A Realist Representation of Social Identity Data

Amanda Hicks, Ph.D.
Department of Health Outcomes and Policy
University of Florida
Gainesville, USA
aehicks@ufl.edu

Abstract—**Social identities merit special treatment in realist ontologies. Their ontological status is unsettled, so we should model them in a manner that is agnostic with respect to their ontological status. Nevertheless, there is a clear criterion for determining whether a specific person has a particular identity, namely, whether that person asserts that they do. This social act forms the basis for a realist representation, not of social identities themselves, but of data about social identities. We report the representation of social identities in the Ontology of Medically Related Social Entities and show that it supports data integration and retrieval.**

Keywords—data integration; demographic information; ethnicity; gender identity; identity; Ontology of Medically Related Social Entities; race

I. Introduction

Demographic information is widely used in information systems. In medical and health information systems they support a variety of biomedical and informatics tasks such as cohort discovery, statistical comparison of groups of people, and record linkage [2]. Common demographic data collected in medical settings include birth date, preferred language, race, ethnicity and sex or gender. In 2011 the Institute of Medicine recommended collecting information on sexual orientation and gender identity (as distinct from biological sex) in electronic health records [3], and Stage 3 for Meaningful Use requires that electronic health records (EHR) certified for meaningful use have fields for collecting information on sexual identity by 2018 [4-6]. It is, therefore, increasingly important to semantically represent gender identity and other social identities coherently to support data retrieval and integration. [2] discusses previous work on realist representations of demographic information in general in the Ontology of Medically Related Social Entities (OMRSE).

This paper describes social identities as a special subset of demographic information and describes a realist representation of social identities to support data retrieval and data integration. This representation supports integration and retrieval of data about people according to their social identities. For the purpose of this paper, social identities include (but are not be limited to) race, ethnicity, and gender identity.

Section Two describes the background assumptions, and hypothesis of this paper. Section Three provides background on data collection for gender identity, sexual orientation, race and ethnicity, drawing important distinctions for understanding the semantics of terms used to describe these types of social identities. Section Four describes a framework for ontologically representing social identities in OMRSE to support semantic integration of demographic data. Section Five describes the results of the validation of our representation with competency questions. Section Six discusses results and future work.

II. Background Assumptions and Hypotheses

[2] notes that demographic data are about a heterogeneous group of things; they may include data about preferred language, biological sex, gender identity, race, date of birth, and marital status. The ontological status of some of these entities is clear. Biological sex is a quality of an organism [7]; date of birth is a time interval; and marital status is the result of a contractual act. However, the ontological status of race, ethnicity, and gender identity is controversial [8, 9]. For this reason, this paper does not attempt to answer the question, what kind of things are race, ethnicity, and gender identities? Instead, it places the process of asserting an identity at the center of a realist represention of social identity data in OMRSE.

We begin our work with the assumption that there is a difference between demographic data such as gender identity, race, ethnicity, on the one hand, and sex, birth date, and marital status on the other. Although the latter group is heterogeneous, its members do share something significant in common; statements about each can be verified as inter-subjective facts about the world. Although we often gather data about a person by asking questions such as *Are you male or female?*, *What is your birth date?*, and *Are you married?*, biological sex, birth date, and marital status refer to inter-subjective features of the world. If by ‘sex’ we mean karyotypic or phenotypic sex, we can perform genetic testing to determine a person’s karyotype or a physical examination to determine phenotype. While we cannot directly observe the date of a person’s birth, once the event is completed, a birth date is something that multiple people observe and come to consensus on. We can determine that a person is married by producing a marriage certificate; if there is no marriage certificate, there is no marriage. In this sense, reports of one’s own sex, birth date, and marital status are corrigible in the face of facts about the inter-subjective world. However, reports of one’s own gender identity, race, and ethnicity are not similarly corrigible. That is, if Jane says that she is a black, Latina, woman, she has already provided all the information we can hope to acquire to determine and verify her race, ethnicity, and gender identity. There is nothing in either the physical or social the world that we can consult to

This work was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 and by award CDRN-1501-26692 from the Patient Centered Outcomes Research Institute (PCORI). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH/NCATS or PCORI.

verify the truth of these claims unless it is to return to Jane herself and ask her to verify these statements.

Nevertheless, it seems that it is possible for Jane to provide misinformation about at least some aspects of her identity. For example, one might object that if Jane has white, non-Latino parents who insist that Jane herself is neither black nor Latina, that this constitutes intrasubjective evidence that her claims are false. This scenario underscores the importance of the context of data collection for determining the meaning of the data collected. As we will see in the next section, the race and ethnicity data collection practices and guidelines prevalent in U.S. healthcare system explicitly rule out defining race and ethnicity in terms of "blood" quotas or other inclusion criteria. Furthermore, the definitions that do exist for these terms are seldom presented to respondets. The result is that the data that are currently, routinely collected only tell us how the person actually identifies themselves. Notice how this affects the case where Jane's parents are white, non-Latino. In the absence of clear inclusion and exclusion criteria for "white" and "Latino", all we know is that Jane's parents identify themselves as white and non-Latino. This does not rule out Jane having reasons to identify some other way. Finally, we may be concerned that Jane has deliberately provided misinformation about her identity. There are two things to note about this scenario. First, no ontology can get around the problem of potential dishonesty or bad data collection practices, nor are they intended to. Second, even in the broader context of data management we do not regard this as a pressing issue since, we have no reason to suspect that providing deliberately misleading inforamtion about one's identity is a common enough pratice to effect the results of data quality and data analysis significantly.

Our hypothesis was that representing social identity data with respect to the process of identifying rather than in terms of identities themselves can support data integration and retrieval in a realist framework while avoiding controversial ontological commitments.

III. Data collection for Gender Identity, Race, and Ethnicity

For the purpose of this work we have adopted the definition and characterization of gender identity in [1]. For race and ethnicity we use the Office of Management and Budget (OMB) definitions and guidelines[10] since this standard is already widely used in biomedicine. Most medical terminologies, coding schemes, and surveys use terms that are intended to comply with the Office of Management and Budget (OMB) minimum set of categories for race and ethnicity [11, 12].

A. Gender identity

Table 1 contains definitions of terms related to sex and gender as presented in [1]. These definitions have been influential in shaping the discussion of the collection of data about gender identity [11] and conform to standard usage where the distinctions between (a) sex and gender and (b) gender expression and gender identity are observed.

By examining these definitions we can see that the verification criteria for gender identity is the individual's own subjective report of their identity rather than an objective or inter-subjective criterion.

Gender identity does not refer to biological and physiological characteristics since it is distinct from biological sex. Furthermore, gender identity cannot be ascertained or verified by gender expression. Consider two cases. 1) Some trans individuals have not socially transitioned to their perceived identity. A biological male who lives as a man but has a subjective sense of being a woman may have a masculine gender expression that would not be indicative of their feminine gender identity. 2) Some people adopt the cultural norms associated with a particular gender expression, but identify differently. For example, a non-binary person may have a masculine gender expression without identifying as a man.

B. Race and Ethnicity

The Office of Management and Budget (OMB) has defined a minimal set of categories for collecting data on race and ethnicity in the U.S. Census. These categories are also used in health care settings and health research in the U.S. [11, 12]. It is important to note that, while the OMB defines the minimum race and ethnicity categories partially in terms of genealogy, they explicitly do not regard the categories as naturalistic, anthropological, or scientific, but instead as social-constructs. Furthermore, they encourage self-identification in the data collection process wherever possible [11].

TABLE I. Definitions from the IOM 2011 Report on the Health of LGBT People

TERM	DEFINITION
Sex	a biological construct, referring to the genetic, hormonal, anatomical, and physiological characteristics on whose basis one is labeled at birth as either male or female
Gender	the cultural meanings of patterns of behavior, experience, and personality that are labeled masculine or feminine
Gender Expression	the manifestation of characteristics in one's personality, appearance, and behavior that are culturally defined as masculine or feminine
Gender Identity	a person's subjective sense of his or her gender

The OMB definitions for race characterize racial categories on the basis of their descent from the original peoples of some geographic region (Table 2). This characterization poses problems for a realist representation. First, the criterion is ambiguous insofar as it does not define 'original peoples'. At what point in human history are original peoples determined? Second, the criterion is not applied consistently. 'American Indian or Alaska Native' is defined as a person who has origins in any of the original peoples of North and South America (including Central America), *and maintains tribal affiliation or community attachment* (emphasis added). This is the only race category that has the extra requirement of a social relationship, which renders the categories not exhaustive. For example, Mexican-Americans who have origins in the original peoples of South or Central America but do not maintain a tribal affiliation or community attachment do not fit any of OMB categories for race.

However, despite the genealogical criterion in the definitions of these terms, the OMB guidelines stress interpreting statements about race as socio-cultural characteristics that involve ancestry rather than as biological or genetic characteristics. This connection to ancestry suggests that the verification criterion for an OMB-based statement about racial identity is about a historical fact since ancestry is determined by inter-subjective criteria. However, this contrasts with additional guidelines for data collection that indicate that that the verification criteria are the subject's response to OMB questions about race.

- "Respect for individual dignity should guide the processes and methods for collecting data on race and ethnicity; ideally, respondent self-identification should be facilitated to the greatest extent possible, recognizing that in some data collection systems observer identification is more practical."
- "do **not** establish criteria or qualifications (such as blood quantum levels) that are to be used in determining a particular individual's racial or ethnic classification." (original emphasis)
- "do **not** tell an individual who he or she is, or specify how an individual should classify himself or herself." (original emphasis) [11].

TABLE II. DEFINITIONS FOR THE OFFICE OF MANAGEMENT AND BUDGET MINIMUM CATEGORIES FOR RACE

OMB CATEGORY	OMB DEFINITIONS
American Indian or Alaska Native	A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian	A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American	A person having origins in any of the black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American."
Native Hawaiian or Other Pacific Islander	A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White	A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

Similarly to race, the OMB's definition of ethnicity also invokes genealogy. The term 'Hispanic' refers to persons who trace their origin or descent to Mexico, Puerto Rico, Cuba, Central and South America, and other Spanish cultures.

However, the same caveats that were discussed for race apply to ethnicity, namely, 1) 'ethnicity' should not to be interpreted as referring to biological or genetic characteristics, but rather as referring to ancestry, and 2) the verification criterion for OMB-based statements about ethnicity is the subject's response to OMB-based questions about ethnicity.

Finally, we should not expect existing data on race and ethnicity to reflect a consistent, genealogical criterion since most patients are not presented with definitions of racial and ethnic terms during the intake process at a clinic or on a survey and because the language used to describe these categories may vary at the discretion and preference of the person(s) designing the form. For example, 'black', 'African American', and 'black or African American' can all be used to describe the same racial category.

In short, the ontological types of things that a race and ethnicity datum might be about are heterogeneous, and to make matters worse, there is often not a single type that is common to all of them that would provide either necessary or sufficient conditions. Furthermore, these categories are not historically stable and stem from contingent circumstances. Even if an ontologist were confident that there are universals for social identities, the historical contingency of identity categories makes ontologically representing these social identities as stable universals impractical. Nevertheless, ontologists can provide a realistic representation of how people actually identify when asked to do so. The lack of inter-subjective verification criteria for identity statements *in tandem* with the stress on self-identification in the instructions provides a principled basis for representing social identity data differently from data with an inter-subjective or objective verification criterion such as birth date and diagnosis.

IV. A REALIST REPRESENTATION OF IDENTIFICATION PROCESSES AND IDENTITY DATA

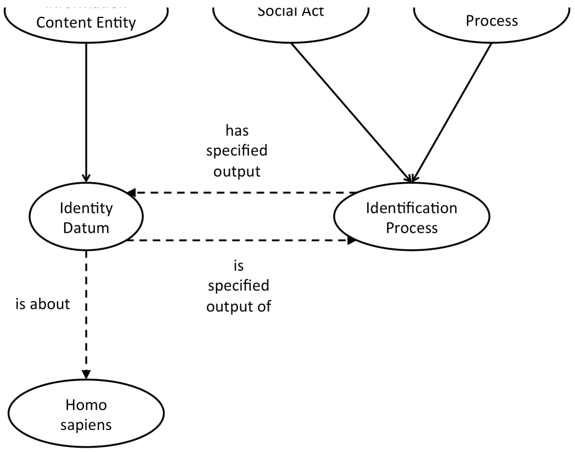

Fig. 1. Representation of Identification Data and Identification Processes

In light of the fact that it is not clear what kinds of things identities are, OMRSE does not model identities as such. However, we do know how identity data are collected and that their verification criterion involves the process of identifying. For this reason, we make the processes of asserting an identity central to representing social identity data, rather than identities themselves. An identification process is a planned process that might utilize a specific vocabulary or common data model, such as the OMB minimal categories for race and ethnicity. However, some identification processes might not use a common vocabulary or common data elements. For example, some may only utilize a free text field. Identification processes, as we represent them here, are planned process that record an identity statement about an individual person. They should not be confused with the private and internal mental or emotional process that involve or give rise to a subject sense of one's identity. Identification processes, as we describe them here, are planned, social, and result in identity data. OMRSE represents these data as information content entities that are the outputs of identification processes. Conversely, all identity data are the specified outputs of an identification processes. Fig. 1 illustrates the representation of identity data and identity processes in OMRSE.

Subclasses of identification process include racial identification process, ethnic identification process, and gender identification process. Identification processes that use a particular set of terms or coding scheme can be the basis of further descendent classes of identification process. For example, OMB racial identification process and PCORnet racial identification process are subclasses of racial identification process (Fig. 2). The latter represents racial identification used in the PCORnet Common Data Model (CDM), a data standard for representing clinical patient data from clinical sites across the US for use in the National Patient-Centered Clinical Research Network (PCORnet) [13].

Table 3 contains definitions related to representing OMB's categories related to OMB Asian as an example of how identities that employ a common data model or common vocabulary are represented with this approach.

A. Extended categories

The OMB guidelines for race and ethnicity allow data collectors to use a larger number of race and ethnicity categories as long as they are extensions of and mappable to the OMB minimum categories, i.e., as long as they do not introduce new categories but are equivalent or subcategories to those in the minimal set [10]. In cases where the expanded set includes subcategories of OMB classes, corresponding identity data can be introduced as a subclass of the appropriate OMB datum. For example, Fig. 3 shows CDC Spanish Basque datum as a subclass of OMB Hispanic or Latino datum.

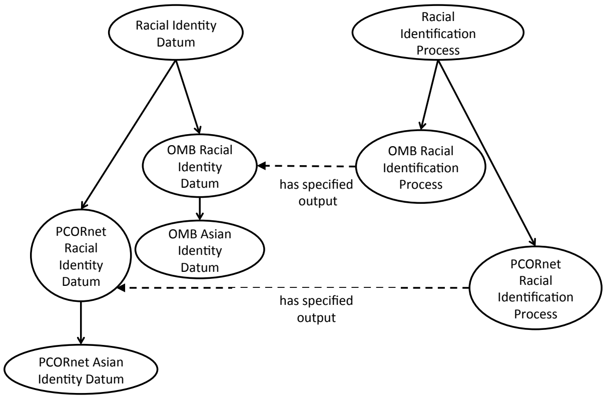
Racial Identity Datum
Racial Identification Process
OMB Racial Identity Datum
OMB Racial Identification Process
has specified output
OMB Asian Identity Datum
PCORnet Racial Identity Datum
PCORnet Racial Identification Process
has specified output
PCORnet Asian Identity Datum

Fig. 2. An example of how to represent heterogeneous social iden data using

B. Integrating Heterogeneous Data

Despite the similar categories and identical definitions, the PCORnet CDM and the OMB racial categories describe different classes of people. The OMB guidelines allow people to select more than one race [14]. PCORnet CDM does not. Instead, the PCORnet CDM has a class for multiple race. Consider a person who identifies as both Black and Asian according to the OMB definitions. According the OMB guidelines in which a person can select more than one race, someone could identify as both Black and as Asian, and that person would be retrieved by a query for people who identified as Black, people who identified as Asian, and people who identified as both. If the same person were filling out a medical intake form using the PCORnet CDM guidelines, they would be instructed to choose only one race. They could, therefore, choose either Black or Asian or multiple race, but they could not choose both Black and Asian. With OMB standards, the classes of people who identify as Black and who identify as Asian can overlap. For the PCORnet CDM, they are disjoint. Therefore, the class of people who can identify with OMB Asian is not identical with the class of people who can identify PCORnet Asian but is actually a superclass class. It is worth noting that transforming OMB compliant racial data into the PCORnet CDM results in an irretrievable loss of information. Namely, persons who have identified with multiple OMB races will be indicated as identifying with the semantically less rich category "multiple races" in the PCORnet CDM. This loss of information is revealed by accurately representing the semantics of these coding schemes, but, in such cases of loss of information, not even a good ontology can not recover information that has not been stored.

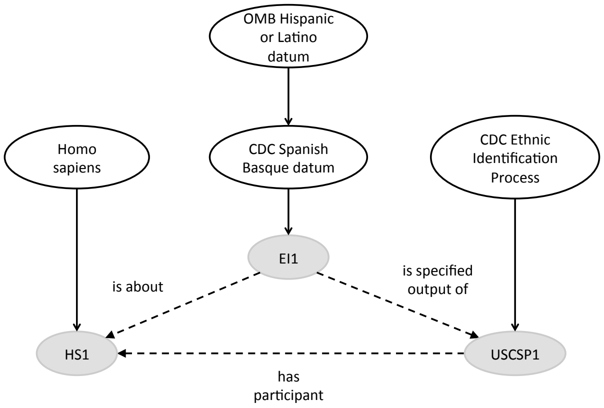

Fig. 3. Representation of Instance Level Social Identity Data

We developed a strategy for representing social identity data that supports integrating OMB and PCORnet CMD data. This strategy is not idiosyncratic to these data models, but is generalizable. This representation involves articulating the relations among classes of people who identify with OMB Asian and those who identify with PCORnet Asian, as an example. The OMB category Asian means the person has declared some Asian descent. The PCORnet CDM category Asian means the person has declared only Asian descent. Fig. 2 illustrates how identification processes and identification data that result from these two heterogeneous coding schemes are related. Notice that PCORnet racial identity datum is not a subclass of OMB racial identity datum. Since the PCORnet racial identity categories actually have a different meaning from the OMB racial identity categories, it would be inappropriate to use subclass relations to connect them. We are currently considering using SKOS:broader and SKOS:narrower to describe the relations between the intentional meanings of the terms, but it is not clear that this will support data retrieval.

V. Validation and Results

Competency questions are frequently used to validate modeling decisions in ontologies. They are questions that reflect the needs of the end user and that the ontology ought to be able to support. We partially validated the suitability of this representation for data retrieval and data integration with the following competency questions below. This validation is only partial since there are outstanding competency questions that require additional modelling decisions. We generated an OWL file with synthetic individuals and constructed Description Logic queries that answered three out of four of the competency questions. These queries in Manchester syntax are listed below. The OWL file with synthetic individuals is available at https://github.com/ufbmi/socid.

1. Which people are racially identified as Asian according to the OMB criteria?

TABLE III. Sample definitions for representing racial identity data

Ontological Definitions	
OMB racial identity datum	A racial identity that is the output of a racial identification process that uses OMB terminology for race or terminology that is mapped the OMB race terms.
OMB Asian identity datum	An OMB racial identity datum about a person who is identified as having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent.
Subject of an OMB Asian identity datum	A human being who is the subject of an OMB Asian identity datum
Subject of a self-identified OMB Asian identity	A human being who is the subject of an OMB Asian identity datum and who is the agent of the planned process for which that identity is a specified output.

inverse 'is about' some 'Asian identity'

2. Which people are racially identified with multiple races according to OMB criteria?

inverse 'is about' min 2 'OMB racial identity'

3. Which people are racially identified with more than one race in either OMB or PCORnet CDM?

inverse 'is about' min 2 'OMB racial identity' or inverse 'is about' some 'PCORnet multiple race identity

4. Which people are racially identified only as Asian according to OMB or PCORnet criteria?

Competency Question 4 requires indicating that each of the OMB race categories are different. For example, we must decide whether the classes OMB Asian identity datum and OMB Alaska Native or Native American datum are disjoint. Adding a disjointness axiom would rule out the possibility of a single identity datum item that indicates that person has both identities, but may support this competency question. Future work will focus on the best way to represent this situation.

We have included this representation of identity data in OMRSE, available at www.github.com/ufbmi/omrse.

VI. Discussion

This proposal diverges from traditional realist approaches insofar as it advocates representing social identities in terms of their verification criteria rather than according to their ontological properties. This approach has the advantage of supporting data integration and retrieval according to realist principles, without making dubious ontological commitments. It also does not sacrifice clear semantics, interoperability of data, or data retrieval. While our competency questions only address racial identity, they do show that different types of social identity data that have been gathered according to different criteria can be adequately represented according to the general ontological principles described in this paper. Analogous questions involving ethnicity and gender identity can be expected to be handled by this approach since they have the same logical form.

Future work includes representing relations between types of identity data to handle the remaining competency question, developing a set of gender identity terms to include in OMRSE, and query real patient data to assess the impact of this representation on cohort discovery tasks that include race and ethnicity.

VII. Conclusions

Our hypothesis was that representing social identity data with respect to processes of identifying rather than identities themselves can support data integration and retrieval in a realist framework while avoiding controversial ontological commitments.

We have produced a BFO-based representation of race and ethnicity identities and developed strategies for semantically integrating social identity data that have been collected using a) the OMB minimal categories for race and ethnicity, b) extensions of the OMB minimal categories for race and ethnicity, and c) common data models such as the PCORnet CDM whose semantics differ from the OMB minimum categories due to pick one/pick many discrepancies. We have added this representation to the OMRSE and produced a synthetic data set in an OWL file to test our competency questions. Our representation to date handles three out of four of our competency questions.

Acknowledgments

Thanks to William R. Hogan for reviewing and commenting on the manuscript and to the Clinical and Translational Science Ontology Group for providing feedback on a presentation of earlier work at the Charleston, SC meeting in September 2015.

References

[1] Institute of Medicine (US) Committee on Lesbian G, Bisexual, and Transgender Health Issues and Research Gaps and Opportunities,. The health of lesbian, gay, bisexual, and transgender people: Building a foundation for better understanding. Washington (DC): National Academies Press (US); Buffalo, New York: 2011.

[2] Hogan WR, Garimalla S, Tariq SA, editors. Representing the reality underlying demographic data. International Conference on Biomedical Ontologies (ICBO); 2011.

[3] Institute of Medicine (US) Committee on Lesbian G, Bisexual, and Transgender Health Issues and Research Gaps and Opportunities. Collecting sexual orientation and gender identity data in electronic health records: Workshop summary. Washington DC: The National Academies Press, 2013 0309268044 9780309268042.

[4] Cahill SR, Baker K, Deutsch MB, Keatley J, Makadon HJ. Inclusion of sexual orientation and gender identity in Stage 3 Meaningful Use Guidelines: A huge step forward for LGBT health. LGBT health. 2015.

[5] Department of Helath and Human Services CfMaMS. 42 cfr parts 412 and 495, [cms-3310-fc and cms-3311-fc], rins 0938-as26 and 0938-as58. Medicare and Medicaid programs; Electronic Health Record Incentive Program—Stage 3 and modifications to Meaningful Use in 2015 through 2017. 2015 October 7.

[6] Department of Helath and Human Services CfMaMS. 45 cfr part 170, rin 0991-ab93. 2015 edition health information technology (healthit) certification criteria, 2015 edition based electronic health record (ehr) definition, and onc health it certification program modification. 2015 October 6, 2015.

[7] Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied Ontology. 2010;5(3-4):139.

[8] James M. Race 2016 [updated March 16, 2016; cited 2016 April 19]. Available from: http://plato.stanford.edu/archives/spr2016/entries/race/.

[9] Mikkola M. Feminist perspectives on sex and gender 2016 [updated January 29, 2016; cited 2016 April 19]. Available from: http://plato.stanford.edu/archives/spr2016/entries/feminism-gender/.

[10] Revisions to the standards for the classification of federal data on race and ethnicity, (1997).

[11] Helsing K, editor Capturing social and behavioral domains and measures in electronic health records. 143rd APHA Annual Meeting and Exposition (October 31-November 4, 2015); 2015: APHA.

[12] Racial and ethnic categories and definitions for NIH diversity programs and for other reporting purposes, NOT-OD-15-089 (2015).

[13] PCORnet Common Data Model (cdm) [updated Last updated on December 18, 2015 cited 2015 April 25]. Available from: http://www.pcornet.org/pcornet-common-data-model/.