Systematic Integration of Large-scale Clinical and Phenotype Datasets: The SickleInAfrica Registry

Wilson Mupfururirwa, Jack Morrice, Jade Hotchkiss, Raphael Sangenda, Annemie Stewart, Victoria Nembaware,Mario Jonas Gaston K. Mazandu†, Nicola Mulder, Ambroise Wonkam contributing members of the SickleInAfrica Database and Research Working Groups, Coordinators ae

Affiliations

1 Division of Human Genetics, Faculty of Health Sciences, University of Cape Town, South Africa.

2 Computational Biology Division, Faculty of Health Sciences, University of Cape Town, South Africa.

3 Sickle Cell Programme, Muhimbili University of Health and Allied Sciences (MUHAS), Dar es Salaam, Tanzania.

**SickleInAfrica Database and Research Working Group contributing members in the table

Abstract:

Decision making in public health is largely dependent on data analytics results. Advances in data collection have yielded large-scale heterogeneous clinical and phenotype datasets from different geographical locations, however, harmonizing these datasets retrospectively for integrative analyses to potentially increase prediction power is still challenging. We present omsim, a model-based ontology mapping and text graph-based similarity information retrieval technique, for automated generation of harmonized datasets from disparate research patient registries. We tested omsim on sickle cell patient multinational research datasets in sub-Saharan Africa.

Key words: Clinical data, sickle cell disease, data harmonization, ontology mapping, data integration, NLP, SickleInAfrica, research data.

Generating large numbers of samples to achieve sufficient predictive power [1] is essential to increase the confidence level of study outputs and minimize potential false positive rates. Advances in high-throughput technology have led to the generation of large-scale genomic, clinical and phenotype datasets from different research sites, in different geographical locations. Harmonizing these datasets across the different sites to facilitate the integration of these datasets can increase sample size and boost the predictive power. Genomic datasets are generally standardized and easy to harmonize and integrate considering existing genomic data standards [2], however, clinical and phenotype datasets are highly heterogeneous and systematically harmonizing such datasets remains a challenge. This indicates that integrating biomedical research datasets with the electronic health record (EHR) is currently very expensive and challenging, limiting health metadata interoperability and research translation.

Several collaborative research studies [3,4,5], including Human Heredity and Health in Africa (H3Africa) [6] and SickleInAfrica [7], have collected large scale phenotype and clinical datasets using advanced clinical data management systems, such as the Research Electronic Data Capture (REDCap) platform [8]. These systems have enabled the development of detailed case report form (CRF), resulting in big datasets with large numbers of variables from different locations, generally using different variable names and codes, which do not allow direct prospective or retrospective comparisons. Such datasets need to be harmonized prior to integration for potential analyses and inferences. However, considering the number of variables included and dataset sizes, performing

manual harmonization and integration is daunting, time-consuming and highly prone to errors. Collaborative consortia [4,9,10,11], such as the World Endometriosis Research Foundation (WERF) Endometriosis Phenome and Biobanking Harmonisation Project (EPHect) [9], developed guidelines for standardization and harmonization of phenotype and clinical research data and biological sample collection. Unfortunately, most existing guidelines rely heavily on manual tasks.

A semi- or full automated variable mapping process with subsequent data integration, referred to as automatic data harmonization, should help minimize manual interference and errors, and possibly enhance data quality [12]. Thus, in addition to the lack of an appropriate data harmonization formalism, automatic data harmonization models are scarce and only a few models, including DataSHaPER [3] and BiobankConnect [13], have been suggested. These two models are semi-supervised and were tested using retrospective epidemiological datasets from 53 large population research studies of the Public Population Project in Genomics (P^3G) [4]. Although DataShaper was built based on the guidelines and a core set of variables pooled for harmonizing datasets generated by these 53 population-based studies, it showed a low proportion of variables that could be harmonized (only 38%) and 47% of the variables were impossible to map. BiobankConnect uses ontology (searching for synonyms) and lexical or string pattern matches to semi-automatically map targeted data elements, producing average precision and recall scores, when assessed to a manually curated set of relevant matches [13]. Even though BiobankConnect showed some improvement as compared to DataSHaPER, it has produced poor precision and recall values in several datasets.

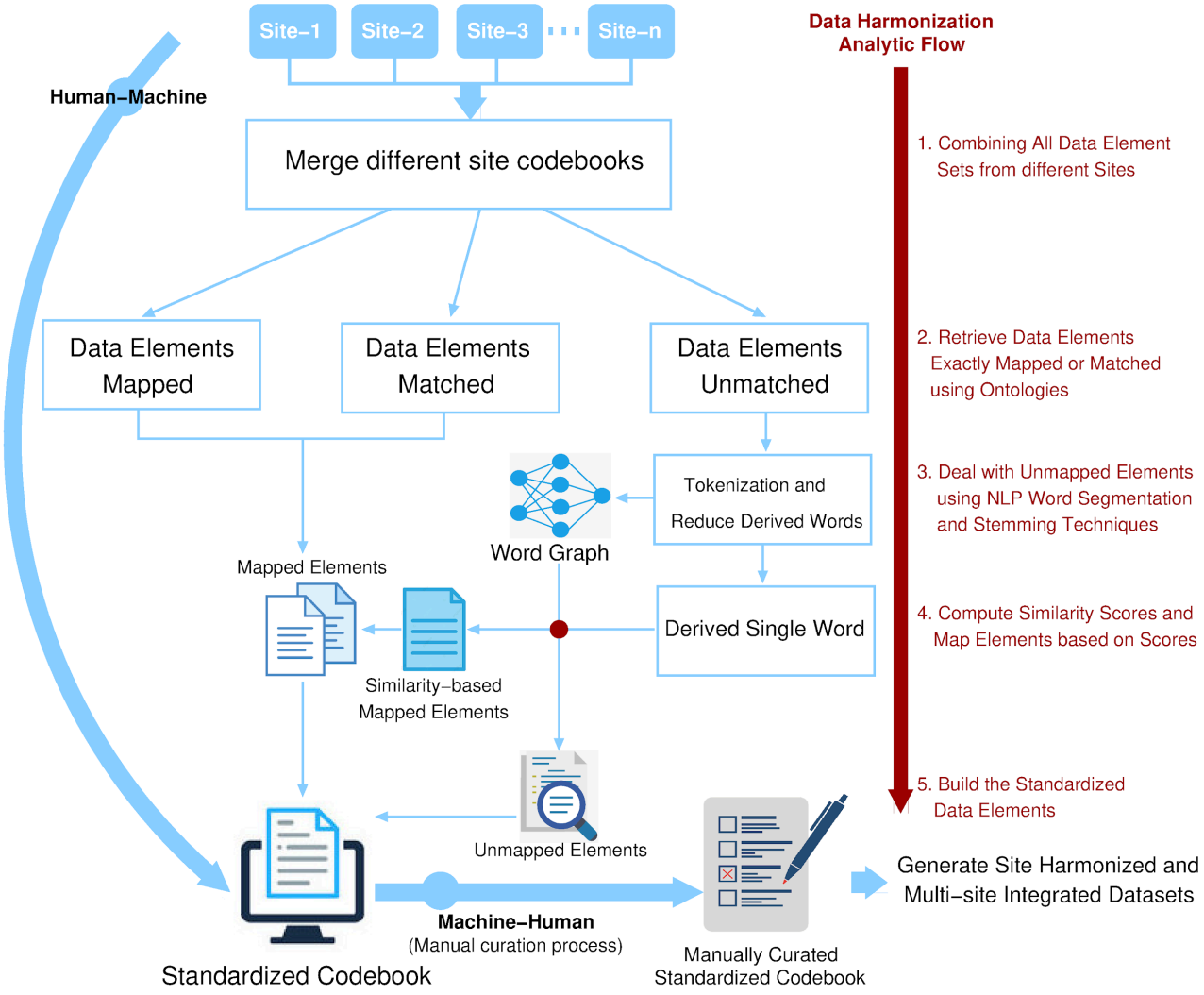

Figure 1: Automated data harmonization workflow. Different processes and steps for merging, harmonizing and integrating datasets from different sources to generate a consistent standardized data element set and multi-source integrated dataset.

In fact, a lexicon may substantially contribute to transforming data into knowledge when efficiently implemented [14], however, a simple lexical or string pattern match model may not perform as it would be expected, especially in the context of big data, dealing with very large heterogeneous sets [12]. Within natural language processing (NLP) domain, graph and semantic similarity-based algorithms have been suggested for word sense disambiguation (WSD), i.e., automatically inferring the meanings (senses) of a given target word, specific to the context without recourse to a given dictionary, providing an unsupervised model without requiring data hand labeled with correct word senses [15]. We introduce omsim, summarized in **Figure 1**, an ontological search matching coupled with NLP-based word graph similarity information retrieval model for automatic data harmonization and integration in two main systematic processes: human-to-machine (H2M) and machine-to-human (M2H). In the H2M process, the data element mapping is performed without human or manual interference and M2H consists essentially of a manual curation process, mainly in cases where more than one matching variable is identified, necessitating domain expert judgment. In this case, all candidate matches are sorted based on similarity scores, so the domain expert can quickly decide on a more appropriate match.

The main aim of the omsim tool is to speed up phenotype and clinical research data pooling (harmonization and integration), automatically mapping given data element sets to a standardized data element set built from targeted sets (see ONLINE METHODS for details) as shown in **Figure 1.** Considering that the complete meaning of a word is always contextual [16] and, subsequently, the meaning of the word sequence is reflected by the words in the sequence [17], WSD for unmatched words is performed using the word graph in which stem word sequences are paths representing possible interpretations of the word sequence to be disambiguated. Graph nodes correspond to words and word co-occurrence represents an edge between two nodes, which can be used for comparing words in different data dictionaries [18]. In this context, one possible approach to define a synonym distance metric is to assume that two words are close from being synonyms if they co-occur in the same word sequences and have the same words in their definition. Thus, we define a distance between two words using the adjacency matrix of the word graph constructed, representing word co-occurrence. From this distance, a semantic similarity model is derived using a linear transformation (see ONLINE METHODS for more details). The semantic similarity score between sequence words is computed using best match average [19] and words with high similarity scores meeting the threshold are mapped to the word for which WSD is applied.

The omsim model is tested on harmonizing SickleInAfrica data dictionaries [7], the largest pan-African sickle cell disease (SCD) research networks to date, currently comprising Tanzania, Nigeria and Ghana. The SickleInAfrica platform contains a multi-national registry containing clinical data of over 10 000 SCD patients in over 27 data collection instruments, embedding approximately 351 data elements (see **Figure 2(A)**). A set of manually selected best matches between sites was constructed by the SickleInAfrica database working group (See online Supplementary File 1 and ONLINE METHODS for the standard operating procedure used for creating the matches and the quality assurance procedures followed). Using this set, we evaluated the prioritization of matches in the results generated by omsim (see **Figure 2(B)**) using known performance metrics, including area under the curve (AUC) of receiver operating characteristics (ROC), referred to as AUC-ROC, as well as average accuracy, positive predictive value or precision, sensitivity and specificity scores. We calculated omsim performance scores for different mapped data elements across the three research sites. Overall, we observed omsim shows a very good performance matching with

AUC-ROC, average accuracy, precision, sensitivity and specificity scores greater than 94% (**Figure 2(B)**). The numbers of shared best matches and omsim suggested matches are highlighted in **Figure 2(A)** and **(C)**, respectively. Looking at different omsim mapped elements, text graph-based similarity search, with a similarity threshold set to 0.7, led to the best performance with approximately half numbers of matches that would otherwise have been missed, providing a significant change with p-value < 0.0018 (χ^2 test).

We have described a systematic and comprehensive harmonization scheme and implemented omsim, for variable mapping between research site data and standardized variables to achieve an optimal harmonization. We applied this scheme to recent large SCD datasets generated from different research sites in sub-Saharan Africa, yielding an optimized harmonized datasets across research studies with performance scores greater than 98, 96, 94, 97 and 98% AUC-ROC, average accuracy, precision, sensitivity and specificity, respectively. omsim is a tool speeding up the harmonization and integration of heterogeneous phenotype and clinical datasets, with potential for use in other big dataset integration challenges. This also enables a potential connection between retrospective disparate research patient registries and patient electronic health records and sets up a possible follow-up SCD cohort at the health-care level. Finally, this tool can be used to ease the harmonization process and will be extensively used to automatically harmonize worldwide global SCD metadata [20] from the Global SCD Registry Portal (https://www.sickleinafrica.org/scd-registry-form) to speed up a refined global standardized data element set for collecting SCD phenotype and clinical datasets. omsim is a Python package available at https://github.com/sickle-in-africa/data-harmonisation, with a strong commitment to maintaining and updating the resource quarterly.

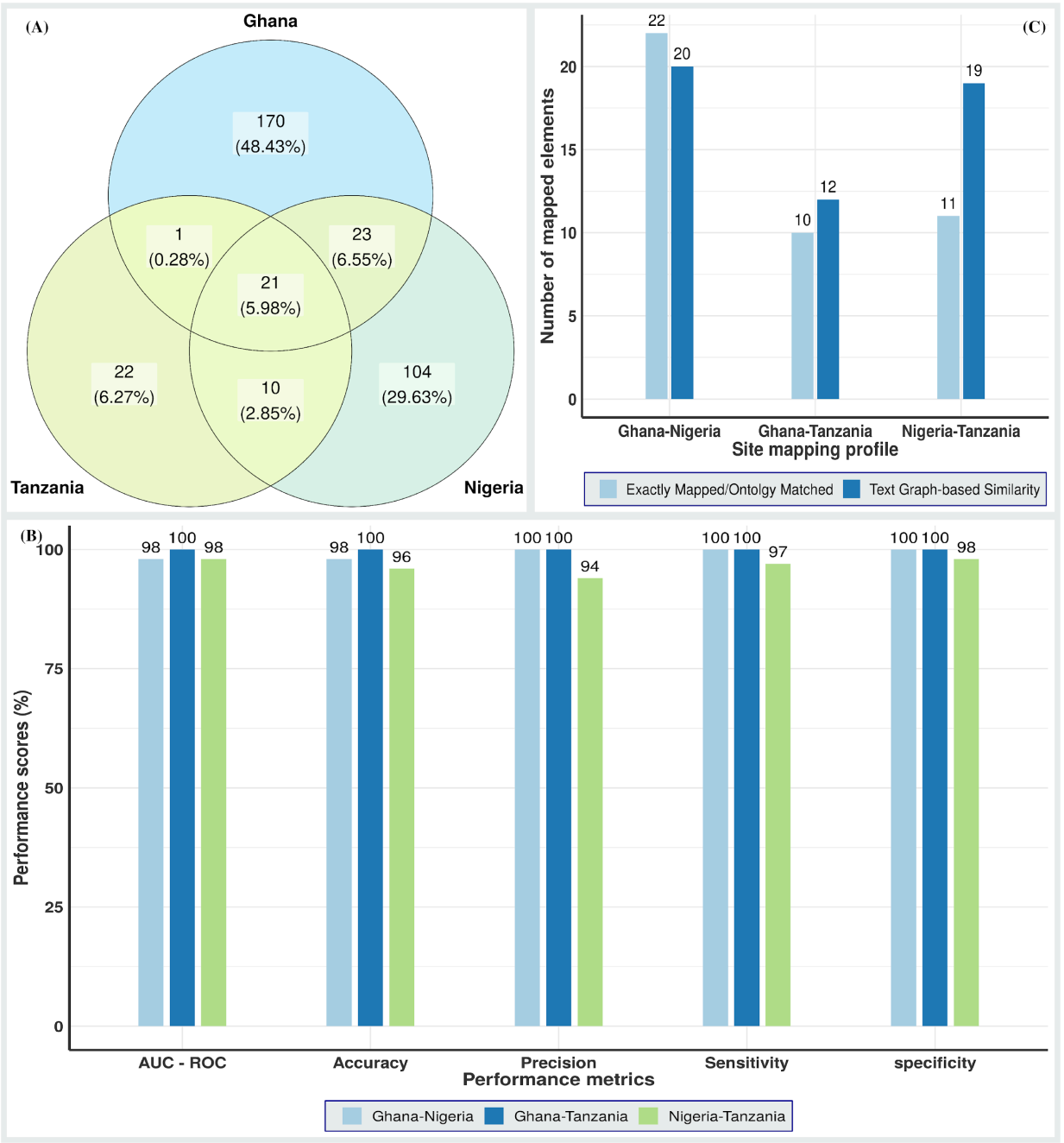

Figure 2: The omsim performance evaluation. (A) Numbers of data elements used by different sites showing numbers of mapped elements between sites. (B) Different omsim performance scores assessing the model classification power. (C) Numbers of mapped data elements retrieved via omsim using ontological search matching (synonyms or exact matching) and text graph-based similarity information retrieval model.

Methods

A comprehensive explanation of the model and associated accession code links are provided in the online version of the manuscript.

Acknowledgements

The SickleInAfrica consortium is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH) under grant numbers U24HL135600 and U24HL13045881.

Author Contributions

SickleInAfrica consortium PIs designed the project. GKM developed the computational framework and continues to maintain omsim via

https://github.com/gkm-software-dev/sickleinafrica-database, led and executed the benchmarking analyses and comparisons with input from RS, AS, VN and the SickleInAfrica Database Working Group who provided datasets for testing the package. GKM drafted the manuscript with input from all authors from the SickleInAfrica Consortium and all authors approved the final version of the paper for submission.

Competing Financial Interests
The authors declare no competing financial interests.

References

1. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, et al. Harmonization of quality metrics and power calculation in multi-omic studies. Nat Commun 2020;11:3092.
2. McGarvey PB, Nightingale A, Luo J, et al. UniProt genomic mapping for deciphering functional effects of missense variants. *Hum Mutat*. 2019;40(6):694-705.
3. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol*. 2010;39(5):1383-1393.
4. Fortier I, Doiron D, Little J, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol*. 2011;40(5):1314-1328.
5. McCarty CA, Huggins W, Aiello AE, et al. PhenX RISING: real world implementation and sharing of PhenX measures. *BMC Med Genomics* 2014;7:16.
6. Joubert BR, Berhane K, Chevrier J, et al. Integrating environmental health and genomics research in Africa: challenges and opportunities identified during a Human Heredity and Health in Africa (H3Africa) Consortium workshop. *AAS Open Res*. 2019;2:159.
7. Makani J, Sangeda RZ, Nnodu O, et al. SickleInAfrica. The Lancet Haematology 2020;7(2):e98-e99.
8. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208.
9. Vitonis AF, Vincent K, Rahmioglu N, et al. World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonization Project: II. Clinical and covariate phenotype data collection in endometriosis research. *Fertil Steril*. 2014;102(5):1223-1232.
10. Fortier I, Dragieva N, Saliba M, Craig C, Robson PJ; with the Canadian Partnership for Tomorrow Project's scientific directors and the Harmonization Standing Committee. Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis. *CMAJ Open*. 2019;7(2):E272-E282.
11. Griffith LE, van den Heuvel E, Raina P, et al. Comparison of Standardization Methods for the Harmonization of Phenotype Data: An Application to Cognitive Measures. *Am J Epidemiol*. 2016;184(10):770-778.
12. Eine B, Jurisch M, Quint W. Ontology-Based Big Data Management. Systems 2017;5:45.
13. Pang C, Hendriksen D, Dijkstra M, et al. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. *J Am Med Inform Assoc*. 2015;22(1):65-75.
14. Kuiler EW. From Big Data to Knowledge: An Ontological Approach to Big Data Analytics. Review of Policy Research 2014;31(4).

15. Navigli R, Lapata M, An experimental study of graph connectivity for unsupervised word sense disambiguation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 678-692.
16. Firth JR. Meaning by collocation. Transactions of the philological society 1935;34(1):36-73.
17. Li Y, McLean D, Bandar ZA, et al. Sentence similarity based on semantic nets and corpus statistics. IEEE transactions on knowledge and data engineering 2006;18(8):1138-1150.
18. Lesk M. Automatic sense disambiguation using machine readable dictionaries. In Proceedings of the 5th annual international conference on Systems documentation (SIGDOC'86), ACM Press 1986; pages 24-26.
19. Mazandu GK, Chimusa ER, Mulder NJ. Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Briefings in bioinformatics 2017;18(5):886-901.
20. Nembaware V, Mazandu GK, Hotchkiss J, et al. The Sickle Cell Disease Ontology (SCDO): Enabling Collaborative Research and Co-Designing of New Planetary Health Applications. OMICS-A Journal of Integrative Biology 2020;24(10): 559-567.

ONLINE METHODS

Unified data harmonization formalism: logical scheme

We designed a logical scheme for data harmonization, which includes a global framework, referred to as standardized or mediated data element set. This standardized data element set is a non-redundant set of data elements, combining different data sources and providing the mapping between standardized data element sets and a given data source. So, a data harmonization system, H, can formally be represented by a 3-tuple or triplet $\left(\Sigma, S_k, \Psi\right)$, where

- Σ is a standardized data element set built using consistent or non-redundant data element sets from different research data sources to be harmonized.
- S_k is a consistent data element set from the research site k.
- Ψ is the mapping between Σ and S_k, constituted by a set of assertions of the form (q, p) and (p, q) where p and q are data elements in Σ and S_k, respectively.

It follows that :

$$\Sigma = \cup_{k=1}^{n} S_K$$

with n the number of sources and S_k the consistent data element set from the research site k. The assertion (q, p) in the mapping specifies that the concept q of the research site k data element set corresponds to the concept p in the standardized data element set (similarly for an assertion of (p, q)). These two fundamental models for specifying the mapping in a data harmonization system are referred to as local-view (LV) and global- or mediated-view (GV) references [21,22,23], respectively. It is worth noting that the above definition of data harmonization scheme is general enough to capture virtually all possible models in the literature.

In practice, the research site k describes the structure of the site, where the real data are, while the standardized data element set provides a reconciled, harmonized, and integrated virtual view of different research site data element sets, also referred to as codebooks. This suggests that each data element set is associated with a database (DB) satisfying constraints related to data quality assurance (QA) described in **Table xx**. DB associated with the standardized data element set is referred to as central DB (CDB) and the one associated with a site data element set is called site or source DB (SDB). We particularly assume that the structures constituting these databases, described in the next section, are defined over a fixed domain Ω, which, for the case study here, is the sickle cell disease (SCD) research. Note that specifying constraints in central DB would enhance the GV reference modeling power when expressing standardized data element set as a conceptual data model [23], in which case, the domain knowledge portal, e.g., ontology, controls either implicitly or explicitly integrity constraints in CDB.

Different site databases and central database

We first describe an SDB conforms to some S_k, i.e., satisfying QA in **Table xx**, as well as the LV mapping Ψ, and contributing to datasets contained in CDB. Currently, most research sources are collecting phenotype and clinical datasets using advanced clinical data management systems, including OpenClinica [24] and REDCap [8] platforms, providing an effective web application for building and managing online surveys and databases. SDB uses REDCap and is accessible via a secured data communication through secure sockets layer (SSL) supported web distributing and versioning (WebDAV) protocol. This schema is flexible enough, scalable and dynamic, which results in the ability to efficiently query and retrieve patient information, and can be easily adapted and extended with quality control setting, enabling SBD to satisfy QA constraints. In addition, the

REDCap system has useful plugins for data capturing, imports and exports in different formats, e.g., comma separated value (csv) format for further possible data quality checks and potential analyses. CDB, which is supposed to be built after harmonizing and integrating different research site datasets, will be a relational database using a MySQL database server and accessible via a web server running Linux with PHP/Python. This database should satisfy QA in **Table xx**, as well as the GV mapping Ψ, and integrating datasets from SDB. However, there is a high likelihood of violating data integrity as harmonizing different site data elements often results in a no reconciled set of variables.

Ensuring standardized codebook consistency

In merging codebooks from different research sites, according to the harmonization system, H, defined above, data elements retrieved from sites can be mapped (exact matches), matched (using ontologies to search for synonyms) and unmatched (neither mapped nor matched at the start). The unmatched data elements are further processed using natural language processing (NLP) based word graph similarity information retrieval model as described below to identify similarity-based mapped elements as well as the final set of unmapped elements which could not be reconciled. This harmonization process is very often needed when dealing with retrospective disparate research participant registries which do not use a commonly defined codebook, to enable the central database to comply with the consistency constraint. This also eases the data integration process for a large multiple geographical location observational study analyses, including comparative and analytical analyses. Optimal standardized data element set is generated using two steps: (1) Identify mapped, ontology-based matched and similarity-based mapped elements between research sites and add them to standardized data element set, and then (2) add no reconciled data elements between research sites in the standardized data element set.

Ontology-based data element matching

An ontology provides a framework for sharing information among different disciplines supporting the interoperability requirement to transmit, reuse, and share data with predetermined semantic and syntactic rules [14]. Several large systems adopt ontologies to enable data integration and management with a high degree of automation [12]. In the biomedical context, several ontologies have been developed and online resources, including the National Center for Biomedical Ontology (NCBO) BioPortal [25] with more than 1000 ontologies currently available, have been designed to facilitate access to different ontologies. The selection of an ontology to use in a given application needs to be done carefully, ensuring that it unambiguously describes targeted domain terminology and different concept associations.

For testing omsim, we selected sickle cell disease ontology (SCDO) [26], human phenotype ontology (HPO) [27], human disease ontology (DO) [28] and Monarch Merged Disease Ontology (MONDO) [29], experimental factor ontology (EFO) [30] containing disease-related concepts associated with our case study, and The National Cancer Institute (NCI) Thesaurus (NCIt) [31] and Systematized Nomenclature Of Medicine (SNOMED) Clinical Terms (SNOMED CT) [32] as both store broad ranges of concepts related to the data elements under consideration. These ontologies are used for searching potential annotation synonyms to targeted data elements. This is very useful for data elements which could not be directly mapped by searching similar or more specific concepts that can be considered as proxies. As illustrations, ‘sex’ and ‘gender’, ‘hypertension’ and ‘high blood pressure’, ‘red blood cell’ and ‘RBC’ or ‘RBCs’ can be considered similar based on their meanings (synonyms) and may be potentially matched using ontology searches.

Dealing with unmatched elements and similarity-based data element mapping

Initially, we build an undirected graph $G = (N, L)$ where N is set of nodes, which are induced words from unmatched research site data elements and L the set of links connecting words, expressing semantic relations or co-occurrences, between connected words. First, we consider that word sequences are 'part-of-speech tagged' [15], using content words only (i.e., nouns, verbs, adjectives and adverbs) and eliminating no contextual words including articles, conjunctions, pronouns and preposition, as well as punctuation and special characters. We then use sequence word segmentation or tokenizer, word stemming and construct the word graph:

- The segmentation process consists of splitting the text into smaller units called tokens, which, in this context, word sequence is split into words, and
- The stemming process is applied to each segmented words to reduce the diversity between the derived forms of words, generating a root (base or stem) form of each derived (or infected) word in the word sequence with similar meaning using stemming technique as implemented in the Python NLP toolkit, NLTK [33].
- The output word sequence $\omega = \left(\omega_1, \omega_2, ..., \omega_p\right)$, with ω_j, $1 \leq j \leq p$, word stem (root, base or no infected), is set as a path in the graph, in which case, the number of occurrences for a given word in the site codebook is simply the number of paths passing through the word.

For each unmatched data element in a given research site, we perform word sense disambiguation, identifying the intended meanings (senses) of words specific to the context [15] that is the most word graph-based semantically similar associated word [34,35] in another codebook using a semantic similarity model [34,35,36,37,38] and maximizing semantic similarity score between words x and y, using best match average (BMA) model [19,36,39] given by:

$$S(x,y) = \frac{1}{2}\left(\frac{1}{p}\sum_{=1}^{p} \max_{j=1,...,q} S\left(\pi_x, \pi_{y_j}\right) + \frac{1}{q}\sum_{j=1}^{q} \max_{i=1,...,p} S\left(\pi_{x_i}, \pi_{y_j}\right)\right)$$

where p and q are numbers of word stems in sequence words x and y, respectively, contained in the word graph and π_{w_s} a graph-based word stem w_s profile, which is simply the row or column corresponding to w_s in the adjacency matrix of the word graph, $A = \left(a_{ij}\right)_{1 \leq i,j \leq n}$, where a_{ij} (the element in ith row and jth column) is 1 if node i, j are linked, 0 otherwise. So, formally the profile of a word w_s, with index k in the adjacency matrix, is a $0 - 1$ vector, π_{w_s}, given by:

$$\pi_{w_s} = \left(a_{kj}\right)_{1 \leq j \leq n}$$

Finally, the semantic similarity score, $S\left(\pi_{w_s}, \pi_{z_t}\right)$, between word stem profiles, π_{w_s} and π_{z_t}, is computed by linearly transforming the word graph based normalized distance, $D\left(\pi_{w_s}, \pi_{z_t}\right)$ [19], and given by:

$$S\left(\pi_{w_s}, \pi_{z_t}\right) = 1 - D\left(\pi_{w_s}, \pi_{z_t}\right)$$

with $D\left(\pi_{w_s}, \pi_{z_t}\right)$ computed as follows:

$$D\left(\pi_{w_s}, \pi_{z_t}\right) = \frac{\delta\left(\pi_{w_s}, \pi_{z_t}\right)}{\Delta\left(\pi_{w_s}, \pi_{z_t}\right)}$$

where $\delta\left(\pi_{w_s}, \pi_{z_t}\right)$ is the Minkowski distance of order p (p an integer) [40] given by:

$$\delta\left(\pi_{w_s}, \pi_{z_t}\right) = \left(\sum_{j=1}^{n} \left|a_{kj} - a_{ij}\right|^p\right)^{\frac{1}{p}}$$

where i is the ith index of adjacency matrix corresponding to or profile of word stem z_t, n the dimension of the adjacency matrix. Some instances of this distance include the Manhattan distance ($p = 1$) and the Euclidean distance ($p = 2$), and in the context of this study, the Manhattan distance is used. It is worth noting that the use of Manhattan or Euclidean distance can be interchangeably used as these metrics are equivalent in R^n, which is a Hausdorff, separated or T_2 topological space. The normalization factor, $\Delta\left(\pi_{w_s}, \pi_{z_t}\right)$, of the metric value, $\delta\left(\pi_{w_s}, \pi_{z_t}\right)$, is computed based on the Minkowski inequality [41] and given by:

$$\Delta\left(\pi_{w_s}, \pi_{z_t}\right) = \left(\sum_{j=1}^{n} \left|a_{kj}\right|^p\right)^{\frac{1}{p}} + \left(\sum_{j=1}^{n} \left|a_{ij}\right|^p\right)^{\frac{1}{p}}$$

Note that, in some specific contexts, $D\left(\pi_{w_s}, \pi_{z_t}\right)$ could also be normalized as follows:

$$D\left(\pi_{w_s}, \pi_{z_t}\right) = \frac{\delta\left(\pi_{w_s}, \pi_{z_t}\right) - \delta_{min}}{\delta_{max} - \delta_{min}}$$

where δ_{max} and δ_{min} are the maximum and minimum distance value of the k and ith dimensions. Furthermore, instead of using semantic similarity based on linearly transformation of the word graph based normalized distance, one adopts to use the cosine similarity model [19,37], given by:

$$S\left(\pi_{w_s}, \pi_{z_t}\right) = \langle \pi_{w_s}, \pi_{z_t} \rangle = \sum_{j=1}^{n} a_{kj} a_{ij}$$

which may be normalized using usual and Tanimoto coefficient normalization schemes [19].
As pointed out previously, the similarity score formula above is applied for word sequence or connected word stems in the word graph. For isolated nodes, w_s and z_t in the word graph, the similarity score is computed Jaccard similarity model [19,42] as shown below:

$$S\left(w_s, z_t\right) = \frac{\left|w_s \cap z_t\right|}{\left|w_s \cup z_t\right|}$$

where $\left|w_s \cap z_t\right| = max\left\{q: \exists x_q \; a \; q \; - \; gram \; occurring \; in \; w_s \; and \; z_t\right\}$ with q-gram [43] the sub-string of length q and $\left|w_s \cup z_t\right| = max\left\{length\left(w_s\right), length\left(z_t\right)\right\}$.

Finally, a given word or word sequence is mapped to words or word sequences with high semantic similarity scores satisfying the threshold, which is data-driven and determined based on the frequency distribution of similarity scores generated.

Assessing performance of the data harmonization scheme suggested

For a given automated data harmonization tool, high sensitivity and specificity metrics are needed in the recognition of matching relevant variables. These metrics allow to assess a machine performance which, for a target data element, chooses the matching one in the source variable, with low risk of having missed a correct or mapping to a wrong source variable. Thus, to assess the model classification power, we computed different performance scores for different mapped data elements across data sources, including area under the curve (AUC) of the receiver operating characteristics (ROC), as well as average precision, sensitivity and specificity scores using the

standard ROCR package version 1.0.11 of the R programming language. We also use a χ^2 test to check the improvement significance contribution of text graph-based similarity model.

Generating harmonized site codebooks and multi-site integrated dataset

After an automated harmonization of the data elements in the H2M process, a data dictionary table mapping matched variables between different sites which is used to generate a non-redundant standardized codebook integrating different site data elements. These two datasets are optimized during the M2H process by performing a manual curation, mainly focusing on cases where more than one matching variable were identified, necessitating domain expert judgment. In this case, all candidate matches are sorted based on similarity scores to enable the domain expert to quickly decide on a more appropriate match. This optimized standardized codebook may be used to build a case report form (CRF) for collecting biomedical datasets in the context of prospective cohort studies. More importantly, it enables the integration of multi-site datasets by mapping different site data elements to the optimized standardized and merge these datasets, thus fostering a large-scale analysis, including comparative and analytical analyses across different geographical locations. It is worth mentioning that the harmonization model suggested is flexible, adaptable and generalizable and constitutes a tool for biomedical data-exchange and interoperability standard, allowing external third-party applications to integrate it into a standardizing biomedical dataset workflow.

References

21. Ullman JD. Information integration using logical views. In Proceedings of the 6th Int. Conf. on Database Theory (ICDT'97), Lecture Notes in Computer Science 1997; volume 1186, pp 19-40.
22. Halevy AY. Answering queries using views: A survey. Very Large Database J. 2001;10(4):270-294.
23. Calì A, Calvanese D, De Giacomo G, et al. Accessing data integration systems through conceptual schemas. Lecture Notes in Computer Science 2001; volume 2224.
24. Cavelaars M, Rousseau J, Parlayan C, et al. OpenClinica. J Clin Bioinforma. 2015;5(Suppl 1):S2.
25. Martínez-Romero M, Jonquet C, O'Connor MJ, Graybeal J, Pazos A, Musen MA. NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. *J Biomed Semantics*. 2017;8(1):21.
26. Sickle Cell Disease Ontology Working Group. The Sickle Cell Disease Ontology: enabling universal sickle cell-based knowledge representation. *Database (Oxford)*. 2019;2019:baz118.
27. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47(D1):D1018-D1027.
28. Schriml LM, Mitraka E, Munro J, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 2019;47(D1):D955-D962.
29. Mungall CJ, McMurry JA, Köhler S, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2017;45(D1):D712-D722.
30. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*. 2010;26(8):1112-1118.
31. de Coronado S, Wright LW, Fragoso G, et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009 Jun;42(3):530-539.

32. Chu L, Kannan V, Basit MA, et al. SNOMED CT Concept Hierarchies for Computable Clinical Phenotypes From Electronic Health Record Data: Comparison of Intensional Versus Extensional Value Sets. JMIR Med Inform. 2019;7(3):e14654
33. Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc., 1st Edition, ISBN-0596516495, 2009.
34. AlMaayah M, Sawalha M, Abushariah MAM. Towards an automatic extraction of synonyms for Quranic Arabic WordNet. International Journal of Speech Technology 2016;19:177-189.
35. Dongsuk O, Kwon S, Kim K, Ko Y. Word Sense Disambiguation Based on Word Similarity Calculation Using Word Vector Representation from a Knowledge-based Graph. Proceedings of the 27th International Conference on Computational Linguistics 2008; pages 2704-2714.
36. Wang X, Zuo W, Wang Y. A Novel Approach to Word Sense Disambiguation Based on Topical and Semantic Association. The Scientific World Journal 2013; 2013(Article ID 586327):8 pages.
37. Orkphol K, Yang W. Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet. Future Internet 2019;11:114. doi:10.3390/fi11050114
38. Wang Y, Wanga M, Fujitab H. Word Sense Disambiguation: A comprehensive knowledge exploitation framework. Knowledge-Based Systems 2020;190:105030.
39. Mazandu GK, Mulder NJ. DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. *BMC Bioinformatics* 2013;14:284.
40. Xu DG, Zhao PL, Yang CH, et al. A novel Minkowski-distance-based consensus clustering algorithm. International Journal of Automation and Computing 2017;14:33-44.
41. Voitsekhovskii MI. Minkowski Inequality. Hazewinkel, Michiel, Encyclopedia of Mathematics, EMS Press 2001, Springer. https://encyclopediaofmath.org/index.php?title=Minkowski_inequality
42. Mazandu GK, Chimusa ER, Rutherford K, et al. Large-scale data-driven integrative framework for extracting essential targets and processes from disease-associated gene data sets. Brief Bioinform. 2018;19(6):1141-1152.
43. Ukkenen E. Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science 1992;92:191-211.