IMPatienT: an integrated web application to digitize, process and explore multimodal patient data.

Corentin Meyer, Norma Romero, Teresinha Evangelista, Brunot Cadot, Jocelyn Laporte, Anne Jeannin-Girardon, Pierre Collet, Kirsley Chennen, Olivier Poch

TITLE

IMPatienT: an integrated web application to digitize, process and explore multimodal patient data.

Authors

Corentin Meyer[1], Norma Beatriz Romero[2], Teresinha Evangelista[2], Brunot Cadot[3], Jocelyn Laporte[4], Anne Jeannin-Girardon[1], Pierre Collet[1], Kirsley Chennen[1], Olivier Poch[1*]

[1] *Complex Systems and Translational Bioinformatics (CSTB), ICube Laboratory, UMR 7357, University of Strasbourg, 1 rue Eugène Boeckel, 67000 Strasbourg, France.*

[2] *Neuromuscular Morphology Unit, Myology Institute, Reference Center of Neuromuscular Diseases Nord-Est-IDF, GHU Pitié-Salpêtrière, Paris, France*

[3] *Sorbonne Université, INSERM, Center for Research in Myology, Myology Institute, GHU Pitié-Salpêtrière, Paris, France*

[4] *Department Translational Medicine, IGBMC, CNRS UMR 7104, 1 rue Laurent Fries, 67404 Illkirch, France.*

* Correspondence: olivier.poch@unistra.fr

ABSTRACT

Background

Medical acts, such as imaging, generally lead to the production of several medical text reports that describe the relevant findings. Such processes induce multimodality in patient data by linking image data to free-text data and consequently, multimodal data have become central to drive research and improve diagnosis of patients. However, the exploitation of patient data is

challenging as the ecosystem of available analysis tools is fragmented depending on the type of data (images, text, genetic sequences), the task to be performed (digitization, processing, exploration) and the domain of interest (clinical phenotype, histology…). To address the challenges, the analysis tools need to be integrated in a simple, comprehensive, and flexible platform.

Results

Here, we present IMPatienT (d**I**gitize **M**ultimodal **PATIEN**t da**T**a), a free and open-source web application to digitize, process and explore multimodal patient data. IMPatienT has a modular architecture, including four components to: (i) create a standard vocabulary for a domain, (ii) digitize and process free-text data by mapping it to a set of standard terms, (iii) annotate images and perform image segmentation, and (iv) generate an automatic visualization dashboard to provide insight on the data and perform automatic diagnosis suggestions. Finally, we demonstrate the usefulness of IMPatienT on a corpus of 40 simulated muscle biopsy reports of congenital myopathy patients.

Conclusions

IMPatienT is a platform to digitize, process and explore patient data that can handle image and free-text data. As it relies on a user-designed vocabulary, it can be adapted to fit any domain of research and can be used as a patient registry for exploratory data analysis (EDA). A demo instance of the application is available at https://impatient.lbgi.fr.

KEYWORDS

Patient data, free-text medical reports, NLP, OCR, data formatting, data processing, image segmentation, exploratory data analysis

BACKGROUND

Patient data now incorporates the results of numerous modalities, including imaging, next-generation sequencing and more recently wearable devices. Most of the time, medical acts produce imaging data, such as echography, radiology or histology result in the production of medical reports that describe the relevant findings. Thus, multimodality is induced in patient data, as imaging data is inherently linked to free-text reports.

Useful tools to centralize, process and explore multimodal data are essential to drive research and improve diagnosis. The use of multimodal data has been shown to increase disease understanding and diagnosis (1–4). For example, Venugopalan *et al.* integrated genetic data with image data and medical records (free-text data) to improve diagnosis of Alzheimer's disease (4). In Mendelian diseases, integration of multiple levels of information is key to the establishment of a diagnosis. For instance, in congenital myopathies (CM), a combination of muscle biopsy analysis / histological analysis (imaging information) with medical records and sequencing data is essential for differential diagnosis between CM subtypes (5–7). Centralization of multimodal data using dedicated software is essential to implement such an approach.

However, the ecosystem of tools for the exploitation of patient data is heavily fragmented, depending on the type of data (images, text, genetic sequences), the task to be performed (digitization, processing, exploration) and the domain of interest (clinical phenotype, histology…). Exploitation tools can be divided in two main categories: (i) tools to process the data and (ii) tools to explore the data.

Clinical reports (free-text) processing relies on the use of a standard vocabulary, such as the Unified Medical Language System (UMLS)(8) or the Human Phenotype Ontology (HPO)(9). Several tools have been developed to easily manage and extend these standard vocabularies, such as OBOEdit(10). Text mining processes have been developed based on these standard vocabularies,

that can automatically detect keywords from the free-text data. For example, Doc2HPO(11) can extract a list of HPO terms from free-text medical records. Other software packages, such as Phenotips(12), have been developed to centralize and process general patient information, such as demographics, pedigree, common measurements, phenotypes and genetic results. Finally, for imaging data, software to process and annotate gigapixel-scale microscopy images are widely used, including Cytomine(13) and SlideRunner(14).

A wide range of tools has been developed to analyze and explore patient data. For example, based on a list of HPO terms describing a patient's specific phenotypic profile, Phenolyzer(15) and Phenomizer(16) can be used to help prioritize candidate genes or rank the best-matching diseases. However, these tools are restricted to the use of HPO terms to describe the patient's profile and are not compatible with other ontologies. Ontology agnostic algorithms have also been developed that predict an outcome based on a list of terms from any normalized vocabulary, such as the Bayesian Ontology Query Algorithm (BOQA) (17). For multimodal approaches, ClinPhen (18) and Exomiser (19) have successfully combined multiple levels of information with both phenotype information (HPO terms) and genetic information (variants) to rank candidate genes in Mendelian diseases.

This large ecosystem of tools highlights the need for an integrated tool that can: (i) both process and explore patient data, (ii) manage multimodal data (text and images), and (iii) work in any domain of interest.

In this study, we present IMPatienT (d**I**gitize **M**ultimodal **PATIEN**t da**T**a), a free and open-source web application that aims to be an integrated tool to digitize, process and explore multimodal patient data**.** Using a modular architecture, we developed four components to: (i) create a standard vocabulary describing a domain of interest, (ii) digitize and process free-text records by automatically mapping them to a set of standard terms, (iii) annotate and segment images with

standard vocabulary, and (iv) generate a dashboard with automatic visualizations to explore the patient data and perform automatic diagnosis suggestions.

Finally, we demonstrate the usefulness of IMPatienT on a set of congenital myopathy (CM) cases. CM are a family of rare genetic diseases, including multiple distinct subtypes, that still lack proper diagnosis with more than 50% of patients without a genetic cause identified(20). We exploited IMPatienT to create a list of standard muscle-histology terms that were then used to process patient histological records and annotate biopsy images. Finally, multiple exploratory visualizations were automatically generated.

IMPLEMENTATION

IMPatienT is a web application developed with the Flask micro-framework, which is a Python-based web framework. Figure 1 illustrates the global organization of the web application. The web application is composed of four modules: (i) Standard Vocabulary Creator, (ii) Report Digitization, (iii) Image Annotation, and (iv) Automatic Visualization Dashboard. All modules incorporate free, open-source and well-maintained libraries that are described in detail in the corresponding sections.

Module 1: Standard Vocabulary Creator

The standard vocabulary creator module allows to create and modify a hierarchical list of vocabulary terms with rich definitions that can be used as an image annotation class, for text report processing, or suggestion of diagnosis.

Figure 2 shows a screenshot of the page used to create and manage the standard vocabulary tree. The ergonomic drag and drop system using the graphical user interface (GUI) allows the user to intuitively and quickly edit and reorganize the vocabulary to add new terms or modify existing ones. Also, the vocabulary term (node) details form makes it easy to edit term properties.

The tree is generated and rendered with the JavaScript library JSTree (version 3.3.12). Each node (term) can have only one parent. For each created node (vocabulary terms), the user can assign a name and organize the tree structure (hierarchy) through the drag and drop interface. Each term in the tree is associated with night optional properties. Four properties are defined by the user: description, list of synonyms, translation in another language, show the term as annotation class. Two properties are automatically generated: the term's unique identifier (ID) and the hexadecimal color associated with the term (for image annotation). Additional term properties (associated diagnosis/disease class, associated genes, list of positively correlating terms (*i.e.* co-occurring terms in reports)) are automatically extracted from patient records registered in the database.

Finally, if the user defined an alternative translation for terms, there is an "invert vocabulary language" button to conveniently switch between standard vocabulary languages. For instance, the user can create a vocabulary in any language and define the translation in English, then switch between the two display modes easily.

Module 2: Report Digitization

The standard vocabulary terms are used to process documents that are in a free-text format. Module 2 uses a semi-automatic approach for digitization and processing of free-text reports that combines fast automatic detection of terms with manual reviewing of the detection. The interface of Module 2 is a form divided into four parts (Figure 3).

In the first part of the digitization form (Fig 3A), a PDF file of the free-text report can be uploaded for natural language processing (NLP) of the content and pre-filling of the form. The text of the PDF report is automatically extracted and detected vocabulary terms are highlighted (see corresponding section below "Optical Character Recognition and Vocabulary Terms Detection"). Highlighted terms allow to easily identify any misdetection or terms that have been missed during the automatic analysis.

The second part (Fig 3B) of the digitization form contains patient information, such as patient ID, document ID, age of the patient. IMPatienT exploits well-established ontologies to normalize the genetic diagnosis and phenotypes (Fig 4). Gene symbols are retrieved from the HUGO Gene Nomenclature Committee (HGNC)(21) and mutation notations are retrieved from the HGVS sequence variant nomenclature(22). Phenotypes are normalized using the HPO ontology. None of these fields contain patient identifying data and are optional.

The third part of the digitization form (Fig 3C) contains the standard vocabulary tree viewer with an absence/presence slider. Each vocabulary term can be marked as present, absent or no information. For terms marked as present, the slider is used to indicate a notion of quantity or certainty of the term. For example, the statement 'There is a small number of fibers containing rods' can be annotated by hand by setting the vocabulary 'Rods' to the value 'Present' with a low quantity value. For terms that have been automatically detected, this slider value is automatically set to 0 (present in a negated sentence) or 1 (present).

Finally, the fourth part (Fig 3D) of the digitization form concerns the suspected diagnosis chosen by the user (disease name from the Orphanet(23) knowledge base) and commentaries or notes. It also includes an automatic suggestion of the diagnosis based on already registered patients using BOQA (23) (see corresponding section below "Patient Disease Suggestions Method").

Optical Character Recognition and Vocabulary Term Detection

The patient report digitization in module 2 is facilitated by the automatic text recognition and keyword detection method. The user uploads a PDF version of the text report to perform Optical Character Recognition (OCR), followed by Natural Language Processing (NLP) to automatically detect vocabulary terms from the report. Figure 5 describes the workflow of the vocabulary terms detection method. First the PDF file is converted to plain text using Tesseract OCR (implemented in python as pyTesseract). Then, the text is processed with Spacy, an NLP python library, by splitting the text into sentences and then into individual words. The resulting list of

sentences are then processed to detect negation using a simple implementation of the concept of NegEx (24). A *n*-gram (unigrams, bigrams, and trigrams) procedure is applied to the list of words to identify contiguous words in the context of all the sentences of the report. The *n*-grams are then mapped against the user-created standard vocabulary using fuzzy partial matching (using Levenshtein distance) with a score threshold of 0.8. Matched keywords are kept and shown on the interface with a green or red highlight of the detected text using Mark.JS JavaScript library (green indicates the presence of the keyword, red indicates the presence in a negated sentence). Keywords are also automatically marked as present or absent (negated) in the vocabulary tree.

Disease Suggestions

The report digitization module 2 contains a disease recommendation algorithm inspired by the BOQA algorithm described by Bauer *et al.* (23). Basically, the algorithm computes the similarity between a list of input vocabulary terms annotated as “present” for a patient (the query) and a simulated patient profile for each disease class (model report) that is generated based on the data from already registered patients.

We implemented this algorithm in python, and we modified it to use the frequencies of vocabulary terms per disease for the generation of the model report instead of the initial deterministic way (not frequency aware). This means that the model report is generated based on the probability (frequency) of each vocabulary term. For example, if disease A is annotated with vocabulary term B at a frequency=0.9 and vocabulary term C at a frequency=0.1, the generated model report for disease A will have a probability=0.9 of containing vocabulary term B and a probability=0.1 of containing vocabulary term C.

Due to the stochastic nature of the generation of the model report, for any given prediction, the generation and calculation of the similarity with the query is repeated 50 times. For each repetition, if a disease has a prediction probability >0.5, it is considered to be the best prediction,

otherwise the prediction is “no prediction”. Finally, of the 50 repetitions, the prediction with the highest occurrence is taken as the final prediction.

Module 3: AI-Assisted Image Annotation Using Automatic Segmentation

To process patient image data, we developed the image annotation module (module 3) to upload, annotate and perform image segmentation with standard vocabulary terms. This module is based on the *“interactive image segmentation with Dash and Scikit-image”* demonstration application (25–27). The original source code was modified to be compatible with the standard vocabulary tree and the database.

The interactive interface to annotate image features with standard vocabulary terms is presented in figures 6A and 6B. The interface allows to draw a free-shape area (annotation) associated with a standard vocabulary term (class). Then, with a minimal number of user annotations, the whole image is segmented based on the annotations (shapes) provided by the user.

To perform image segmentation, on the server side, local features (intensity, edges, texture) are extracted from the labeled areas of the image and are used to train a dedicated AI random-forest classifier model. This dedicated model is then applied to predict similar areas in the whole image. Finally, every pixel of the image is labeled with a standard vocabulary term corresponding to the AI prediction based on the annotations.

The segmentation is entirely interactive. After the initial segmentation, the user can correct the classification by adding more annotation shapes to the image and can modify the paintbrush width setting to make more precise annotation marks. In addition, the stringency range parameter of the model can be adapted using the slider to modify the model behavior and automatically recompute the segmentation in real time.

Results of the segmentation are retrievable as a single archive including the raw image, the annotations (JSON), the random-forest trained classifier, the blended image and the segmentation mask image.

Module 4: Automatic Visualization Dashboard

The automatic visualization dashboard module is designed to perform exploratory data analysis by generating multiple graphs based on the patient data in the database. All visualizations are created using Plotly, a python graph library, that allows to make interactive graphs.

Application Security and Personal Data

IMPatienT is developed as a free and open-source project meaning that the code can be audited by anyone in the GitHub code repository. The code is regularly scanned for known issues and outdated libraries to mitigate security issues. There is no patient-identifying data kept in the database, only a custom identifier and age. No name or date of birth are required or stored. Additionally, access to all modules and data entered via the web application is restricted by a login-page and user accounts can only be created by the administrator of the platform. No user information is stored except for the username, email and salted and hashed passwords.

RESULTS

IMPatienT is an interactive and user-friendly web application that integrates a semi-automatic approach for text and image data digitization, processing, and exploration. Due to its modular architecture and its standard vocabulary creator, it has a wide range of potential uses.

Web Application Workflow

Figure 1 shows how the user can interact with the web application to digitize, process, and explore patient data. The first step is to create a standard vocabulary using the Standard Vocabulary Creator interface (module 1). The user only needs to create a few terms (nodes) to

begin using the web application. Defining the properties of the terms (definition, synonyms…) is optional, and organizing them in a hierarchical structure is also optional.

Then, the user can start digitizing patient reports using module 2 (step 2). This can be done manually by filling out the form in module 2 and checking terms as present or absent in a given report, or the user can employ the Vocabulary Term Matching method by uploading a PDF version of the report. Using module 3, the user can also upload, annotate, and segment image data.

Finally, the user can view multiple exploratory graphs (histograms, correlation matrix, confusion matrix, frequency tables) that are automatically generated in module 4. All data entered via the web application are retrievable in standard formats, including the whole database of reports as a single SQLite3 file or CSV files, the images and their segmentation models and masks as a GZIP archive, the standard vocabulary with annotation as a JSON file and various graphs and tables as JSON or PNG files.

Use Case: Congenital Myopathy Histology Reports

As a use case of IMPatienT, we focused on congenital myopathies (CM). We used the standard vocabulary creator to create a sample muscle histology standard vocabulary based on common terms used in muscle biopsy reports from the Paris Institute of Myology. Then, we inserted 40 generated digital patients in the database with random sampling of standard vocabulary terms and associated a gene and disease class among a list of common CM genes and three recurring CM subtypes (nemaline myopathy, core myopathy and centronuclear myopathy). All these data are available on the demo instance of IMPatienT (https://impatient.lbgi.fr/).

For the image data, figure 7 shows an example of the segmentation of a biopsy image, where we annotated the cytoplasm of the cells (green), intercellular spaces (black) and cell nuclei (red). The raw image (Fig 7A) is annotated with free-shape areas associated with standard vocabulary terms

(Fig 7B). Then, the whole image is automatically segmented based on the annotations, producing the segmentation mask where each pixel is associated with a class (Fig 7C 7D).

The automatic visualization dashboard was used to generate the six visualizations provided in Figure 8. These visualizations include a breakdown of the patients in the database by age, genes, or diagnosis (Fig 8A). A correlation matrix (using Pearson correlation coefficient) between the occurrence of standard vocabulary terms is generated (Fig 8B), which can serve as a starting point for exploration of co-occurrence of features in patients. The confusion matrix of the final diagnosis of patients versus the suggested diagnosis with BOQA (Fig 8C) allows the user to monitor the accuracy of the disease suggestion function. In addition, a histogram showing the classification of patients without a final diagnosis is provided to indicate possible prognosis of undiagnosed patients (Fig 8D). Finally, the frequency of each standard vocabulary term by gene and by disease is automatically calculated and shown in two tables (see Additional file 1 and 2).

DISCUSSION

IMPatienT is a platform that simplifies the digitization, processing, and exploration of both textual and image patient data. The web application is centered around the concept of a standard vocabulary tree that is easy to create and used to process text and image data. This allows IMPatienT to work with patient data from domains that still lack a consensus ontology and rely on well-established ontologies for patient data, such as HPO for phenotypes, Orphanet for disease names or HGCN/HGVS for genetic diagnoses.

The semi-automatic approach implemented in IMPatienT offers faster digitization processes while ensuring accuracy through manual review. This is achieved by analyzing text data using OCR and NLP, then matching the text to the standard vocabulary, followed by manual correction. For image data, the user first provides sparse annotations on the image, which are then used to calculate an automatic segmentation of the whole image. For data exploration, IMPatienT uses a

fully automatic approach including various visualizations as well as diagnosis suggestions, while allowing the user to extract the processed data in a standard format for further analysis (database, images, frequency tables).

IMPatienT aims to integrate multiple approaches in a unified platform with two main objectives: universality (*i.e* not restricted to a specific domain) and multimodality (*i.e.* integration of multiple data types). To our knowledge, other tools similar to IMPatienT do not fulfill both objectives. For example, Doc2HPO is a tool that also uses a semi-automatic approach to digitize clinical text according to a list of HPO terms, based on NLP methods and negation detection. However, as Doc2HPO is restricted to HPO, it does not provide custom vocabulary tree facilities. In contrast IMPatienT is suitable for digitization of text data not described by HPO.

Cytomine is a powerful software package for image annotation and analysis, that includes an ontology builder and complex image processing tools. However, it is restricted to image data only and thus it does not provide a multimodal approach. In contrast, IMPatienT uses an integrated approach that couples images with text reports in a single platform.

Finally, in IMPatienT we reimplemented the diagnosis suggestion algorithm called BOQA that is also used in Phenomizer, a tool to rank a list of the top matching diseases based on a list of input HPO terms. We modified the algorithm to consider frequencies of terms by disease to have meaningful predictions. However, BOQA uses binary states for terms (terms are marked as present or absent) and is not compatible with numeric features. In the future, it will be necessary to implement a more complex system such as explainable AI with learning classifier systems (28). This should improve accuracy, explainability, and handling of quantitative values, although at the cost of computational power.

In the future, we plan to further develop IMPatienT by exploring the automatization of the standard vocabulary creation with the analysis of a complete corpus of text. For text analysis, we

wish to implement additional context comprehension, *i.e.* not only negation but also hypothetical statements, uncertainty and family context as well as better text-vocabulary terms matching. For image data, we plan to add support for DICOM images and whole slide images.

CONCLUSIONS

With IMPatienT, we have developed an integrated web application to digitize, process and explore multimodal patient data. Thanks to its standard vocabulary creator module, it can be adapted to any domain that currently lacks a standard vocabulary. It provides automatization of the task of processing free-text patient data and annotating images. It also provides automatic data exploration with the diagnosis suggestion algorithm and the visualization dashboard. IMPatienT can serve as a research tool to find new associations of patient features that might be relevant for diagnosis. A demonstration instance of the web application is available at https://impatient.lbgi.fr.

Availability and requirements

Project name: IMPatienT v1.4.3

Project home page: https://github.com/lambda-science/IMPatienT

Project Version: v1.4.3 https://github.com/lambda-science/IMPatienT/releases/tag/v1.4.3

Operating system(s): Platform independent

Programming language: Python 3.9, HTML5, CSS3, JS

Other requirements: Flask 2.0, Spacy 3.2, Dash 2, Plotly 5, Scikit-Learn 1.0, Scikit-Image 0.19, Tesseract 4, JSTree 3.3, Mark.JS 8.11, JQuery 3.6, Bootstrap 5

License: GNU AGPL

Any restrictions to use by non-academics: /

Figures titles and legends

Figure 1: IMPatienT web application workflow and organisation

Figure 2: Screenshot of the Standard Vocabulary Creator module (module 1).

(A) The hierarchical structure viewer and editor tool that supports drag and drop modification and creation/deletion/modification using the mouse.

(B) The properties of the selected term node with its unique ID, display name, alternative language translation, synonyms, description, associated genes and diseases and correlating terms extracted from the application instance database.

Figure 3: Screenshot of the report digitalization module.

(A) PDF upload section for automatic keyword detection in the text. Detected keywords have a green background, detected and negated keywords have a red background.

(B) Patient information section (age, document ID, gene, mutation, phenotype).

(C) Standard vocabulary tree viewer to select keywords with associated slider to manually indicate keyword value (absence or presence level). Keywords marked as present are indicated with a green check mark, absent keywords are marked with a red cross.

(D) Final section with an overview of all annotated terms, diagnosis selection and commentary part with automatic diagnosis suggestion using BOQA algorithm.

Figure 4: Overview of the ontologies used by IMPatienT to process patient data in the report digitization module (module 2).

Figure 5: Optical character recognition and vocabulary term detection method used in the report digitization module (module 2) to automatically analyse free-text reports.

Figure 6: Screenshot of the image annotation module.

(A) Image viewer used to navigate, zoom and annotate the histology image.

(B) Menu interface to select the annotation label, brush width and segmentation parameters.

Figure 7: Image segmentation process in the image segmentation module.

(A) Raw image input before annotation.

(B) Image with limited manual annotation of cytoplasm (green), cell nucleus (red) and intercellular space (black).

(C) Blended image of the raw image and segmented image after automated segmentation with a random-forest classifier.

(D) Segmented image mask alone.

Additional files

Additional File 1:

- File Name: table_frequencies_per_gene.csv
- File Format: CSV
- Title of data: Table of frequencies of standard vocabulary per genes
- Description of data: This CSV file contains all frequencies of standard vocabulary terms for each gene with the total number of reports per gene and the number of occurrences of each term if not 0.

Additional File 2:

- File Name: table_frequencies_per_diag.csv
- File Format: CSV
- Title of data: Table of frequencies of standard vocabulary per diagnosis

- Description of data: This CSV file contains all frequencies of standard vocabulary terms for each diagnosis with the total number of reports per diagnosis and the number of occurrences of each term if not 0.

List of abbreviations

AI: Artificial Intelligence

BOQA: Bayesian Ontology Query Algorithm

CM: Congenital Myopathies

EDA: Exploratory Data Analysis

GUI: Graphical User Interface

HGCN: HUGO Gene Nomenclature Committee

HGVS: Human Genome Variation Society

HPO: Human Phenotype Ontology

NLP: Natural Language Processing

OCR: Optical Character Recognition

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and analyzed during the current study are available in the IMPatienT repository, https://github.com/lambda-science/IMPatienT.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by the Agence Nationale de la Recherche (ANR), 80 | Prime funds from the CNRS (MYO-xIA Project), the University of Strasbourg and INSERM.

Authors' Contributions

AJG, PC, KC and OP conceived the research topic. CM and KC designed the application and wrote the original draft. CM implemented the application. NBR, TE, BC and JL provided biopsy reports and performed data curation and software evaluation. All authors read and approved the final manuscript.

Acknowledgements

We thank the BiGEst-ICube platform for their assistance.

References

1. Kerr WT, Hwang ES, Raman KR, Barritt SE, Patel AB, Le JM, et al. Multimodal diagnosis of epilepsy using conditional dependence and multiple imputation. In: 2014 International Workshop on Pattern Recognition in Neuroimaging. 2014. p. 1–4.

2. Yan R, Ren F, Rao X, Shi B, Xiang T, Zhang L, et al. Integration of Multimodal Data for Breast Cancer Classification Using a Hybrid Deep Learning Method. In: Huang D-S, Bevilacqua V, Premaratne P, editors. Intelligent Computing Theories and Application. Cham: Springer International Publishing; 2019. p. 460–9. (Lecture Notes in Computer Science).

3. Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. Transl Res J Lab Clin Med. 2018 Apr;194:56–67.

4. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep. 2021 Feb 5;11(1):3254.

5. North KN, Wang CH, Clarke N, Jungbluth H, Vainzof M, Dowling JJ, et al. Approach to the diagnosis of congenital myopathies. Neuromuscul Disord NMD. 2014 Feb;24(2):97–116.

6. Cassandrini D, Trovato R, Rubegni A, Lenzi S, Fiorillo C, Baldacci J, et al. Congenital myopathies: clinical phenotypes and new diagnostic tools. Ital J Pediatr. 2017 Nov 15;43(1):101.

7. Böhm J, Vasli N, Malfatti E, Le Gras S, Feger C, Jost B, et al. An integrated diagnosis strategy for congenital myopathies. PloS One. 2013;8(6):e67527.

8. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267–70.

9. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1207–17.

10. Day-Richter J, Harris MA, Haendel M, The Gene Ontology OBO-Edit Working Group, Lewis S. OBO-Edit—an ontology editor for biologists. Bioinformatics. 2007 Aug 15;23(16):2198–200.

11. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and accurate HPO concept curation. Nucleic Acids Res. 2019 Jul 2;47(W1):W566–70.

12. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. Hum Mutat. 2013;34(8):1057–65.

13. Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016 May 1;32(9):1395–401.

14. Aubreville M, Bertram C, Klopfleisch R, Maier A. SlideRunner - A Tool for Massive Cell Annotations in Whole Slide Images. ArXiv180202347 Cs. 2018;309–14.

15. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015 Sep;12(9):841–3.

16. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009 Oct;85(4):457–64.

17. Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics. 2012 Oct 1;28(19):2502–8.

18. Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet Med. 2019 Jul;21(7):1585–93.

19. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015 Dec;10(12):2004–15.

20. H J, S T, F Z, A S, J O, C S, et al. Congenital myopathies: disorders of excitation-contraction coupling and muscle contraction. Nat Rev Neurol [Internet]. 2018 Mar [cited 2022 Mar 16];14(3). Available from: https://pubmed.ncbi.nlm.nih.gov/29391587/

21. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D939–46.

22. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016;37(6):564–9.

23. INSERM. Orphanet: an online database of rare diseases and orphan drugs [Internet]. 1997 [cited 2022 Feb 13]. Available from: http://www.orpha.net

24. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform. 2001 Oct 1;34(5):301–10.

25. Gouillart E. Interactive Machine Learning - Image segmentation [Internet]. GitHub. 2020 [cited 2021 Nov 23]. Available from: https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation

26. Walt S van der, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014 Jun 19;2:e453.

27. Hossain S. Visualization of Bioinformatics Data with Dash Bio. Proc 18th Python Sci Conf. 2019;126–33.

28. Urbanowicz RJ, Moore JH. ExSTraCS 2.0: Description and Evaluation of a Scalable Learning Classifier System. Evol Intell. 2015 Sep;8(2):89–116.