wrong or imprecise. Finally, a given gold standard can be later replaced by a better one: this shows that the disease cannot be defined by a positive result to a gold standard - otherwise, there could not be, by definition, a "better" gold standard. Thus, while a diagnosis of a disease represents the best knowledge by some health or research professional of the presence of the disease in a particular patient, a diagnosis is not equivalent to a disease: it is rather "about" a disease. This formalization is compatible with IAO (Information Artifact Ontology [16]) and OGMS (Ontology for General Medical Sciences).

The question of how probabilistic notions can be represented in ontologies has been tackled from different perspectives in the past. For example, [12] has proposed the alternative PR-OWL format that extends the classical OWL format; we take here a different approach, which does not aim at changing the OWL format. Soldatova and colleagues [13] have described a model in which probabilities can be assigned to research statements. We build here upon an alternative approach [14], in which probabilities can be assigned to dispositions.

Sensitivity and specificity have been recently introduced in the Ontology of Biological and Clinical Statistics (OBCS [15]) as subclasses of *Data item*. We will partly endorse and refine this classification, by considering estimates of sensitivity and specificity as subclasses of *Data Item*, and extend this classification to PPV and NPV. A data item, as defined by the Information Artifact Ontology (IAO) [16], is intended to be a truthful statement about something. In order to formalize IPs, one should thus clarify which entities in the real world they are about.

Proportion measurements are data items that are obtained from some processes named "proportion measures", which involve performing two kinds of tests (the index test and the reference test) in a sample. On the other hand, we have defined a real sensitivity value $f_2(\mathbf{g},IT,M)$ as the proportion of people who would get a positive result by *IT* among those who have the disease *M*. But note here the conditional structure: what is referred to is the proportion of true positives among diseased *if IT* was performed on them. In realistic situations, however, as explained above, the sensitivity value will be estimated by performing the test on a sample of the population only – not the entire population **g**; thus, $f_2(\mathbf{g},IT,M)$ is the value of a non-actual proportion.[9] However, possible-but-non-actual situations cannot be straightforwardly represented in a realist ontology like BFO. To solve this problem, we will formalize the real IP value as the probability assigned to a disposition borne by an instance of group of individuals; and estimates of IPs as data items which are about such a disposition. This will provide a formal characterization of IPs and their estimates based on proportion measurements.

Results

The formalization that will be presented here can be visualized on Fig. 2 and Fig. 3, in which classes are in rectangles, instances in boxes with rounded edges, and the numerical value assigned by datatype properties in ellipses. Unless specified otherwise, all the relations used here belong to BFO 1.1 [11].

Test results and sensitivity estimate

Let us first start with the formalization of test results and the IP estimates they lead to (see Fig. 1).[10] A *Medical_test* will be here considered as a subclass of *Planned_process* (as defined by OBI, the Ontology for Biomedical Investigations [17]) which consists in the observation of a given feature to infer the presence of another feature – in the case of interest, a pathological entity such as a disease. Consider a medical test[11] IT_1 and a disease *M*:

Medical_test is_a Planned_process
IT_1 *is_a Medical_test*
M is_a Disease

Suppose that we are interested in the sensitivity and specificity of test IT_1 for diagnosing *M* in a group $\mathbf{g_1}$. This group $\mathbf{g_1}$ will be formalized as a collection of humans (for more on collections, see [18]). To estimate this sensitivity and specificity, one can select a sample $\mathbf{s_1}$

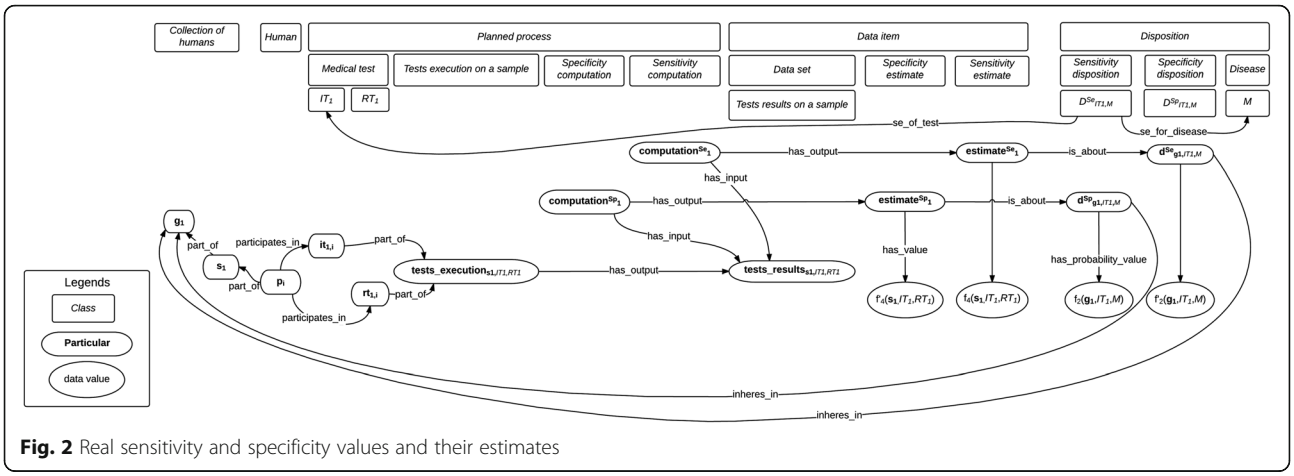

Fig. 2 Real sensitivity and specificity values and their estimates

considered to be representative of $\mathbf{g_1}$ (which will be called the reference class). Thus:

g_1 instance_of *Collection_of_humans*
s_1 instance_of *Sample_of_humans*
Sample_of_humans is_a Collection_of_humans
s_1 part_of g_1

Let's now introduce the class of tests RT_1 which are reference tests for M:

RT_1 *is_a Medical_test*

$\mathbf{s_1}$ is composed of n humans, named $\mathbf{p_1}$, $\mathbf{p_2}$,…,$\mathbf{p_n}$. Two[12] tests will be performed on each $\mathbf{p_i}$: an instance of RT_1, named thereafter $\mathbf{rt_{1,i}}$, and an instance of IT_1, named $\mathbf{it_{1,i}}$; thus, for every i between 1 and n:

p_i instance_of *Human*
p_i part_of s_1
p_i participates_in $rt_{1,i}$
p_i participates_in $it_{1,i}$

We introduce $\mathbf{tests\_execution_{s1,}}_{IT1,RT1}$ which has as part all the tests $\mathbf{rt_{1,i}}$ and $\mathbf{it_{1,i}}$ for i between 1 and n and the recording of which members of the sample are true positives (those who have been tested positive both by IT_1 and RT_1), true negatives (those who have been tested negative both by IT_1 and RT_1), false positives (those who have been tested positive by IT_1 but negative by RT_1)

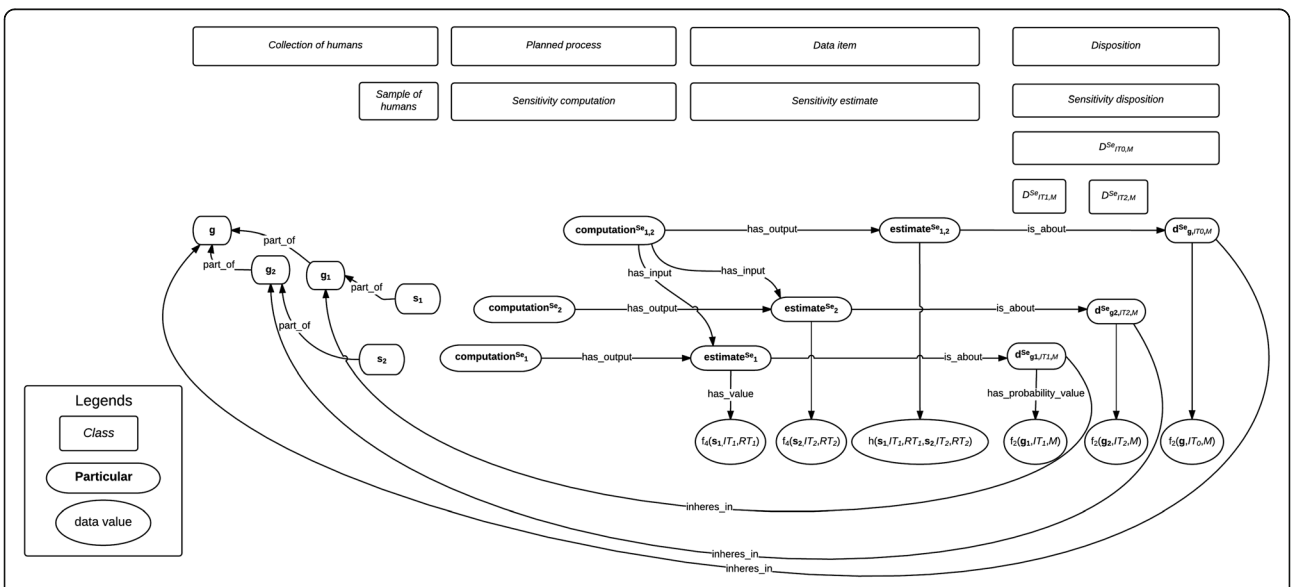

Fig. 3 Aggregation of several sensitivity estimates

and false negatives (those who have been tested negative by IT_1 but positive by RT_1). This recording leads (OBI:-**has_specified_output**) to the creation of the instance of *Data_set* named $\mathbf{tests\_results_{s1,}}_{IT1,RT1}$:

tests_execution$_{s1,IT1,RT1}$ instance_of *Planned_process*
rt$_{1,i}$ part_of tests_execution$_{s1,IT1,RT1}$
it$_{1,i}$ part_of tests_execution$_{s1,IT1,RT1}$
tests_results$_{s1,IT1,RT1}$ instance_of *Data_set*
tests_execution$_{s1,IT1,RT1}$ has_specified_output
tests_results$_{s1,IT1,RT1}$

The $\mathbf{tests\_results_{s1,}}_{IT1,RT1}$ will then serve as input (OBI:**has_specified_input**) to a planned process noted $\mathbf{computation_1^{Se}}$ which computes a sensitivity estimates noted $\mathbf{estimate_1^{Se}}$, by calculating the proportion of true positives among positives:[13]

computation$_1^{Se}$ is_a *Planned_process*
estimate$_1^{Se}$ is_a *Data_item*
computation$_1^{Se}$ has_specified_input
tests_results$_{s1,IT1,RT1}$
computation$_1^{Se}$ has_specified_output estimate$_1^{Se}$

Finally, we can use the datatype property OBI:**has_specified_value** to relate $\mathbf{estimate_1^{Se}}$ with its numerical value $f_4(\mathbf{s_1},IT_1,RT_1)$:

estimate$_1^{Se}$ has_specified_value $f_4(s_1,IT_1,RT_1)$

Similar strategies can hold for representing Specificity, PPV and NPV and their estimates.[14]

Aggregation of sensitivity estimates

We will now show how various sensitivity estimates can be aggregated for a finer sensitivity estimate (cf. Fig. 3). Suppose that we have another sample $\mathbf{s_2}$ (also a **part_of g**), composed of n' humans named $\mathbf{q_1}$, $\mathbf{q_2}$, ..., $\mathbf{q_{n'}}$. We can perform another measure of sensitivity for a related (possibly identical to IT_1) index test IT_2 for M in **g** on this sample, using a related (possibly identical to RT_1) reference test RT_2, by performing instances of RT_2 named $\mathbf{rt_{2,}}_j$ (for j between 1 and n') and instances of IT_2 named $\mathbf{it_{2,j}}$ on each member $\mathbf{q_j}$ of $\mathbf{s_2}$. One can then define the entity $\mathbf{tests\_execution_{s2,}}_{IT2,RT2}$ as a planned process which has as part those tests $\mathbf{rt_{2,}}_j$ and $\mathbf{it_{2,j}}$, and which has as output $\mathbf{tests\_results_{s2,}}_{IT2,RT2}$; the latter serves as input to another computation of sensitivity $\mathbf{computation_2^{Se}}$, which has as output another estimate of sensitivity $\mathbf{estimate_2^{Se}}$, to which the value $f_4(\mathbf{s_2},IT_2,RT_2)$ can be assigned (the latter being the proportion, among people who have been tested positive by RT_2 in $\mathbf{s_2}$, of people who had a positive result to IT_2).

As explained earlier, various sensitivity estimates can be combined to estimate the value of the sensitivity of a test for M in **g**. If IT_1 and IT_2 on one hand, and RT_1 and RT_2 on the other hand, are similar enough (in particular, if they are identical), those results might be gathered to come up with a finer estimate of the sensitivity value. More specifically, if IT_1 and IT_2 can be subsumed under a common index test class IT_0, and RT_1 and RT_2 can also be subsumed under a common reference test class RT_0, then their values can be compiled mathematically (for example by meta-analysis methods) to come up with the value of a (hopefully finer) estimate named $\mathbf{estimate_{1,2}^{Se}}$, whose value is given by a function $h(\mathbf{s_1},IT_1,RT_1,\mathbf{s_2},IT_2,RT_2)$. This can be generalized to the aggregation of more than two former estimates.

We can here introduce a planned process of computation of sensitivity named $\mathbf{computation_{1,2}^{Se}}$, which takes as input both $\mathbf{estimate_1^{Se}}$ and $\mathbf{estimate_2^{Se}}$, and the output of such a process, a data item named $\mathbf{estimate_{1,2}^{Se}}$:

computation$_{1,2}^{Se}$ instance_of *Planned_process*
estimate$_{1,2}^{Se}$ instance_of *Data_item*
computation$_{1,2}^{Se}$ has_specified_input estimate$_1^{Se}$
computation$_{1,2}^{Se}$ has_specified_input estimate$_2^{Se}$
computation$_{1,2}^{Se}$ has_specified_output estimate$_{1,2}^{Se}$
estimate$_{1,2}^{Se}$ has_specified_value $h(s_1,IT_1,RT_1,s_2,IT_2,RT_2)$

We will not aim at giving the details of this function h, which is the responsibility of the statistician, not the ontologist – who focuses on how to represent such values.

Finally, since $\mathbf{estimate_1^{Se}}$ or $\mathbf{estimate_{1,2}^{Se}}$ are informational entities, they must be about some entities. To determine what those entities are about, we will need to formalize the entity to which is assigned the "real sensitivity value".

Real sensitivity value

As said earlier, estimates of sensitivity of IT for M in **g** aim at estimating the real sensitivity value, which is given by the proportion of members of **g** who would get a positive result to IT among those who have M. However, the condition of performing the test IT on the members of **g** is never realized, because the test is performed (at best) on one or several samples of the population, not on the whole population **g**: the performance of test IT on the members of **g** is a *possible* (leaving aside practical difficulties), *non-actual* condition. Interpreting specificity, PPV, and NPV along the former lines would also imply such possible, non-actual conditions.

BFO's realist methodology [19] implies that all instances should be *actual* entities. Thus, one cannot represent directly such a possible-but-not-actual condition in an ontology based on BFO. In order to solve this difficulty, we will introduce a strategy named "randomization", which will clarify the nature of the real sensitivity value as a probability assigned to an actual entity, namely a disposition. This will also clarify what an estimate of

sensitivity is about, namely about this disposition. Thus, it will enable to represent IPs in a realist fashion, compliant with BFO's methodology.

From proportions to objective probabilities: the randomization strategy

We will explain first how the proportion of a subgroup in a group can be formalized as a probability value assigned to a disposition; this will help explaining later how the proportion of a subgroup in a group undergoing a possible, non-actual condition can be formalized along similar lines.

Dispositions are entities that can exist without being manifested; an example of disposition is the fragility of a glass, which can exist even when the glass does not break. We will use Röhl & Jansen's model of disposition [20] in BFO, which associates to every instance of disposition one or several instances of realizations, and one or several instances of triggers (a trigger is the specific process that can lead to a realization occurring). In this model, the fragility of a glass is a disposition of the glass to break (the breaking process is the realization) when it undergoes some kind of stress (the process of undergoing such a stress is the trigger); this disposition inheres in the glass. Starting with the definition of these entities and their relations at the instance level, Röhl & Jansen proceed to formalize them at the universal level. Previous work [14] has shown how to adapt this model to probabilistic dispositions. Thus, an instance of balanced coin is the bearer of an instance of disposition to fall on heads (the realization process) when it is tossed (the trigger process), to which an objective probability 1/2 can be assigned.

We will now extend the scope of this model to the situation at hand. Consider the prevalence Prev(**g**,*M*), which was defined above as the proportion of persons having *M* in the actual population **g**. We can define the disposition $\mathbf{d}^{\text{Prev}}_{\mathbf{g},M}$, borne by the group **g**, that a person randomly drawn in **g** has *M*. More specifically, let's write $T_{\mathbf{g}}$ the process "randomly drawing a person in **g**", and $R_{\mathbf{g},M}$ the process "drawing by $T_{\mathbf{g}}$ someone who has *M*": the triggers of $\mathbf{d}^{\text{Prev}}_{\mathbf{g},M}$ are instances of $T_{\mathbf{g}}$ and its realizations are instances of $R_{\mathbf{g},M}$. Following the lines of previous work [14], one can thus define the probability assigned to the disposition[15] $\mathbf{d}^{\text{Prev}}_{\mathbf{g},M}$, which is the probability of drawing randomly someone who has *M* in **g**. This probability is equal to the proportion of individuals who have *M* in **g**, that is, to Prev(**g**,*M*): if there are e.g., 10 % diseased people in **g**, then the probability of drawing randomly a diseased person in **g** is 10 %. Thus, the prevalence value can be identified to the objective probability assigned to the disposition $\mathbf{d}^{\text{Prev}}_{\mathbf{g},M}$. We name this strategy the "randomization" of the proportion of persons having *M* in **g**.

The randomization strategy may not be necessary to formalize a proportion in an actual group, such as the prevalence. But this strategy can also be applied to proportions of people in groups which are subject to a *possible, non-actual* condition – and thus, be relevant to formalize sensitivity and other IPs, and their estimates. As a matter of fact, the real sensitivity value $f_2(\mathbf{g},IT,M)$ was defined as the proportion of people who would get a positive result to *IT* among *M*'s bearers in **g**. This value can be "randomized" as follows. We can define $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M}$ as the disposition[16] to draw randomly, among the individuals of ***g*** who have *M*, someone who is tested positive by *IT*. More specifically, let's define the process $T^{\text{Se}}_{\mathbf{g},IT,M}$ as the "performance of test *IT* on the individuals in **g**, and random draw of an individual among those who have the disease *M*";[17] and the process $R^{\text{Se}}_{\mathbf{g},IT,M}$ as the "drawing by $T^{\text{Se}}_{\mathbf{g},IT,M}$ of someone who got a positive result to *IT*". The triggers of $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M}$ are instances of $T^{\text{Se}}_{\mathbf{g},IT,M}$, and its realizations are instances of $R^{\text{Se}}_{\mathbf{g},IT,M}$. As it happens, the real sensitivity value $f_2(\mathbf{g},IT,M)$ is the objective probability assigned to this disposition $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M,}$: indeed, if there are e.g., 15 % of the diseased people in **g** who would get a positive result by *IT*, then the probability of randomly drawing someone who got a positive test result by *IT* among diseased people in **g** if test *IT* would be performed on them is equal to 15 %.

Specificity value can be defined along similar lines, as probabilities assigned to actual dispositions borne by the group g noted $\mathbf{d}^{\text{Sp}}_{\mathbf{g},IT,M}$ (and similarly for the PPV and NPV). Although both $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M}$ and $\mathbf{d}^{\text{Sp}}_{\mathbf{g},IT,M}$ are dispositions inhering in **g**, they have different triggers and different realizations; the process $T^{\text{Sp}}_{\mathbf{g},IT,M}$ is the "performance of test *IT* on the individuals in **g**, and random draw of an individual among those who *do not* have the disease *M*" and the process $R^{\text{Sp}}_{\mathbf{g},IT,M}$ is the "drawing by $T^{\text{Sp}}_{\mathbf{g},IT,M}$ of someone who got a *negative* result to *IT*".

Assignment of real sensitivity values to dispositions

Let us now consider how to formalize these probability values in ontologies. $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M}$ is a disposition individual inhering in the group **g**; and a probability value can be assigned to this disposition using a datatype property **has_probability_value** [15]. This probability value is what we called the real sensitivity value:[18]

$\mathrm{d}^{\text{Se}}_{\mathrm{g},IT,M}$ has_probability_value $f_2(g,IT,M)$

Thanks to our analysis above, we can now answer our original question, and state what sensitivity estimates such as $\mathbf{estimate}^{\mathbf{Se}}_{\mathbf{1}}$ or $\mathbf{estimate}^{\mathbf{Se}}_{\mathbf{2}}$ are about[19] - namely, about this disposition:

$\text{estimate}^{\text{Se}}_{1}$ is_about $\mathrm{d}^{\text{Se}}_{\mathrm{g}1,IT1,M}$
$\text{estimate}^{\text{Se}}_{2}$ is_about $\mathrm{d}^{\text{Se}}_{\mathrm{g}2,IT2,M}$

Also, if the samples $\mathbf{s_1}$ and $\mathbf{s_2}$ are considered by the statistician as representative enough of a general population $\mathbf{g_0}$ encompassing $\mathbf{g_1}$ and $\mathbf{g_2}$, if RT_1 and RT_2 are considered as similar enough to be representative in the same way of the disease M, and if IT_1 and IT_2 are considered as similar enough to be representative of a more general index test IT_0, then:

$\text{estimate}^{\text{Se}}_{1,2}$ is_about $\text{d}^{\text{Se}}_{\text{g0},IT0,M}$

As $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M}$ is an individual, it cannot be related directly to the classes IT and M, but only indirectly, through the following formalization. First, $\mathbf{d}^{\text{Se}}_{\mathbf{g},IT,M}$ can be seen as an instance of a disposition class written $D^{Se}_{IT,M}$, which has as trigger the process class $T^{\text{Se}}_{IT,M}$: "performance of test IT on the members of a group, and random draw of a person among those who have the disease M"; and as realization the process class $R^{\text{Se}}_{IT,M}$ defined as "drawing by $T^{\text{Se}}_{IT,M}$ of someone who got a positive result to IT". We can then introduce two new relations *sensitivity_disposition_of_test* and *sensitivity_disposition_for* (abreviated as *se_of_test* and *se_for_disease*) relating $D^{\text{Se}}_{IT,M}$ with IT and M:

$\text{d}^{\text{Se}}_{\text{g},IT,M}$ instance_of $D^{\text{Se}}_{IT,M}$
$D^{\text{Se}}_{IT,M}$ *is_a Disposition*
$D^{\text{Se}}_{IT,M}$ *se_of_test IT*
$D^{\text{Se}}_{IT,M}$ *se_for_disease M*

These two relations *se_of_test* and *se_for_disease* are introduced for pragmatic reasons of facility of use: on a foundational level, $D^{Se}_{IT,M}$ and M (resp. IT) could be related through a complex array of relations and entities that involve the relation *has_trigger* between $D^{Se}_{IT,M}$ and $T^{Se}_{IT,M}$, as well as a sequence of relations between $T^{\text{Se}}_{IT,M}$ and M (resp. IT). Such an analysis would raise interesting theoretical questions, as instances of $D^{Se}_{IT,M}$ can exist even if no instance of M or IT do exist - we therefore face here issues similar to the ones addressed by [20] and [21].

Figure 2 represents classes and particulars involved in formalizing tests execution and results, sensitivity estimates, the disposition this estimate is about, and the real sensitivity value. Figure 3 represents the classes and particulars involved in formalizing aggregation of sensitivity estimates into a finer estimate. Specificity, PPV and NPV can be formalized along similar lines, as data items about dispositions related to tests and diseases through relations that could be labeled *sp_of_test*, *sp_for_disease*, *ppv_of_test, ppv_for_-disease*, *npv_of_test,* and *npv_for_disease.*

Example of application

An example will now illustrate this formalization. McTaggart and colleagues [8] have performed a meta-analysis to determine the accuracy of point-of-care tests for detecting albuminuria (let's call IT_0 the class of such index tests), using as reference test a laboratory test albumin-creatinine ratio-ACR (let's call RT_0 the class of such reference tests).

They take into account ten studies in their article. Consider for example Lloyd et al. [22], which measures the accuracy of semiquantitative Clinitek® microalbumin urine dipstick with a cutoff value indicating albumineria at 3.4 mg/mmol (let's call IT_1 the class of such index tests), with a laboratory ACR test with the same cutoff value as a reference (let's call RT_1 the class of such reference tests). A sample $\mathbf{s_1}$ of 204 diabetic patients (labelled here $\mathbf{p_{1,1}}$, $\mathbf{p_{1,2}}$,…, $\mathbf{p_{1,204}}$) was considered. On each of those patients, one measurement of IT_1 called $\mathbf{a_{1,i,1}}$ and one of RT_1 called $\mathbf{rt_{1,i,1}}$ is performed. The 2x204 = 408 processual entities are all part of a general tests execution process labelled $\mathbf{tests\_execution_{s1},}_{IT1,RT1}$, which leads after computation to the informational entity $\mathbf{estimate^{Se}_1}$, giving the proportion of measure pairs in which IT_1 led to a positive result among those in which RT_1 led to a positive result. This proportion is 83.8 %, and therefore, the value $f4(\mathbf{s_1},IT_1,RT_1)$ of the informational entity $\mathbf{estimate^{Se}_1}$ is 0.838.

Writing $\mathbf{g}$ the human population, we have $\mathbf{s_1}$ **part_of** $\mathbf{g}$; also, RT_1 *is_a* RT_0 and IT_1 *is_a* IT_0. Therefore, $f_4(\mathbf{s_1},IT_1,RT_1)$ provides an estimate of $f_2(\mathbf{g},IT_0,RT_0)$, which is the sensitivity value of a point-of-care test in detecting albuminuria in the general population. However, other studies are pooled with this one by McTaggart and colleagues [8] to provide a better estimate of $f_2(\mathbf{g},IT_0,RT_0)$. All together, they lead to the value $h(\mathbf{s_1},IT_1,RT1,\ldots,\mathbf{s_{10}},IT_{10},RT_{10})$ which provides an estimate of the value of $f_2(\mathbf{g},IT_0,RT_0)$.

Note that the ten studies taken into account in this meta-analysis include different kinds of patients. Seven studies involve each a different sample of patients (let's call them $\mathbf{s_1}$, $\mathbf{s_2}$, …., $\mathbf{s_7}$) with diabetes mellitus, one of them ($\mathbf{s_7}$) involving young patients with type 1 diabetes. Two studies consider samples of patients ($\mathbf{s_8}$ and $\mathbf{s_9}$) with kidney disease, diabetes mellitus, or both. Finally, one study includes a sample ($\mathbf{s_{10}}$) of patients treated for advanced chronic kidney disease in a renal outpatient clinic. Let's call $\mathbf{g}$ the human population, $\mathbf{g_1}$ the members of g who have diabetes mellitus, $\mathbf{g_2}$ the members of $\mathbf{g}$ who have a kidney disease and $\mathbf{g_0}$ the members of $\mathbf{g}$ who have either diabetes mellitus or a kidney disease (that is, $\mathbf{g_0}$ is the mereological sum of $\mathbf{g_1}$ and $\mathbf{g_2}$). All $\mathbf{s_i}$ are part of $\mathbf{g}$, the human population. Thus, the meta-analysis made by McTaggart and colleagues [8] provides an estimation of $f_2(\mathbf{g},IT_0,RT_0)$ or $f_2(\mathbf{g_0},IT_0,RT_0)$. If the meta-analysis had been performed on $\mathbf{s_1}$-$\mathbf{s_7}$ only, then it would have provided an estimation of $f_2(\mathbf{g_1},IT_0,RT_0)$; and if it had been performed on samples of patients with kidney disease only, then it would have provided an estimation of $f_2(\mathbf{g_2},IT_0,RT_0)$.

Note also that various cutoff values can be used to define the presence of albuminuria, varying between 2.65 mg/mmol to 3.4 mg/mmol, and those values are chosen by the medical sub-community who is conducting the study (the same cutoff value is taken for both IT_0 and RT_0 in each study). Therefore, the classes IT_0 and RT_0, which mention 'detecting albuminuria' without specifying a cutoff value, are not scientifically defined: those classes are not universals, but rather collection of particulars [19] whose nature is partly social ([8] acknowledge this limitation in their meta-analysis).

Alternative meta-analysis could use a subset of those studies to estimate various sensitivities, for example the sensitivity $f_2(\mathbf{g_1},IT_1,RT_1)$ of point-of-care test with a reference of laboratory ACR test, with albuminuria defined as ACR greater than 3.4 mg/mmol, in the reference class of patients with diabetes mellitus; or the sensitivity $f_2(\mathbf{g_2},IT_2,RT_2)$ of point-of-care test, with a reference of laboratory ACR test, with albuminuria defined as ACR greater than 2.65 mg/mmol, in the reference class of patients with kidney disease; etc. A well-founded semantic representation of sensitivity should thus make clear what is the reference class, as well as the class of index test and reference test.

Discussion and conclusions

We have thus provided a practically tractable formalization of IPs in a realist ontology, which clearly dissociates IPs' real values, their estimates and the related proportion measurements. It has defined the central entities that are concerned by an IP estimation in a way that is compliant with OBO Foundry. In particular, it addresses the difficulty of considering possible, non-actual conditions in a realist ontology based on BFO by introducing dispositions.

This model could then be extended in three directions. A first step would be to clarify the ontological status of the two following entities: sample sizes on one hand; and 95 % confidence interval for sensitivity and specificity values on the other hand. A second step would be to clarify the relations *se_of_test* and *se_for_disease*, which could be reduced to basic relations and entities already accepted in the OBO Foundry. A third step would be to use this model in an ontology-based diagnostic system that would compute positive predictive values or negative predictive values from the prevalence, sensitivity and specificity values. More generally, it could be articulated with medical Bayesian networks. As a matter of fact, the notion of medical test used here could be generalized to a very general notion of test consisting in inferring the presence of an entity on the basis of the knowledge of the presence of another entity; as such, it could serve as a foundation for the integration of Bayesian reasoning into ontologies.

This model could be used in two kinds of computer applications targeted at two different kinds of audiences. First, clinicians could determine more easily which kind of sensitivity and specificity (or PPV and NPV) estimates they could use when diagnosing a disease for a given patient, by having a clearer view of the subjects' characteristics in each samples on which those IP estimates are based. As a matter of fact, section 3.4 illustrates how an ontological analysis can make explicit what are the index test, the reference test and the sample associated with a sensitivity estimation. Universal qualities that are instantiated by all members of the sample - such as having diabetes mellitus, being a man, being more than 65 years old, etc. - would enable to determine what could be the reference class $\mathbf{g}$ associated with a sensitivity estimate. This enables to determine, when applying some given IP values to a specific patient with given characteristics, whether this application is warranted or not.

Second, statisticians could determine more easily which kind of sensitivity estimates they could aggregate together. If several estimations of IPs are represented ontologically according to the structure shown above, one could use this ontological structure to determine which estimations of IPs could be combined to obtain a finer estimate. First, one would have to find a group $\mathbf{g_0}$ that would encompass the reference classes (such as $\mathbf{g_1}$ and $\mathbf{g_2}$) associated with those studies. Second, one would have to analyze whether there exists some general index test class such as IT_0 (resp. some general reference test class such as RT_0) which would subsume the various index tests classes such as IT_1 and IT_2 (resp. reference tests such as RT_1 and RT_2) that are used in those studies. Once those are found, one could use meta-analytic methods to derive a value for $f_2(\mathbf{g_0},IT_0,RT_0)$ from the other studies. Future work will aim at building an ontology of medical tests to facilitate finding such encompassing index and reference test classes.

As it takes into account the dependence of IPs upon the group of people considered, it has the potential to contribute to the development of precision medicine [23] in context of learning health systems [24, 25], an emerging approach that takes into consideration patients characteristics and dispositions, including individual variability in genes, to offer more personalized preventive, diagnostic and therapeutic strategies.

Endnotes

[1]These will be abbreviated in the following as "a test *IT*" and "the patient has *M*". Note that a test may aim at diagnosing a disease, in which case it can be called "indicator of diagnostic performance". However, it may also aim at evaluating the presence of a disorder, a pathological process [26], a predisposition to a disease, a sign, a symptom, or other various medically relevant entities

(such as a glycemia higher than 1.26 g/l). Several tests results can then be considered to draw a diagnostic conclusion for a disease. Therefore, in the general case, indicators of performance are indicators of *assay* performance rather than indicators of *diagnostic* performance (we thank an anonymous reviewer for this suggestion of terminology). Also, a test does not need to be performed on a human – it can be performed on a non-human animal. In the following, we will consider tests aiming at diagnosing a disease on a human, but our considerations can be straightforwardly adapted to tests aiming at evaluating another medically relevant entity on a human or non-human animal.

[2]In practice, such a test is not perfect; thus, it could be analyzed as a chain of two tests: one that detects the rheumatoid factor on the basis of e.g., some chemical reaction, and another one that detects rheumatoid arthritis on the basis of the presence of the rheumatoid factor.

[3]More specifically, it should be interpreted as the expected value of such a proportion – but we will ignore here this additional subtlety.

[4]The article will concentrate on the case of sensitivity, but it can be similarly adapted to other IPs.

[5]Here again (see footnote 3), this should be interpreted as the expected value of such a proportion.

[6]At least for all practical purposes: from a theoretical point of view, every measurement can be wrong, even pure observations.

[7]If one assumes that the sample is representative of the target population, there should be no selection bias (which occurs when proper randomization is not achieved). However, the sensitivity values that would be obtained using two different samples could be slightly different since randomness at the selection process will yield slightly different samples. That is why statisticians use confidence interval for characterizing sensitivity and specificity.

[8]We might also speak of a "sensitivity in a sample" for the function $f_2(\mathbf{s},IT,M)$, that is, the proportion of people who are tested positive by *IT* among the diseased person in the sample **s**. But it might be confusing to speak of both the "sensitivity in a target population" and the "sensitivity in a sample"; and the first and the second arguments above may justify keeping the label "sensitivity" for this proportion in a target population **g** – that is, for $f_2(\mathbf{g},IT,M)$.

[9]Let us summarize. On one hand, $f_2(\mathbf{g},IT,M)$ is the value of a non-actual proportion (because the test *IT* is not performed on all members of **g**), which cannot be known with certainty, but only estimated. On the other hand, both $f_4(\mathbf{s},IT,RT)$ and $f_2(\mathbf{s},IT,M)$ (see footnote 8) are values of actual proportions (because the tests *IT* and *RT* are performed on all members of **s**); and although $f_2(\mathbf{s},IT,M)$ cannot be known with certainty (because we cannot know with certainty who has the disease: we can only use a reference test – at best the gold standard – to determine who are those individuals), $f_4(\mathbf{s},IT,RT)$ can be known with certainty for all practical purposes (because we can know with certainty who got a positive result to *RT*).

[10]We have created an ontology according the lines of what is described below, built on OBI, called BIPO (Bayesian Indicator of Performance Ontology). It can be found at https://github.com/OpenLHS/BIPO. It contains 24 classes, 12 object properties, 2 data properties and 42 logical axioms.

[11]We will not take a stance on whether *Medical_test* should be interpreted as identical to OBI:*Assay*, as proposed by [27].

[12]Note that in some cases, several pairs of tests will be performed on a person. See e.g., Kimberger et al. (2007), which measures the accuracy of a temporal artery thermometer in detecting fever (defined as a temperature greater than 37.8 °C), with respect to a reference standard given by a bladder thermometer: four measurement pairs of temporal artery temperature and bladder temperature are performed on each of the seventy patients of the sample considered by the authors. To represent such a case, one can introduce for every human $\mathbf{p_i}$ a sequence of four reference tests $\mathbf{rt}_{1,i,1}$, $\mathbf{rt}_{1,i,2}$, $\mathbf{rt}_{1,i,3}$ and $\mathbf{rt}_{1,i,4}$ .and four index tests $\mathbf{it}_{1,i,1}$, $\mathbf{it}_{1,i,2}$, $\mathbf{it}_{1,i,3}$ and $\mathbf{it}_{1,i,4}$; but the formalization that is described below remains similar.

[13]See e.g., http://vassarstats.net/clin1.html for an example of webpage supporting this kind of computation.

[14]As a reminder, not only the values of PPV and NPV but also the values of sensitivity and specificity depend on the group under consideration (this is the spectrum effect), and it is not the task of the ontologist to determine which ones should be idealized as constant (for all practical matters) across groups and which ones should be considered as variable: the task of the ontologist is to represent those values and the entities those values depend upon.

[15][15] assigned a probability to a triplet $(\mathbf{d},T,R)$ rather than to a disposition **d**, because it had to take into account dispositions that may have several classes of triggers or realizations (that is, multi-trigger and multi-track dispositions [20]). However, in the present situation, $\mathbf{d}^{Se}_{\mathbf{g},M}$ is simple-trigger and simple-track: all its triggers are instances of $T^{Se}_{\mathbf{g}}$, and all its realizations are instances of $R^{Se}_{\mathbf{g},M}$. Therefore, the probability value assigned to $(d^{Se}_{\mathbf{g},M},T^{Se}_{\mathbf{g}},R^{Se}_{\mathbf{g},M})$ can be, for practical matters, assigned directly to $d^{Se}_{\mathbf{g},M}$.

[16]Such dispositions should not be confused with other dispositions in the medical domain. First, diseases have been formalized as dispositions by the Ontology for General Medical Sciences (OGMS) [26]. Second, there can be predispositions to diseases that could be

formalized as disposition. However, the disposition to draw randomly, among the individuals of ***g*** who have *M*, someone who is tested positive by *IT*, exists independently of whether the disease (or a predisposition to this disease) is formalized or not as a disposition. Note also that this disposition inheres in a group of people, whereas a disease as a disposition (as formalized by OGMS), or a predisposition to a disease, inheres in a single person.

[17]In general, we cannot determine in practice with certainty which individuals of *g* have *M*, and which do not (see the discussion about gold standard tests above); but the practical impossibility to realize this trigger does not preclude to define this entity.

[18]We could also introduce the entity **real_sensitivity**$_{\mathbf{g},IT,M}$ instance of *Data_item*, as a sibling of **estimate**$_1^{Se}$ such that **real_sensitivity**$_{\mathbf{g},IT,M}$ **has_specified_value** $f_2(\mathbf{g},IT,M)$ (cf. [14], in which **real_sensitivity**$_{\mathbf{g},IT,M}$ was denoted $\mathbf{se}_{\mathbf{g},IT,M}$). However, the value $f_2(\mathbf{g},IT,M)$ assigned to such an entity will never be known with certainty. We could substitute to this value the best estimate of the sensitivity value, as was proposed in [14]; however, such a model could not represent in a single ontology various estimates of the same sensitivity – whereas it is possible in the present framework, which also makes unnecessary the introduction of the informational entity **real_sensitivity**$_{\mathbf{g},IT,M}$.

[19]It is important to differentiate what a sensitivity estimate is about (namely a disposition) from how it has been mathematically obtained (for example, by weighting different proportion measurements) – as explained earlier, the latter will not be represented in the ontology, as various mathematical methods can be used.

Abbreviations

General abbreviations for indicators of performance
IP: (Bayesian) Indicators of performance; NPV: Negative predictive value; PPV: Positive predictive value; Prev: Prevalence; Se: Sensitivity; Sp: Specificity
Other general abbreviations
ACR: Albumin-creatinine ratio; RF: Rheumatoid factor
Classes and instances abbreviations for disposition-related entities
$\mathbf{d}_{\mathbf{g},M}^{Prev.}$: Disposition (borne by the group **g**) that a person randomly drawn among the individuals in **g** would have *M*; $\mathbf{d}_{\mathbf{g},IT,M}^{Se}$: Disposition (borne by the group **g**) that a person randomly drawn among the individuals of **g** who have *M* would have a positive result to *IT*; this is an instance of $D_{IT,M}^{Se}$; $D_{IT,M}^{Se}$: A subclass of *Sensitivity disposition* such that $D_{IT,M}^{Se}$ *se_for_disease M* and $D_{IT,M}^{Se}$ *se_of_test IT*; $D_{IT,M}^{Sp}$: A subclass of *Specificity disposition* such that $D_{IT,M}^{Sp}$ *sp_for_disease M* and $D_{IT,M}^{Sp}$ *sp_of_test IT*; $T_{\mathbf{g}}$: The process of drawing randomly a person in **g**; the triggers of $\mathbf{d}_{\mathbf{g},M}^{Prev}$ are instances of T_g; $T_{\mathbf{g},IT,M}^{Se}$: The process of performing test *IT* on the individuals in **g**, and then drawing randomly an individual among those who have the disease *M*; the triggers of $\mathbf{d}_{\mathbf{g},IT,M}^{Se}$ are instances of $T_{\mathbf{g},IT,M}^{Se}$; $R_{\mathbf{g},M}$: The process of drawing by $T_{\mathbf{g}}$ someone who has *M*; the realizations of $\mathbf{d}_{\mathbf{g},M}^{Prev}$ are instances of $R_{\mathbf{g},M}$; $R_{\mathbf{g},IT,M}^{Se}$: The process of drawing by $T_{\mathbf{g},IT,M}^{Se}$ someone who got a positive result to *IT*; the realizations of $\mathbf{d}_{\mathbf{g},IT,M}^{Se}$ are instances of $R_{\mathbf{g},IT,M}^{Se}$;
Other classes abbreviations
IT / IT_0 / IT_1 / IT_2: A subclass of *Medical test* which is an index test (test whose indicator of performance is being estimated); *M*: A subclass of *Disease*; *RT* / RT_0 / RT_1 / RT_2: A subclass of *Medical test* which is a reference test
Other instances abbreviations
g / $\mathbf{g_0}$ / $\mathbf{g_1}$ / $\mathbf{g_2}$: An instance of *Collection of humans* which is a general human population; $\mathbf{p_i}$ / $\mathbf{q_j}$: An instance of *Human*; $\mathbf{it_{1,i}}$ (resp. $\mathbf{it_{2,j}}$): An instance of (index) *Medical test* performed on person $\mathbf{p_i}$ (resp. $\mathbf{q_j}$); $\mathbf{rt_{1,i}}$ (resp. $\mathbf{rt_{2,j}}$): An instance of (reference) *Medical test* performed on person $\mathbf{p_i}$ (resp. $\mathbf{q_j}$); **s** / $\mathbf{s_1}$ / $\mathbf{s_2}$: An instance of *Sample of humans*;
Functions abbreviations
$f_1(IT,M)$: Proportion of individuals who get a positive result to *IT*, among individuals who have *M*; $f_2(\mathbf{g},IT,M)$: Proportion, among members of **g** who have *M*, of those who would get a positive result to *IT if the test IT was realized on them*; this is the real sensitivity value of *IT* for *M*; $f_3(\mathbf{g},IT,RT)$: Proportion, among members of **g** who would get a positive result to *RT if the test RT was realized on them*, of those who would get a positive result to *IT if the test IT was realized on them*; $f_4(\mathbf{s},IT,RT)$: Proportion, among members of sample **s** who had a positive result to *RT*, of those who got a positive result to *IT;* this is an estimate of the real sensitivity value of *IT* for *M*, performed on a sample **s**, with *RT* as a reference test; $f'_2(\mathbf{g},IT,M)$: Proportion, among members of **g** who don't have *M*, of those who would get a negative result to *IT if the test IT was realized on them*; this is the real specificity value of *IT* for *M*; $f'_4(\mathbf{s},IT,RT)$: Proportion, among members of sample **s** who had a negative result to *RT*, of those who got a negative result to *IT;* this is an estimate of the real specificity value of *IT* for *M*, performed on a sample **s**, with *RT* as a reference test; $h(\mathbf{s_1},IT_1,RT_1,\mathbf{s_2},IT_2,RT_2)$: Estimate of the sensitivity value obtained by aggregating the estimate on sample $\mathbf{s_1}$ and the estimate on sample $\mathbf{s_2}$

Acknowledgements
We would like to thank two anonymous reviewers for their comments that led to significant improvements in our model and in the manuscript, as well as assistance during various presentations of this work for their suggestions. ABa would like to thank the bourse de fellowship of the department of medicine of Sherbrooke University for financial support. This manuscript is an extended version of work presented at ICBO (International Conference on Biomedical Ontology) 2015.

Authors' contributions
ABa conceived the formalization of the real indicator of performance values, JFE and ABa conceived the formalization of the estimation of indicators of performances, and ABa and JFE developped the BIPO ontology, in light of inputs from ABu and RD. JFE and RD provided the medical examples supporting the formalization.
ABa drafted the manuscript with important feedbacks from JFE, RD and ABu. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Département de médecine, Université de Sherbrooke, Sherbrooke, Québec, Canada. [2]INSERM UMR 1099, LSTI, Rennes, France. [3]CHU de Martinique, Université Antilles-Guyane, Fort-de-France, France. [4]INSERM UMR_S 1138 Eq 22, Université Paris Descartes, Hôpital européen Georges Pompidou, AP-HP, Paris, France. [5]Centre de recherche du CHUS, CIUSSS de l'Estrie-CHUS, Sherbrooke, Québec, Canada.

Received: 2 February 2016 Accepted: 6 September 2016
Published online: 03 January 2017

References
1. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science. 1959;130:9–21.
2. Peacock J, Peacock P. Oxford Handbook of Medical Statistics. Oxford: Oxford University Press; 2011.
3. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16:981–91.
4. Park HB, Yokota A, Gill HS, El Rassi G, McFarland EG. Diagnostic accuracy of clinical tests for the different degrees of subacromial impingement syndrome. J Bone Joint Surg Am. 2005;87:1446–55.
5. Hanson V, Rexler ED, Kornreich H. The relationship of rheumatoid factor to age of onset in Juvenile rheumatoid arthritis. Arthritis Rheum. 1969;12:82–6.
6. Moons KGM, van Es G-A, Deckers JW, Habbema JDF, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8:12–7.

7. Niven DJ, Gaudet JE, Laupland KB, Mrklas KJ, Roberts DJ, Stelfox HT. Accuracy of peripheral thermometers for estimating temperature: a systematic review and meta-analysis. Ann Intern Med. 2015;163:768–77.
8. McTaggart MP, Newall RG, Hirst JA, Bankhead CR, Lamb EJ, Roberts NW, Price CP. Diagnostic accuracy of point-of-care tests for detecting albuminuria: a systematic review and meta-analysis. Ann Intern Med. 2014;160(8):550–7.
9. Reilly P, Macleod I, Macfarlane R, Windley J, Emery R. Dead men and radiologists don't lie: a review of cadaveric and radiological studies of rotator cuff tear prevalence. Ann R Coll Surg Engl. 2006;88:116–21.
10. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.
11. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. In: Pisanelli D, editor. Ontologies in medicine. Amsterdam: Ios Press; 2004. p. 20–38.
12. Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. Stud Health Technol Inform. 2012;180:68–72.
13. da Costa PCG, Laskey KB, Laskey KJ. PR-OWL: A Bayesian Ontology Language for the Semantic Web. In: Costa PCG, d'Amato C, Fanizzi N, Laskey KB, Laskey KJ, Nickles M, Pool M, editors. Uncertainty Reasoning for the Semantic Web I. Berlin: Springer; 2008. p. 88–107.
14. Soldatova LN, Rzhetsky A, De Grave K, King RD. Representation of probabilistic scientific knowledge. J Biomed Semant. 2013;4 Suppl 1:S7.
15. Barton A, Duvauferrier R, Burgun A. Formalization of indicators of di agnostic performance in a realist ontology. In: Couto F M, Hastings J, editors. Proceedings of 6th International Conference on Biomedical Ontology (ICBO2015). CEUR Workshop Proceedings 1515, CEUR-WS.org; 2015. p. 63-70.
16. Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, He Y. OBCS: The Ontology of Biological and Clinical Statistics. In: Hogan W, Arabandi S, Brochhausen M, editors. Proceedings of the 5th International Conference on Biomedical Ontology. Houston: CEUR Workshop Proceedings; 2014. p. 65.
17. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone S-A, Soldatova LN, Stoeckert Jr CJ, Turner JA, Zheng J, OBI consortium. Modeling biomedical experimental processes with OBI. J Biomed Semant. 2010;1 Suppl 1:S7.
18. Jansen L, Schulz S. Grains, components and mixtures in biomedical ontologies. J Biomed Semant. 2011;2:S2.
19. Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Appl Ontol. 2010;5:139–88.
20. Röhl J, Jansen L. Representing dispositions. J Biomed Semant. 2011;2 Suppl 4:S4.
21. Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector A. An Ontological Analysis of Reference in Health Record Statements. In: Garbacz P, Kutz O, editors. Proceedings of the 8th International Conference on Formal Ontology in Information Systems (FOIS2014). Amsterdam: Ios Press; 2014. p. 289–302.
22. Lloyd, Mariana M, Johannes K, H. Van Jaarsveld. Evaluation of point-of-care tests for detecting microalbuminuria in diabetic patients. South African Family Practice 53.3. 2011;281–286.
23. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. N Engl J Med. 2012;366:489–91.
24. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, Ethier J-F, Kostopoulou O, Kuchinke W, McGilchrist M, Van Royen P, Wagner P. Translational medicine and patient safety in Europe: TRANSFoRm—architecture for the Learning Health System in Europe. BioMed Res Int. 2015;2015:1–8.
25. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, Gunter C, Musen M, Platt R, Stead W, Sullivan K, Van Houweling D. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc. 2014;22(1):43-50.
26. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. San Francisco: Proceedings of the 2009 AMIA Summit on Translational Bioinformatics; 2009. p. 116–20.
27. Jensen M, Cox AP, Bona JP, Duncan W, Ray PL, Diehl AD. Applications of OBI "assay.". In: Hogan W, Arabandi S, Brochhausen M, editors. Proceedings of the 5th International Conference on Biomedical Ontology. Houston: CEUR Workshop Proceedings; 2014. p. 96–7.