The GA4GH Phenopacket schema: A computable representation of clinical data for precision medicine

Julius O. B. Jacobsen[0000-0002-3265-1591]1,* Michael Baudis[0000-0002-9903-4248]2,3 Gareth S. Baynam[0000-0003-4920-9553]4,5,6 Jacques S. Beckmann[0000-0002-9741-1900]7 Sergi Beltran[0000-0002-2810-3445]8,9,10 Orion J. Buske[0000-0002-9064-092X]11 Tiffany J. Callahan[0000-0002-8169-9049]12 Christopher G. Chute[0000-0001-5437-2545]13 Mélanie Courtot[0000-0002-9551-6370]14,15 Daniel Danis[0000-0003-0900-3411]16 Olivier Elemento[0000-0002-8061-9617]17 Andrea Essenwanger[0000-0002-0042-8088]18 Robert R. Freimuth[0000-0002-9673-5612]19 Michael A. Gargano[0000-0002-2157-3591]16 Tudor Groza[0000-0003-2267-8333]20 Ada Hamosh[0000-0002-1780-5230]21 Nomi L. Harris[0000-0001-6315-3707]22 Rajaram Kaliyaperumal[0000-0002-1215-167X]23 Kevin C. Kent Lloyd[0000-0002-5318-4144]24,25 Aly Khalifa[0000-0002-7084-1345]19 Peter M. Krawitz[0000-0002-3194-8625]26 Sebastian Köhler[0000-0002-5316-1399]27 Brian J. Laraway[0000-0002-0450-7074]12 Heikki Lehväslaiho[0000-0002-6263-1356]28 Leslie Matalonga[0000-0003-0807-2570]8 Julie A. McMurry[0000-0002-9353-5498]12 Alejandro Metke-Jimenez[0000-0003-1068-0938]29 Christopher J. Mungall[0000-0002-6601-2165]22 Monica C. Munoz-Torres[0000-0001-8430-6039]12 Soichi Ogishima[0000-0001-8613-2562]30 Anastasios Papakonstantinou[0000-0003-4301-3859]8 Davide Piscia[0000-0002-0468-0408]8 Nikolas Pontikos[0000-0003-1782-4711]31,32 Núria Queralt-Rosinach[0000-0003-0169-8159]23 Marco Roos[0000-0002-8691-772X]23 Julian Sass[0000-0002-2068-7765]18 Paul N. Schofield[0000-0002-5111-7263]33,34,35 Dominik Seelow[0000-0002-9746-4412]36,37 Anastasios Siapos[0000-0001-6753-6764]38 Damian Smedley[0000-0002-5836-9850]1 Lindsay D. Smith[0000-0002-0603-4178]15,39 Robin Steinhaus[0000-0001-6613-4675]36,37 Jagadish Chandrabose Sundaramurthi[0000-0002-6670-9157]16 Emilia M. Swietlik[0000-0002-4095-8489]40,41,42 Sylvia Thun[0000-0002-3346-6806]18 Nicole A. Vasilevsky[0000-0001-5208-3432]43 Alex H. Wagner[0000-0002-2502-8961]44,45 Jeremy L. Warner[0000-0002-2851-7242]46 Claus Weiland[0000-0003-0351-6523]47 Melissa A. Haendel[0000-0001-9114-8737]12,* Peter N. Robinson[0000-0002-0736-9199]16,48,*

1. Queen Mary University of London, William Harvey Research Institute, London EC1M 6BQ, UK
2. University of Zurich, Department of Molecular Life Sciences, Zürich 8057, Switzerland
3. Swiss Institute of Bioinformatics, Computational Oncogenomics Group, Zürich 8057, CH
4. King Edward Memorial Hospital, Western Australian Register of Developmental Anomalies and Genetic Services of WA, Perth 6008, AU
5. University of Western Australia, Faculty of Health and Medical Sciences, Division of Paediatrics, Perth 6008, AU
6. Telethon Kids Institute, Genetic and Rare Diseases, Perth 6008, AU
7. University of Lausanne, Faculty of Biology and Medicine, Lausanne CH-1015, Switzerland
8. CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Bioinformatics Unit, Barcelona 8028, ES
9. Universitat Pompeu Fabra (UPF), Barcelona 8005, ES
10. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona 8028, ES
11. PhenoTips, Toronto M5G 1L5, CA
12. University of Colorado Anschutz Medical Campus, Center for Health AI, Aurora 80045, CO, USA
13. Johns Hopkins University, Schools of Medicine, Public Health, and Nursing, Baltimore 21287, MD, USA
14. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK
15. Ontario Institute for Cancer Research, Adaptive Oncology, Toronto M5G0A3, CA
16. The Jackson Laboratory, Genomic Medicine, Farmington 6032, CT, USA
17. Weill Cornell Medicine, Caryl and Israel Englander Institute for Precision Medicine, New York 10021, NY, USA
18. Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Core Facility Digital Medicine and Interoperability, Berlin 10178, DE

19. Mayo Clinic, Department of Artificial Intelligence and Informatics, Rochester 55905, MN, USA
20. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge CB10 1SD, UK
21. Johns Hopkins University, Department of Genetic Medicine, Baltimore 21287, MD, USA
22. Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology, Berkeley 94720, CA, USA
23. Leiden University Medical Center, Human Genetics, Leiden 2333 ZA, NL
24. UC Davis, Mouse Biology Program, Davis 95618, CA, USA
25. UC Davis School of Medicine, Department of Surgery, Sacramento 95817, CA, USA
26. University Hospital Bonn, Bonn, Germany, Institute for Genomic Statistics and Bioinformatics, Bonn 53113, DE
27. Ada Health GmbH, Berlin 10178, DE
28. CSC – IT Center for Science, Sensitive Data Services, Espoo FI-02101, FI
29. CSIRO, The Australian e-Health Research Centre, Herston 4029, AU
30. Tohoku University, INGEM, Sendai 980-8573, JP
31. University College London, Institute of Ophthalmology, London EC1V 9EL, UK
32. Moorfields Eye Hospital, Genetics Service, London EC1V 2PD, UK
33. University of Cambridge, Dept of Physiology, Development and Neuroscience, Cambridge CB2 3EG, UK
34. The Jackson Laboratory, Mammalian Genetics, Bar Harbor ME 04609, ME, USA
35. The Alan Turing Institute, London NW1 2DB, UK
36. Bioinformatics and Translational Genetics, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin 10178, DE
37. Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Genetics and Human Genetics, Berlin 13353, DE
38. Lifebit Biotech Ltd., London W86BD, UK
39. Global Alliance for Genomics and Health, N/A, Toronto M5G0A3, CA
40. University of Cambridge, Medicine Department, Cambridge CB2 0QQ, UK
41. Addenbrooke's Hospital, Respiratory Medicine Department, Cambridge CB2 0QQ, UK
42. Royal Papworth Hospital, Cambridge Centre for Lung Infection, Cambridge CB2 0AY, UK
43. University of Colorado, Anschutz Medical Campus, Center for Health AI, Aurora 80045, CO, USA
44. Nationwide Children's Hospital, The Steve and Cindy Rasmussen Institute for Genomic Medicine, Columbus 43215, OH, USA
45. The Ohio State University College of Medicine, Departments of Pediatrics and Biomedical Informatics, Columbus 43215, OH, USA
46. Vanderbilt University, Departments of Medicine and Biomedical Informatics, Nashville 37235, TN, USA
47. Senckenberg - Leibniz Institution for Biodiversity and Earth System Research, Data and Modelling Centre, Frankfurt/Main 60325, DE
48. University of Connecticut, Institute for Systems Genomics, Farmington 6032, CT, USA

Consortial authors

Myles Axton[0000-0002-8042-4131]1 Lawrence Babb[0000-0002-2455-2227]2 Cornelius F. Boerkoel[0000-0003-3097-241X]3 Bimal P. Chaudhari[0000-0002-0115-949X]4,5 Hui-Lin Chin[0000-0001-7431-6794]6,7 Michel Dumontier[0000-0003-4727-9435]8 David P. Hansen[0000-0002-2998-4563]9 Harry Hochheiser[0000-0001-8793-9982]10 Veronica A. Kinsler[0000-0001-6256-327X]11,12 Hanns Lochmüller[0000-0003-2324-8001]13,14,15 Alexander R. Mankovich[0000-0002-1258-4184]16 Gary I. Saunders[0000-0002-7468-0008]17 Panagiotis I. Sergouniotis[0000-0003-0986-4123]18 Rachel Thompson[0000-0002-6889-0121]13 Andreas Zankl[0000-0001-8612-1062]19,20,21

1. Wiley, Inc, Research, Hoboken 7030, NJ, USA

2. Broad Institute of MIT and Harvard, Cambridge 2142, MA, USA
3. University of British Columbia, Medical Genetics, Vancouver V6H3N1, CA
4. Nationwide Children's Hospital, The Steve and Cindy Rasmussen Institute for Genomic Medicine, Divisions of Neonatology, Genetics and Genomic Medicine, Columbus 43215, OH, USA
5. The Ohio State University College of Medicine, Department of Pediatrics, Columbus 43210, OH, USA
6. Khoo Teck Puat-National University Children's Medical Institute, National University Hospital, Department of Paediatrics, Singapore 119074, Singapore
7. Women's Hospital of British Columbia, Provincial Medical Genetics, Vancouver V6H3N1, CA
8. Maastricht University, Institute of Data Science, Maastricht 6229 EN, NL
9. CSIRO, Australian e-Health Research Centre, Brisbane 4027, AU
10. University of Pittsburgh, Biomedical Informatics, Pittsburgh 15206, PA, USA
11. Great Ormond St Hospital for Children, Paediatric Dermatology, London WC1N 3JH, UK
12. Francis Crick Institute, Mosaicism and Precision Medicine Laboratory, London NW1 1AT, UK
13. Children's Hospital of Eastern Ontario Research Institute, Molecular Biomedicine, Ottawa K1H 8L1, CA
14. University of Ottawa, Brain and Mind Research Institute, Department of Cellular and Molecular Medicine, Ottawa K1H 8M5, CA
15. The Ottawa Hospital, Neuromuscular Centre, Ottawa K1Y 4E9, CA
16. Philips Research North America, Precision Diagnosis & Image-Guided Therapy, Cambridge 2141, MA, USA
17. ELIXIR, ELIXIR Hub, Cambridge CB10 1SD, UK
18. University of Manchester, Division of Evolution, Infection and Genomics, Manchester M13 9PT, UK
19. The University of Sydney, Faculty of Medicine and Health, Sydney 2006, AU
20. The Children's Hospital at Westmead, Department of Clinical Genetics, Westmead 2145, AU
21. Garvan Institute of Medical Research, Kinghorn Centre for Clinical Genomics and Bone Division, Darlinghurst 2010, AU

To the editor. Despite great strides in the development and wide acceptance of standards for exchanging structured information about genomic variants, the development of standards for computational phenotype analysis for translational genomics has lagged behind. Phenotypic features (signs, symptoms, laboratory and imaging findings, results of physiological tests, etc.) are of essential clinical importance, yet exchanging them in conjunction with genomic variation is often overlooked or even neglected. In the clinical domain, significant work has been dedicated to the development of computational phenotypes.[1] Traditionally, these approaches have largely relied on rule-based methods and large sources of clinical data to identify cohorts of patients with or without a specific disease.[2–5] However, they were not developed to enable deep phenotyping of phenotypic abnormalities, to facilitate computational analysis of interpatient phenotypic similarity, or to support computational decision support. To address this, the Global Alliance for Genomics and Health[6] (GA4GH) has developed the Phenopacket schema, which supports exchange of computable longitudinal case-level phenotypic information for diagnosis of and research on all types of disease including Mendelian and complex genetic diseases, cancer, and infectious diseases (Fig 1).

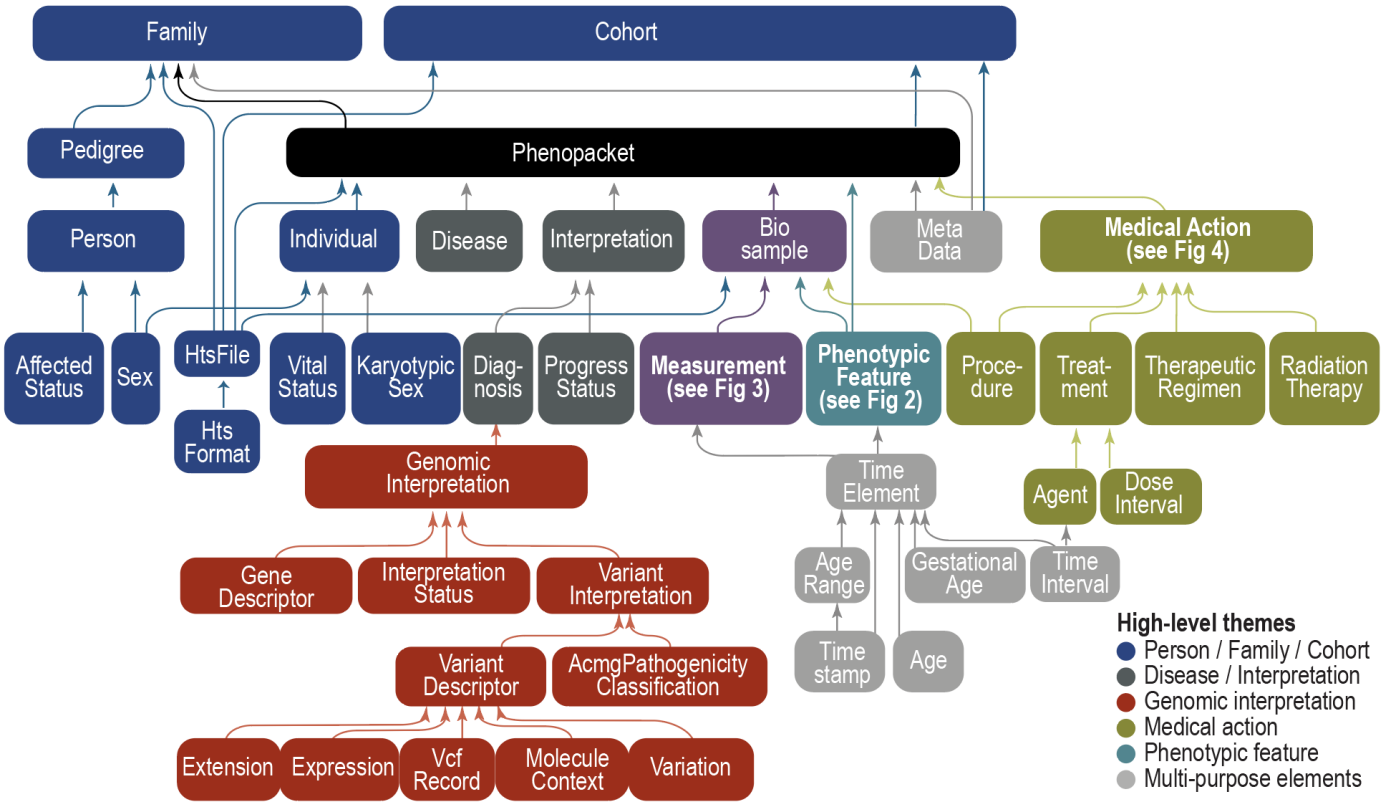


Figure 1. Phenopacket schema overview. *The GA4GH Phenopacket schema consists of several optional elements, each of which contains information about a certain topic such as phenotype, variant, pedigree, etc. An element can contain other elements, which allows a hierarchical representation of data. For instance,* `Phenopacket` *contains elements of type* `Individual`, `PhenotypicFeature`, `Biosample`, *and so on. Individual elements can therefore be regarded as building blocks that are combined to create larger structures. Colors represent the major themes of elements within the schema.*

The *PhenotypicFeature* is the central element of the Phenopacket schema. A *PhenotypicFeature* can be used to describe any phenotypic characteristic (often, but not necessarily, clinical abnormalities) including signs and symptoms, laboratory findings, histopathology findings, imaging, electrophysiological results, etc., along with modifier and qualifier concepts. Each phenotypic feature is described using an ontology term. While the Phenopacket schema does not mandate which ontology to use, it provides recommendations, such as the Human Phenotype Ontology[7] (HPO) for rare diseases and the National Cancer Institute Thesaurus (NCIT) for transmission of information about a cancer specimen, e.g., pathological staging or more detailed information about histology or tumor markers.[8] One can indicate whether an abnormality was excluded during the diagnostic process (e.g., whether a morphological cardiac defect was excluded by echocardiography), or use other optional HPO terms to denote the severity, frequency (e.g., number of occurrences of seizures per week), laterality (e.g., unilateral), or other pattern of a phenotypic feature in the patient being described. Finally, the onset (and if applicable the resolution) of specific features can be indicated. Other key elements are *Measurement*, which is used to capture quantitative (i.e., numerical), ordinal (e.g., absent/present), or categorical measurements; *Biosample*, a description of biological material obtained from the individual represented in the Phenopacket and used for phenotypic, genotypic, or other -omics analysis; and *MedicalAction*, which includes a hierarchical representation of medical actions including medications, procedures, and other actions taken for clinical management. The *Treatment* element is a subelement of *MedicalAction* and represents administration of a pharmaceutical agent, broadly defined as prescription and over-the-counter medicines, vaccines, and other therapeutic agents such as monoclonal antibodies or CAR T-cell-therapy.

The *Interpretation* elements specify interpretations of genomic findings. This element leverages complementary resources developed by the GA4GH Genomic Knowledge Standards Work Stream: the Variation Representation Specification (VRS) and VRS Added Tools for Interoperable Loquacious Exchange (VRSATILE; see Web Resources).[6] Further information on this and other elements is available in the online documentation.

The Phenopacket schema was designed to support a number of use cases. Many of these use cases have been successfully implemented and tested in the community, particularly in the field of rare disease diagnostics and biobanking, while others, such as EHR integration, are in the process of being implemented (Supplemental Table 1).

The Phenopacket schema (version 2.0) was formally reviewed and approved as a GA4GH standard[6] in 2021. It is designed to be interoperable with other relevant standards, including the traditional PED file as well as the GA4GH pedigree standard, the GA4GH Beacon,[9] and the GA4GH Variation Representation Specification. The GA4GH has committed to coordinate its activities and future roadmaps with those of other standards development organizations (SDOs), including International Organization for Standardization (ISO) Technical Subcommittee for Genomics Informatics (ISO/TC215/SC1) and HL7 Clinical Genomics (CG). Consequently, a Fast Interoperable Healthcare Resources (FHIR) implementation guide for Phenopacket interoperability is being developed and the Phenopacket schema is in the process of ISO certification (Supplemental Table 2).

The VCF standard for storing genotyping data allowed a wide range of research groups to write software for analyzing such data.[10] The GA4GH Phenopacket schema aspires to be similarly transformative in the landscape of genome analysis using phenotype data. Multiple providers of phenotypic data include patients and clinicians, via a variety of mechanisms including clinical notes and EHR records, interfaces such as FHIR, app-based entry, and mobile devices. The Phenopacket schema acts as a common model that can capture data from many sources with a unified software representation and in turn can be used by multiple receivers of the phenotypic information, including journals, databases, registries, clinical laboratories. Phenopackets can support diverse users and use cases, including patient matchmaking services, diagnostics, and cohort identification. Software has become an essential resource for genomic medicine. We anticipate that the Phenopacket schema will encourage the development of a collection of software for the analysis of genomic data in the context of clinical information that will accelerate innovation and discovery. Genomic data will become ever more important in translational research and clinical care in the coming years and decades. The Phenopacket schema represents a standard for capturing clinical data and integrating it with genomic data that will help to obtain the maximal utility of this data for understanding disease and developing precision medicine approaches to therapy.

Core Phenopacket resources - Software availability:

Phenopacket schema source code: https://github.com/phenopackets/phenopacket-schema

Phenopacket schema documentation: https://phenopacket-schema.readthedocs.io/

Phenopacket tools: https://github.com/phenopackets/phenopacket-tools

Related Standards - Web Resources:

GA4GH Beacon project: https://beacon-project.io/

GA4GH Phenopacket FHIR implementation guide: https://github.com/phenopackets/core-ig

GA4GH Pedigree standard: https://github.com/GA4GH-Pedigree-Standard/pedigree

GA4GH Variation Representation Specification (VRS): vrs.ga4gh.org

VRS Added Tools for Interoperable Loquacious Exchange (VRSATILE): vrsatile.readthedocs.io

Phenopacket RDF model: https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema/wiki

Genomics Informatics — Phenopackets: A Format for Phenotypic Data Exchange (ISO): https://www.iso.org/standard/79991.html

Contact: j.jacobsen@qmul.ac.uk; melissa@tislab.org; peter.robinson@jax.org

Acknowledgements

The authors gratefully acknowledge insight and feedback from Marian H. Adly, Pier Luigi Buttigieg, Nour Gazzaz, Janine Lewis, Manuel Posada de la Paz and Maria Taboada.

Funding

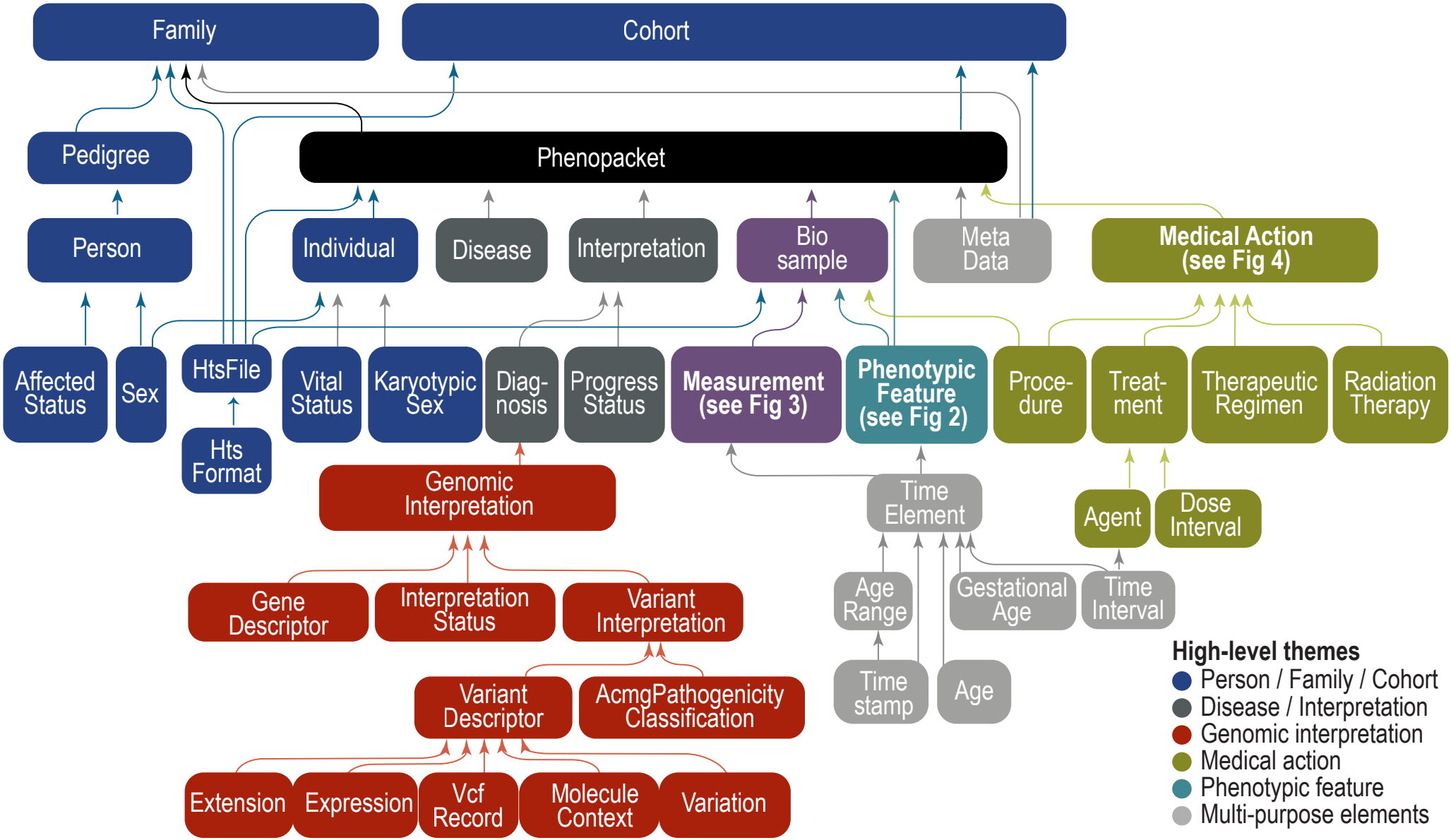
This work was supported by 7RM1HG010860-02 (NHGRI). Additional funding was as follows. PNR was supported by NLM contract #75N97019P00280, NIH NHGRI RM1HG010860, NIH OD R24OD011883, NIH NICHD 1R01HD103805-01. HH was supported by NIH OD R24OD011883. GIS was supported by ELIXIR, the research infrastructure for life-science data. CGC was supported by NIH NCATS U24TR002306. KCL was supported by NIH OD 5UM1OD023221. MB was supported by BioMedIT Network project of Swiss Institute of Bioinformatics (SIB) and Swiss Personalized Health Network (SPHN). AHW was supported by NIH NHGRI K99HG010157, NIH NHGRI R00HG010157. CJM, MAH, MCM-T, JAM, DD were supported by NIH NHGRI RM1HG010860, NIH OD R24OD011883. AM-J was supported by Australian Genomics. Australian Genomics is supported by the National Health and Medical Research Council (GNT1113531). DS, JOBJ were supported by NIH NHGRI RM1HG010860, NIH OD R24OD011883, NIH NICHD 1R01HD103805-01. MD was supported by NIH NHGRI U54HG004028, NIH NHGRI 5U01HG008473-03, NIH NCATS OT2TR003434-01S1U54HG008033-01. GSB was supported by Roy Hill Community Foundation, Angela Wright Bennett Foundation, McCusker Charitable Foundation, Borlaug Foundation, Stan Perron Charitable Foundation. LB was supported by NIH NHGRI U41HG006834 (Clinical Genome Resource). MC was supported by EMBL-EBI Core Funds and Wellcome Trust GA4GH award number 201535/Z/16/Z. AH was supported by NIH NHGRI 1U41HG006627, NIH NHGRI 1U54HG006542, NIH NHGRI 1RM1HG010860. PNS was supported by The Alan Turing Trust. NLH was supported by NIH NHGRI RM1HG010860, NIH OD R24OD011883, U.S. Department of Energy Contract DE-AC02-05CH11231. NP was supported by Moorfields Eye Charity. NQ-R was supported by EU Horizon 2020 research and innovation programme grant agreement 825575 (EJP-RD). OE was supported by NIH grants UL1TR002384, R01CA194547, P01CA214274 LLS SCOR grants 180078-01, 7021-20, Starr Cancer Consortium Grant I11-0027. HL was supported by CIHR Foundation Grant on Precision Health for Neuromuscular Diseases FDN-167281. RT was supported by CIHR postdoctoral fellowship award MFE-171275. LDS was supported by Genome Canada and NIH NHGRI U24HG011025. SO was supported by AMED. DP, LM, AP, SB, MR, RK were supported by EU Horizon 2020 research and innovation programme grant agreements 779257 (Solve-RD) and 825575 (EJP-RD). RRF was supported by NLM contract #75N97019P00280.

Conflicts of interest

SK is an employee of Ada Health GmbH. NP is a director of Phenopolis Ltd. OE is supported by Janssen, Johnson and Johnson, Volastra Therapeutics, AstraZeneca and Eli Lilly research grants. He is scientific advisor and equity holder in Freenome, Owkin, Volastra Therapeutics and One Three Biotech. ARM is an employee of Philips Research North America. OJB is an employee of PhenoTips. MA is an editor employed by Wiley. AS is an employee of Lifebit Biotech Ltd.

References

1. Richesson, R. & Smerek, M. Electronic health records-based phenotyping. *Rethinking clinical trials: A living textbook of pragmatic clinical trials* **2016**, (2014).
2. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* **20**, 117–121 (2013).
3. Shivade, C. *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* **21**, 221–230 (2014).
4. Wei, W.-Q. & Denny, J. C. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med.* **7**, 41 (2015).
5. Richesson, R. L., Sun, J., Pathak, J., Kho, A. N. & Denny, J. C. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif. Intell. Med.* **71**, 57–61 (2016).
6. Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom* **1**, (2021).
7. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
8. Sioutos, N. *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* **40**, 30–43 (2007).
9. Fiume, M. *et al.* Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* **37**, 220–224 (2019).
10. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).


Family
Cohort
Pedigree
Phenopacket
Person
Individual
Disease
Interpretation
Bio sample
Meta Data
Medical Action (see Fig 4)
Affected Status
Sex
HtsFile
Vital Status
Karyotypic Sex
Diag-nosis
Progress Status
Measurement (see Fig 3)
Phenotypic Feature (see Fig 2)
Proce-dure
Treat-ment
Therapeutic Regimen
Radiation Therapy
Hts Format
Genomic Interpretation
Time Element
Agent
Dose Interval
Gene Descriptor
Interpretation Status
Variant Interpretation
Age Range
Gestational Age
Time Interval
Variant Descriptor
AcmgPathogenicity Classification
Time stamp
Age
Extension
Expression
Vcf Record
Molecule Context
Variation
High-level themes
Person / Family / Cohort
Disease / Interpretation
Genomic interpretation
Medical action
Phenotypic feature
Multi-purpose elements