UC San Diego
SCHOOL OF MEDICINE
Department of BioMedical Informatics

# Review and Evaluation of the State of Standardization of Computable Phenotype

Stephanie Feudjio Feupe, MSc, Ko-Wei Lin, PhD, Tsung-Ting Kuo, PhD, Chun-Nan Hsu, PhD, Hyeon-Eui Kim, PhD
*Department of Biomedical Informatics, UC San Diego, La Jolla, USA*

## Abstract

*Inadequate standardization of phenotype data in biomedical database and electronic medical record (EMR) has created challenges in data reuse for patient care, quality improvement, new scientific discovery, and validation of existing knowledge. We conducted literature review on this topic published in the past 10 years to assess the phenotype standardization efforts and summarized the findings here.*

## Background/Introduction

A phenotype is an observable or measurable characteristic or trait in an organism. The ever growing need to investigate the associations among phenotypic data and other types of data such as genetic, pharmaco-genetic, environmental, psycho-behavioral data makes it crucial to have standard for phenotype representation across databases. While the availability of phenotype data in the EMR, clinical databases and other supporting sources bring new opportunities for large scale data driven research such as genome wide association studies and cohort discovery; the lack of clarity in phenotype definitions and the idiosyncrasies in data representations make it a non-trivial task to render phenotype data interoperable and sharable. Given the significance of phenotypic data and the growing efforts to make phenotype data computable, we reviewed and analyzed published literature to access current state of the research in order to understand what efforts have been done, to identify potential gaps and unexplored areas for developing further research focus on this topic.

## Methods and Materials

### Search strategy and outcome

| Pubmed query | WOS query |
|---|---|
| Phenotype [MeSH] AND (standardization OR "vocabulary, controlled" [MeSH] OR "terminology as topic" [MeSH]) AND Published year: past 10 years AND Language: English AND Abstract available AND Human studies | TOPIC: (Phenotype OR "computable phenotype") AND TOPIC: ("standardization , definitions" OR "vocabulary, controlled" OR " terminology as topic") AND YEAR PUBLISHED: (2004-2015) AND Language: English |
| 611 articles found | 440 articles found |
| Total number of articles found: | 1051 |

## Conclusions

15 relevant articles met our eligibility criteria, thus were included in our review. The articles were summarized using a standardized review form to include methods, results, limitation, etc., as shown in the table here. Phenotypes are computable and integrable when they can be described to a more granular level as possible. Methods for such representation ranged from human experts annotation to computerized algorithms. In the past year, efforts like eMERGE, PhenX, HPO have attempted to standardize phenotype data from omics research databases; whereas, the combination of information models and terminology system like UMLS and SNOMED-CT has been used to assure EHR interoperability and phenotype identification. Most methods for rendering phenotype computable in these reviewed articles focused on specific types of phenotypes or the data from a single data repository, which constitutes a factor against the generalizability.

**(Scan the barcode to see the full results table).**

## Results/Evaluation

| *Author/ Year* /Title | Data Source | Methods | Results |
|---|---|---|---|
| ***Pathak J, 2013*** Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. | EHR (both structured & narratives) | Mapping EHR data to information model (CEM) and standardized terminologies. Developing an NLP pipeline for narrative texts (6 templates based on CEM) → rules & supervised learning Tested with QDM (Quality Data Model) use case | SHARPn architecture and toolkit |
| ***McCray AT, 2014*** Modeling the autism spectrum disorder phenotype | 24 standardized screening or diagnostic instruments for autism spectrum disorders | Ontology building using "hybrid" method: combination of top-down and bottom-up (cluster analysis), manual review & refinement | Total 283 concepts under 3 top classes Mapped to UMLS, ICF, MeSH 5000 questions mapped to the ontology and normalized into 3395 Fully integrated with the Autism Consortium database |
| ***Doan S , 2014*** PhenDisco: phenotype discovery system for the database of genotypes and phenotypes | dbGaP phenotype data dictionary | lite NLP based standardization pipeline was built using Metamap and heuristic rules | standardization pipeline |
| ***Burgun A , 2009*** Two approaches to integrating phenotype and clinical information | - Mammalian Phenotype Ontology (MPO) - OMIM - Unified Medical Language System (UMLS) | For terminology mapping, authors used exact and string matching. For mapping through gene annotations, authors used UMLS dictionary as a resource to map between MPO and OMIM. | 1,469 MPO concepts (22%) were mapped successfully to UMLS's concepts, of which 869 were present in OMIM. 1,968 genes were associated with both MPO and OMIM annotations. |
| ***Pathak J, 2011*** Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGENetwork experience. | Phenotype data dictionaries from 5 different eMERGE Network medical sites studying different diseases. | 157 data elements were mapped to cdDSR and SNOMED CT using both lexical (search for relevant pre-coordinated concepts and data elements) and semantic (post-coordination) techniques. New data elements were curated when feasible. | 60% target data elements can be mapped using lexical techniques. After post-coordination with curating new caDSR CDEs and new NCI thesaurus concepts |
| ***Cohen R, 2011*** CSI-OMIM - Clinical Synopsis Search in OMIM | OMIM syndrome entries. Clinical synopsis (CS): in structured text that outlines signs, symptoms of the disease. | (1)Define phenotype areas (2)Apply Natural Language Processing methods tagging each phrase with semantic information of UMLS and MESH, and clustering (3)Compute pair-wise similarity. | Define 26 phenotype areas. 79770 new connections were discovered, adding 16 new connections per syndrome on average. Precision 93.5%. Web application CSI-OMIM was created. |
| ***Kerkhof HJ, 2010*** Recommendations for standardization and phenotype definitions in genetic studies of osteoarthritis | OA phenotypes of the 28 studies | (1)Assess whether different OA definitions result in different association results (2) Create consensus OA definition from definitions used within the consortium in the Rotterdam Study-I (3)tested the association of hip OA with gender, age and BMI using one-way ANOVA (4)provide recommendation for OA definition for future studies | A standardization of radiographic OA definitions was made to reduce heterogeneity but SOA phenotypes standardization show to be difficult because of the difference and specific of the OA by studies. A list of recommendation for future OA studies, including mainly a more precise definition if the OA. |
| ***Hoehndorf R, 2010*** Interoperability between phenotype and anatomy ontologies. | Unspecified | Reusing classes and relations from different ontologies to provide means to formalize phenotypic traits. Describe methods for integrating phenotypic descriptions with canonical ontologies. Assess of the method. | Provide the means to integrate phenotypic descriptions with ontologies of other domains. The framework leads to the capability to represent disease phenotypes, perform powerful queries that were not possible before and infer knowledge. |
| ***Riggs ER, 2012*** Phenotypic information in genomic variant databases enhances clinical care and research: the International Standards for Cytogenomic Arrays Consortium experience | Survey from (a) Practicing clinicians (b) Research groups | The origin of this NCBI-sponsored ClinVar database is about genotypes. Mainly based on 'trigger phrases' that are mapped to the most specific representative HPO term, and also consider negations and uncertain finding. The collected phenotypes are normalized using HPO and linked to MedGen / Orphanet / OMIM. | Provide a database ClinVar, as well as annotation-form and external data submission tools. |

Scan for a complete view of the results

**Acknowledgements**
This research was supported by the NLM grant # 5T15LM11271-2 to the University of California, San Diego

sfeudjiofeupe@ucsd.edu