Data-driven method to enhance craniofacial and oral phenotype vocabularies

Rashmi Mishra, BDS, MPH; Andrea Burke, DMD, MD; Bonnie Gitman, DMD, MD; Payal Verma, DMD, MPH; Mark Engelstad, DDS, MD; Melissa A. Haendel, PhD; Ilias Alevizos, DMD, MMSc; William A. Gahl, MD, PhD; Michael T. Collins, MD; Janice S. Lee, DDS, MD; Murat Sincan, MD

Supplemental material is available online.

ABSTRACT

Background. A significant amount of clinical information captured as free-text narratives could be better used for several applications, such as clinical decision support, ontology development, evidence-based practice, and research. The Human Phenotype Ontology (HPO) is specifically used for semantic comparisons for diagnostic purposes. All these functions require quality coverage of the domain of interest. The authors used natural language processing to capture craniofacial and oral phenotype signatures from electronic health records and then used these signatures for evaluation of existing oral phenotype ontology coverage.

Methods. The authors applied a text-processing pipeline based on the clinical Text Analysis and Knowledge Extraction System to annotate the clinical notes with Unified Medical Language System codes. The authors extracted the disease or disorder phenotype terms, which were then compared with HPO terms and their synonyms.

Results. The authors retrieved 2,153 deidentified clinical notes from 558 patients. Finally, 2,416 unique diseases or disorders phenotype terms were extracted, which included 210 craniofacial or oral phenotype terms. Twenty-six of these phenotypes were not found in the HPO.

Conclusions. The authors demonstrated that natural language processing tools could extract relevant phenotype terms from clinical narratives, which could help identify gaps in existing ontologies and enhance craniofacial and dental phenotyping vocabularies.

Practical Implications. The expansion of terms in the dental, oral, and craniofacial domains in the HPO is particularly important as the dental community moves toward electronic health records.

Key Words. Natural language processing; evidence-based dentistry; craniofacial and oral phenotypes; ontology.

JADA 2019:150(11):933-939
https://doi.org/10.1016/j.adaj.2019.05.029

There has been an explosion of available data as the medical field has moved from paper-based to electronic health records (EHRs), along with other big data sources such as digital imaging, genomics, proteomics, and metabolomics. To make this vast amount of data clinically useful and to support in-depth analysis to understand the molecular basis of diseases, methods and tools are required that accurately integrate and link -omics data with clinical information.[1]

A phenotype is defined as the morphologic, physiological, and behavioral characteristics of a person whereas a genotype is a person's entire genetic makeup.[2-4] A phenotype terminology is a catalog of specific signs, symptoms, imaging findings, and other abnormalities seen in clinical practice. A concept in ontology denotes a single meaning, including all variations from any source that express that same meaning. Each of these concepts is assigned at least 1 semantic type and unique Concept Unique Identifier (CUI), which represents that single meaning.[2] Individual clinical concepts (for example "Class II malocclusion") are referred to as terms. Ontologies provide helpful definitions of these terms, as well as the relationships between them. When compiled, these terms form a foundation for computational searching and analysis and provide the ability to create inferences about information. For example, Crouzon syndrome is a rare craniofacial abnormality whose phenotype includes craniosynostosis, prominent forehead, curved nose, midface hypoplasia, short upper lip, mandibular prognathism with Class 3 malocclusion, crowded teeth, and anterior open bite. However, patients with Crouzon syndrome may exhibit a wide spectrum of these features. To better understand the various subphenotypes in craniofacial conditions, precise and detailed disease-phenotype relationships found in ontologies are needed.

One approach to capturing phenotypic information is the use of biomedical ontologies that are consensus-based vocabularies. Gruber[5] defines ontology as "the specification of conceptualizations, used to help programs and humans share knowledge." The highly structured vocabulary and relationships within ontologies help optimize the exchange of information in and across domains without an uncontrolled explosion due to excess information. The Human Phenotype Ontology (HPO) facilitates the study of disease-phenotype relationships and is developed on the basis of the medical literature. The HPO is a part of the Monarch Initiative, a National Institutes of Health (NIH)—supported international consortium dedicated to semantic integration of biomedical and model organism data with the ultimate goal of improving biomedical research.[6-8] Human phenotype ontologies are especially important in cases of rare and undiagnosed diseases, when clinical experts may use different terms to describe similar clinical phenotypes.[9]

The construction of ontologies requires content expertise and continuous accrual of new knowledge, which is challenging to incorporate in a timely manner. The comprehensiveness of ontology in a given domain is crucial for its usefulness. Thus, there is a need for systems that can effectively and efficiently provide an assessment of domain coverage and potential new content for inclusion. One of these mechanisms includes exploiting the information found in biomedical resources such as the published literature and EHRs. EHRs have several advantages for use in phenotyping, such as cost efficiency and the availability of large amounts of clinical and temporal information.[10] Although structured data, such as *International Classification of Diseases* (ICD) codes,[11] are useful controlled vocabularies, they are limited in detail as they are designed to facilitate medical billing. In contrast, unstructured clinical data (for example, consultation notes, history and physical notes, discharge summaries, and pathology reports) are rich in clinical detail of patient conditions including diagnosis, signs and symptoms, family history, onset, severity, and other clinical findings. Unfortunately, the unstructured nature of this free-text information makes it challenging to extract and interpret the phenotypic information within it. Natural language processing (NLP) can support analysis of data derived from EHRs by converting clinical and genetic information from unstructured text into a computer-accessible form.[12]

In this study, we used existing NLP tools to capture the various terms used to describe clinical phenotypes in the craniofacial, dental, and oral domains. This information was further used to evaluate the completeness of existing phenotype ontologies. We used an open source NLP tool known as the clinical Text Analysis and Knowledge Extraction System (cTAKES), an application first developed at the Mayo Clinic.[13] Our aim was to extract craniofacial, dental, and oral phenotype terms from EHRs at the NIH Clinical Center (CC) and to identify terms that may be missing from the HPO without any feature engineering or retraining of machine learning models. Such successful extraction will enhance the coverage of craniofacial, dental, and oral terms in the HPO and demonstrate the use of a data-driven methodology using an NLP tool within a diverse clinical source.

METHODS

Our methodology included data collection, NLP, identifying craniofacial and dental phenotypes, and comparison with the HPO.

The figure shows the work flow of the text-mining pipeline that extracts the phenotype terms from clinical documents.

Data set

Research participants who are admitted to the NIH CC have their clinical documentation archived in the NIH CC data warehouse system known as the Biomedical Translational Research Information System (BTRIS).[14] BTRIS incorporates data from the CC EHRs and research systems from

ABBREVIATION KEY

BTRIS:	Biomedical Translational Research Information System.
cTAKES:	Clinical Text Analysis and Knowledge Extraction System.
CUI:	Concept Unique Identifier.
EHR:	Electronic health record.
FN:	False-negative result.
FP:	False-positive result.
HPO:	Human Phenotype Ontology.
ICD:	*International Classification of Diseases.*
NIH:	National Institutes of Health.
NIH CC:	National Institutes of Health Clinical Center.
NLP:	Natural language processing.
SNOMED-CT:	Systematized Nomenclature of Medicine-Clinical Terms.
TP:	True-positive result.
UMLS:	Unified Medical Language System.

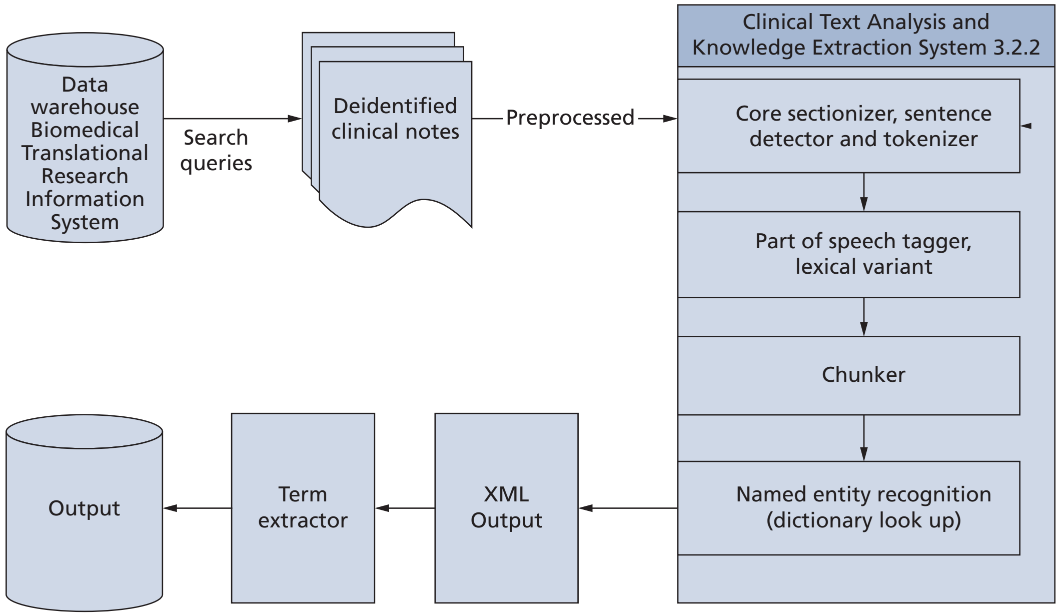
Data warehouse Biomedical Translational Research Information System
Search queries
Deidentified clinical notes
Preprocessed
Clinical Text Analysis and Knowledge Extraction System 3.2.2
Core sectionizer, sentence detector and tokenizer
Part of speech tagger, lexical variant
Chunker
Named entity recognition (dictionary look up)
XML Output
Term extractor
Output

Figure. Text-mining pipeline.

Table 1. Examples of search criteria for craniofacial disorders.

TERM	ICD-9 CODE*
Abnormalities, Craniofacial	376.44
Craniofacial Abnormality	351.8
Craniofacial and Skeletal Abnormality	351.9
Craniomandibular Disorder	438.83
Craniomandibular Diseases	438.83

* ICD: *International Classification of Diseases*.[11]

the 27 NIH institutes and centers. This includes data on common and rare conditions that are routinely studied at the NIH. Of the 27 institutes and centers, clinical data are collected from only 18 institutes that have ongoing human subject protocols. These data include structured as well as unstructured clinical texts available to researchers in a deidentified, secure, and controlled manner. As a first step, a query of BTRIS was performed to search data from 2007 through 2015. The search criteria were based on clinical terms related to the dental, oral, and craniofacial region (for example, hemifacial microsomia) and ICD-9 codes (for example, 756.0) (Table 1).[11] The full search criteria are presented in eTable 1 available online at the end of this article.

The search was done against the BTRIS database, which contains 30 million clinical encounter notes. Table 2 lists the distribution of the types of notes including first registration reports (first admission summary for all new human participants at the NIH CC), dental consultations, all other consultations (for example, dermatology, genetics, pediatrics), discharge summaries, and outpatient notes.

Text-mining pipeline

Establishment of a Reference Standard

To capture oral and craniofacial terms, a reference standard was created on the basis of 200 documents randomly selected from the 2,329-note data set. From this corpus, 112 sentences were selected and independently annotated by 2 clinicians (B.G., R.M.) for disease or disorder terms. In the first round, with no annotation guidelines, an interrater agreement (linear weighted κ) of 0.62 was achieved.[15] Disagreements were reconciled through consensus, and the annotation schema was refined. In the second round, the same 2 clinicians rated the documents independently, using the established guidelines (linear weighted κ, 0.82). An example is given in eTable 2 (available online

Table 2. Distribution of notes.

TYPE OF NOTE	COUNT, NO.
First Registration Reports	254
Dental Consultations	521
Other Consultations	1,108
Discharge Summaries/Outpatient Notes	446

at the end of this article). Given the high interrater reliability of the annotation, 1 clinician (R.M.) extracted the remaining terms from the corpus.

Note Parsing

We installed Oracle VirtualBox 4.3.10r93012 to host a virtual machine with an Ubuntu 14 operating system (includes the open source NLP tool cTAKES 3.2.2). Our application was based on various components adapted from the existing cTAKES and scripts written in Python 3.5.1. Owing to the different components associated with the tool, we explored different analysis engines for various categories of information extraction, and we selected the aggregate plain text Unified Medical Language System (UMLS) processor for our final output. UMLS, developed by the National Library of Medicine, lists comprehensive biomedical and health care terms for developing computer systems capable of understanding the specialized vocabulary.[2] The clinical notes were analyzed by preprocessing them into cTAKES input format with Python scripts. Modules such as sentence boundary detector, tokenizer, normalizer, part-of-speech tagger, shallow parser, and named entity were used in our pipeline. The text analysis continued with dictionary-based lookup, using the UMLS Metathesaurus module, to detect disease or disorder terms. Detection was performed at the token level first and then expanded to match at the noun phrase level. All detected concepts were extracted along with their CUIs, preferred terms, as well as coding schema, using a Python script. Finally, a custom script was used to analyze and filter all unique disease or disorder terms by CUI (see eBox 1 available online at the end of this article for more information).

Data-Driven Concept Screening and Comparison With HPO

A team of clinical experts (A.B., B.G., M.E., J.S.L.) further evaluated the disease or disorder phenotype terms obtained by processing the entire corpus. They screened the craniofacial and dental phenotypes from the list, and these phenotype terms were then manually compared with the HPO database (November 2016 version).

RESULTS

We retrieved 2,329 deidentified notes from 558 patients, based on our search criteria. After running our pipeline on the raw deidentified free text clinical notes, we extracted 2,416 disease or disorder terms. An example of raw input that was converted into output after running cTAKES and our customized Python script is presented in Table 3. The application output was then compared with the reference standard and the extracted information was labeled as true-positive result, false-positive result, true-negative result, or false-negative result. To test the accuracy of our method, the output was used to compute the 3 outcome measures: precision, recall, and *F* score. As shown in Table 4, we achieved an overall *F* score of 0.81 with high precision and moderate recall.

Of the 2,416 disease or disorder terms, 210 terms aggregated with the craniofacial and dental domain. Comparison with the HPO (November 2016 version) helped us find 26 terms that were not present in the database. Overall, we found that 11% of the extracted craniofacial and dental phenotype terms were absent from the HPO. The final list of missing terms can be found in eBox 2.

DISCUSSION

Use of phenotype sets for clinical research depends on many factors. First, the presence of all relevant and necessary terms in an ontology is of utmost importance. Such terms should also map to the medical jargon and terminology found in clinical notes. Second, the presence of a model organism phenotype that is compatible and mapped to the HPO is necessary to facilitate effective matching based on the phenotype similarities between species. This transspecies mapping is critical

Table 3. Example of processed cTAKES[*] output extracted from free text clinical notes.

***Raw Input*:** FIRST_NAME i=120] who comes for an evaluation and possible treatment options for his **Overbite**. There is an **epidermal nevus** of the right face extending onto the neck.
***Processed Output*:**

NORMALIZED FORM	UMLS CUI[†]	PREFERRED TEXT	CODING SCHEMA	CODE
Overbite	C0266063	Deep overbite	SNOMED-CT[‡]	60476005
Epidermal	C0334082	Epidermal nevus	SNOMED-CT	25201003
Nevus	C0334082	Epidermal nevus	SNOMED-CT	25201003

* cTAKES: Clinical Text Analysis and Knowledge Extraction System. † UMLS CUI: Unified Medical Language System Concept Unique Identifier. ‡ SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms.

Table 4. Result of the dictionary-based natural language processing pipeline.

PERFORMANCE MEASURE	RESULTS
Precision = TP[*]/(TP + FP[†])	0.91
Recall = TP/(TP + FN[‡])	0.74
***F* Measure = 2 × (Precision × Recall)/(Precision + Recall)**	0.81

* TP: True-positive result. † FP: False-positive result. ‡ FN: False-negative result.

to further research in developmental variation, evolutionary modifications, and genetic etiology of normal and diseased development. In addition, the presence of searchable central databases or registries that contain deidentified but coded phenotypes of patients should support semantic matching of similar cases.[16-18] This is particularly important in cases of rare or undiagnosed diseases, when only a handful of people around the world may have a similar constellation of features and symptoms. Finally, scientific publications should be annotated with explicit phenotype terms to facilitate identification of phenotypes and information for clinicians and researchers.[16] The power to identify similar phenotypes in patients, in addition to linking to comprehensive biomedical databases that use extensive ontologies and transspecies phenotype alignment, offers the ability to coordinate and identify phenotype-gene relationships and gene-gene and gene—small molecule relationships in a systems biology context to advance the understanding of disease mechanisms.

The results from our study showed that existing free-text electronic medical records can be mined by the cTAKES tool and used to aid in ontology development and maintenance. The precision and recall are comparable with the extraction of phenotype terms by a domain expert (in our case, dental and craniofacial professionals). Reasonable accuracy was achieved in extracting various disease or disorder terms, using our method for detecting disease or disorder phenotype. The data-driven concept screening reduced the number of phenotypes from hundreds to a few dozen, which made the manual evaluation feasible and beneficial for comparing with the existing ontology. Thus, NLP is an excellent tool that ensures consistent, efficient, and timely update of existing ontologies through incorporation of new terms.

Overall, we found 26 craniofacial and dental terms that were not a part of the HPO. The terms compared with the HPO included only the terms that were present in UMLS and not in the HPO. In an evidence-based manner, we showed that there are terms that physicians and craniofacial experts regularly use for the dental, oral, and craniofacial region that are absent in the current version of ontologies. The absence of these terms from current ontologies may limit the use of any ontology. For example, 1 of the terms missing in the HPO is "*posterior crossbite*." A posterior crossbite is a dental malocclusion indicating an abnormal buccolingual relationship of the teeth. It is a deviation from ideal occlusion in the transverse plane of space in the posterior segments. This can occur between a single posterior tooth or a group of teeth, unilaterally or bilaterally. This anomaly is often a part of the phenotypic description of patients with developmental or congenital conditions, such as cleft palate, narrow palate, asymmetric growth of the maxilla or the mandible, as well as some pathologic conditions, such as acromegaly, muscular dystrophy, condylar hypoplasia or hyperplasia, and osteochondroma. It is not self-correcting and requires early detection and intervention as it can worsen over time. Here, we have shown that the HPO may be incomplete in the important subdomains of craniofacial, dental, and oral phenotypes, which can hinder advances in craniofacial and oral disease identification and treatment. An updated HPO ontology may also help in translational research on craniofacial disorders. For example, a clinician may describe a craniofacial phenotype as "narrow maxilla" with similarities to the phenotype of a knockout mouse. The model organism will have a specific gene knocked out and the gene in question may be part of a recognized pathway. This relationship may not be found through simple text searches but may become evident via graph-based searches using ontological relationships.

We were able to achieve reasonable accuracy in extracting various disease or disorder terms by the dictionary look-up method. The precision rate was high (91%) with a moderate recall rate (74%). The moderate recall rate may be due to limitations in abbreviation extraction with both false-positive and false-negative cases. For example, "*bacterial occ*," which is a regularly used abbreviation for occasional amounts of bacteria in the sample (urine or mucus), was extracted as "osteochondritis dissecans." This clearly points to the necessity of integrating abbreviation recognition and disambiguation components into clinical NLP systems to improve our performance in future work. Also, some of the lexical and spelling variations, and misspelling in the clinical notes in our data set, accounted for a moderate recall. One major limitation of this study was the relatively small data set related to craniofacial disease, resulting in fewer clinical notes in our domain of interest. As the craniofacial research team and program at the NIH is being developed, we will be able to test this method in a larger database. The validation of the algorithm in various patient populations or EHR systems from other institutions will be necessary for reproducibility and further interrogation of the method. Finally, manual screening of the concepts was a time-consuming process; however, we are exploring other methodologies that may automate the screening procedure in the future.

Extracting phenotype terms through NLP is restricted particularly if the descriptors are contained in phrases using polysemous words such as "*restriction in mouth opening*." Precise boundary identification is needed to enhance the concept identification that will improve the sensitivity and specificity of the task. Another important area in which NLP could improve the phenotype ontologies is through extracting temporal relations from EHRs. For example, "*alopecia totalis*" does not impart full information as age at onset is an integral component of the disease description that further delineates the importance of the finding (early versus late alopecia totalis may align with a rare condition with natural aging progression, respectively). As such, the treatment plan would be completely different for alopecia totalis at age 10 years than at age 60 years. Therefore, for more complete phenotype description, the methodology must be able to capture other clinical information such as phenotype severity, age at onset, and progression over time. Thus, integration of this information can advance the understanding of disease mechanisms and accelerate development of target therapies.

CONCLUSIONS

As the use of electronic clinical notes in dentistry increases, extracting relevant information from those records by NLP could be beneficial in dental research. As dentistry moves toward precision medicine, we need ontologies to capture comprehensive relationships between genes and diseases. Our data-driven method using NLP can enhance current biomedical databases to ensure consistent updating of existing ontologies such as the HPO by incorporating new terms in a timely and efficient manner. This research will result in the improvement of HPO coverage of dental, oral, and craniofacial terms that are used by clinicians and researchers in an evidence-based manner. Our research discovered phenotype terms using NLP and BTRIS data that were not included in the HPO, and this identification method can be applied widely by other researchers in their domain. An important application would be the successful use and adoption of these NLP tools to automate the phenotyping algorithms. Future work includes use of more heterogeneous data and exploration of advanced techniques such as long parsing with deeper contextual information extraction to encourage the development of robust clinical phenotype extraction from EHRs. ■

SUPPLEMENTAL DATA

Supplemental data related to this article can be found at: https://doi.org/10.1016/j.adaj.2019.05.029.

Dr. Mishra is a clinical research fellow, National Institute of Dental and Craniofacial Research, National Institutes of Health, 10 Center Dr., Room 5-2531 (northeast atrium), MSC 1470, Bethesda, MD 20892-1470, e-mail rashmi1402@gmail.com. Address correspondence to Dr. Mishra.

Dr. Burke is a clinical research fellow, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Dr. Gitman is a staff clinician, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Dr. Verma is a postdoctorate fellow, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Dr. Engelstad is an associate professor, Oregon Health & Sciences University, Oregon University, Portland, OR.

Dr. Haendel is a professor, Oregon Health & Sciences University, Oregon University, Portland, OR.

Dr. Alevizos is the chief, Sjogren's Syndrome Clinic, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Dr. Gahl is the clinical director, National Human Genomic Research Institute, National Institutes of Health, Bethesda, MD.

Dr. Collins is the chief, Skeletal Clinical Studies Unit, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Dr. Lee is the clinical director, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Dr. Sincan is a staff scientist, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD.

Disclosures. Dr. Haendel is a founding member of Pryzm Health. Dr. Sincan is an employee of Sanford Health System since 2017. None of the other authors reported any disclosures.

This work was carried out with funding from the National Institute of Dental and Craniofacial Research, National Institutes of Health Intramural Research Program.

The authors thank the Biomedical Translational Research Information System staff at the National Institutes of Health who assisted with data retrieval. The authors also thank Dr. Leslie Biesecker (National Human Genomic Research Institute, National Institutes of Health) for thorough review and expert comments in preparation of the manuscript. In addition, the authors thank the laboratory members of the Craniofacial Anomalies and Regeneration Section for insightful input on the research methodology.

1. Du P, Feng G, Flatow J, et al. From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*. 2009;25(12):i63-i68.

2. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Database issue):D267-D270.

3. Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet*. 2010;77(6):525-534.

4. Smith CL, Eppig JT. The Mammalian Phenotype ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome*. 2012; 23(9-10):653-668.

5. Gruber TR. A translation approach to portable ontology specification. *Knowledge Acquisition*. 1993;5(2): 199-220.

6. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008;83(5):610-615.

7. Kohler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 2014;42(Database issue):D966-D974.

8. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat*. 2012;33(5):777-780.

9. Posey JE, Harel T, Liu P, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. *N Engl J Med*. 2017;376(1):21-31.

10. Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc*. 2015;22(5):993-1000.

11. Weatherspoon D, Chattopadhyay A. International Classification of Diseases codes and their use in dentistry. *J Dent Oral Craniofac Epidemiol*. 2013;1(4):20-26.

12. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008;35:128-144.

13. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507-513.

14. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. *Stud Health Technol Inform*. 2010;160(Pt 2):1299-1303.

15. Deleger L, Li Q, Lingren T, et al. Building gold standard corpora for medical natural language processing tasks. *AMIA Annu Symp Proc*. 2012;2012:144-153.

16. Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. *Cold Spring Harb Mol Case Stud*. 2015;1(1):a000372.

17. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20(1):117-121.

18. Zemojtel T, Kohler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*. 2014;6(252):252ra123.

eBox 1. Steps for the NLP[*] methodology.

- Install Apache cTAKES[†] 3.2.2 (check for the latest version).
- In the initial setup cTAKES performs limited recognition of named entity because of its limited in-built library. To expand on the library, you can add UMLS credentials to the cTAKES code.
- Obtain UMLS[‡] credential from: https://uts.nlm.nih.gov//license.html.
- Add your credentials: cd $CTAKES_HOME java -Dctakes.umlsuser=user name -Dctakes.umlspw=Password -cp $CTAKES_HOME/desc/:$CTAKES_HOME/resources/:$CTAKES_HOME/lib/* -Dlog4j.configuration=file:$CTAKES_HOME/config/log4j.xml -Xms512M -Xmx3g org.apache.uima.tools.cpm.CpmFrame "$@".
- Process the XML cTAKES output through the named entity extractor to obtain the disease or disorder phenotype output. Command lines can also be changed to get anatomic sites, signs and symptoms, and medications list, using the named entity extractor.
- For more detailed information, visit http://healthnlp.github.io/examples/.
- For more detailed codes, please contact the first author (R.M.).

[*] NLP: Natural language processing. [†] cTAKES: clinical Text Analysis and Knowledge Extraction System. [‡] UMLS: Unified Medical Language System.

eBox 2. Craniofacial and dental terms not in Human Phenotype Ontology.

Alveolar periostitis
Anterior crossbite
Bilateral crossbite
Buccal exostosis
Chiari malformation type II
Class I malocclusion
Class II malocclusion
Class III malocclusion
Cradle cap
Craniolacunia
Deep overbite
Dental plaque
Dilaceration
Distoocclusion of teeth
Hypertrophy of tonsils
Hypertropia
Interfrontal craniofaciosynostosis
Lingual-facial-buccal dyskinesia
Midline deviation of dental arch
Posterior crossbite
Septal hypertrophy
Telogen effluvium
Tooth ankylosis
Tooth attrition
Tooth erosion
Tooth fractures

eTable 1. Full search criteria based on clinical terms related to the dental, oral, and craniofacial region.

MeSH* TERMS
Abnormalities, Craniofacial Abnormality, Craniofacial Craniofacial Abnormality Craniofacial and skeletal abnormality Craniomandibular Disorder Disorder, Craniomandibular Disorders, Craniomandibular Craniomandibular Diseases Craniomandibular Disease Disease, Craniomandibular Diseases, Craniomandibular Disease, Maxillary Diseases, Maxillary Maxillary Disease Craniofacial Dysostoses Dysostoses, Craniofacial Craniofacial Dysostosis Dysostosis, Craniofacial Craniofacial Dysarthrosis Craniofacial Dysostosis, Type 1; CFD1 Crouzon Disease Disease, Crouzon Crouzon's Disease Crouzons Disease Disease, Crouzon's Crouzon Craniofacial Dysostosis Craniofacial Dysostosis, Crouzon Dysostosis, Crouzon Craniofacial Crouzon Syndrome Syndrome, Crouzon Craniofacial Dysostosis Type 1 Craniofacial Dysostosis, Type I Deformities, Dentofacial Deformity, Dentofacial Dentofacial Deformity Dentofacial Abnormalities Abnormalities, Dentofacial Abnormality, Dentofacial Dentofacial Abnormality Dentofacial Dysplasia Dentofacial Dysplasias Dyplasia, Dentofacial Dyplasias, Dentofacial
ICD-9† CODES
376.44 524.03 524.73 351.8 351.9 438.83 758.32 767.5 781.94 524.50 524.59 524.89 524.9

* MeSH: Medical Subject Heading. † ICD: *International Classification of Diseases*.[11]

eTable 2. Manual annotation schema.

TEXT	DISEASE OR DISORDER TERMS
FIRST_NAMEi=120]with PIK3CA related overgrowth who comes for an evaluation and possible surgical treatment options for his facial asymmetry. Normocephalic with obvious right facial enlargement due to soft tissue infiltration, likely lipomatosis. There is an epidermal nevus of the right face extending onto the neck.	Overgrowth Facial asymmetry Normocephalic Facial enlargement Lipomatosis Epidermal nevus