



**HAL**  
open science

## **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

### ► To cite this version:

Corentin Meyer, Norma Romero, Teresinha Evangelista, Brunot Cadot, Jocelyn Laporte, et al.. IMPatient: an Integrated web application to digitize, process and explore Multimodal PATIENT daTa.. 2022. hal-03635350v2

**HAL Id: hal-03635350**

**<https://hal.science/hal-03635350v2>**

Preprint submitted on 27 Sep 2022 (v2), last revised 17 Jun 2024 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# 1 HUMAN MUTATION

## 2 TITLE

3 IMPatient: an Integrated web application to digitize, process and explore  
4 Multimodal PATIENT data.

## 5 Authors

6 Corentin Meyer<sup>1</sup>, Norma Beatriz Romero<sup>2</sup>, Teresinha Evangelista<sup>2</sup>, Brunot Cadot<sup>3</sup>,  
7 Jocelyn Laporte<sup>4</sup>, Anne Jeannin-Girardon<sup>1</sup>, Pierre Collet<sup>1</sup>, Kirsley Chennen<sup>1</sup>, Olivier Poch<sup>1\*</sup>

8 <sup>1</sup> *Complex Systems and Translational Bioinformatics (CSTB), ICube Laboratory, UMR 7357, University of*  
9 *Strasbourg, 1 rue Eugène Boeckel, 67000 Strasbourg, France.*

10 <sup>2</sup> *Neuromuscular Morphology Unit, Myology Institute, Reference Center of Neuromuscular Diseases*  
11 *Nord-Est-IDF, GHU Pitié-Salpêtrière, Paris, France*

12 <sup>3</sup> *Sorbonne Université, INSERM, Center for Research in Myology, Myology Institute, GHU Pitié-*  
13 *Salpêtrière, Paris, France*

14 <sup>4</sup> *Department Translational Medicine, IGBMC, CNRS UMR 7104, 1 rue Laurent Fries, 67404 Illkirch,*  
15 *France.*

16 \* Correspondence should be addressed to: Olivier Poch [olivier.poch@unistra.fr](mailto:olivier.poch@unistra.fr)

17

## 18 **ABSTRACT**

19 Medical acts, such as imaging, generally lead to the production of several medical text reports  
20 that describe the relevant findings. Such processes induce multimodality in patient data by  
21 linking image data to free-text data and consequently, multimodal data have become central to  
22 drive research and improve diagnosis of patients. However, the exploitation of patient data is  
23 challenging as the ecosystem of available analysis tools is fragmented depending on the type of  
24 data (images, text, genetic sequences), the task to be performed (digitization, processing,  
25 exploration) and the domain of interest (clinical phenotype, histology...). To address the  
26 challenges, the analysis tools need to be integrated in a simple, comprehensive, and flexible  
27 platform. Here, we present IMPatientT (**I**ntegrated digital **M**ultimodal **P**ATIENT **d**a**T**a), a free and  
28 open-source web application to digitize, process and explore multimodal patient data. IMPatientT  
29 has a modular architecture, including four components to: (i) create a standard vocabulary for a  
30 domain, (ii) digitize and process free-text data by mapping it to a set of standard terms, (iii)  
31 annotate images and perform image segmentation, and (iv) generate an automatic visualization  
32 dashboard to provide insight on the data and perform automatic diagnosis suggestions. Finally,  
33 we demonstrate the usefulness of IMPatientT on a corpus of 40 simulated muscle biopsy reports  
34 of congenital myopathy patients. IMPatientT is a platform to digitize, process and explore patient  
35 data that can handle image and free-text data. As it relies on a user-designed vocabulary, it can  
36 be adapted to fit any domain of research and can be used as a patient registry for exploratory  
37 data analysis (EDA). A demo instance of the application is available at <https://impatient.lbgi.fr>.

## 38 **KEYWORDS**

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

## 41 INTRODUCTION

42 Patient data now incorporates the results of numerous modalities, including imaging, next-  
43 generation sequencing and more recently wearable devices. Most of the time, medical acts  
44 produce imaging data, such as echography, radiology or histology result in the production of  
45 medical reports that describe the relevant findings. Thus, multimodality is induced in patient  
46 data, as imaging data is inherently linked to free-text reports. The link between image and report  
47 data is crucial as raw images can be re-interpreted during the patient's medical journey with new  
48 domain knowledge or by different experts leading to different reports. Thus, patient multimodal  
49 data needs to be processed in an integrated way to preserve this link in a single database.

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

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

64 Clinical reports (free-text) processing relies on the use of a standard vocabulary, such as the  
65 Unified Medical Language System (UMLS) [8] or the Human Phenotype Ontology (HPO)[9].  
66 Several tools have been developed to easily manage and extend these standard vocabularies,  
67 such as Protégé [10]. Text mining processes have been developed based on these standard  
68 vocabularies, that can automatically detect keywords from free-text data. For example, Doc2HPO  
69 [11] can extract a list of HPO terms from free-text medical records. Other software packages,  
70 such as Phenotips [12] have been developed to centralize and process general patient  
71 information, such as demographics, pedigree, common measurements, phenotypes and genetic  
72 results. SAMS [13] and RD-Connect PhenoStore [14] are other examples of web applications that  
73 aim to perform deep phenotyping of patients by building a single database of standardized  
74 patient data using well-established ontologies such as HPO. Finally, for imaging data, software to  
75 process and annotate gigapixel-scale microscopy images are widely used, including Cytomine  
76 [15], SlideRunner [16] and Ilastik [17]. Cytomine is a powerful software package for gigapixel scale  
77 image annotation and analysis, that includes an ontology builder and complex image processing  
78 tools. However, it is restricted to image data only.

79 A wide range of tools have been developed to analyze and explore patient data. For example,  
80 based on a list of HPO terms describing a patient's specific phenotypic profile, Phenolyzer [18]  
81 and Phenomizer [19] can be used to help prioritize candidate genes or rank the best-matching  
82 diseases. However, these tools are restricted to the use of HPO terms to describe the patient's  
83 profile and are not compatible with other ontologies. Ontology agnostic algorithms have also  
84 been developed that predict an outcome based on a list of terms from any normalized  
85 vocabulary, such as the Bayesian Ontology Query Algorithm (BOQA) [20]. For patient images  
86 exploitation, guidelines and frameworks have been proposed to standardize the measurement of  
87 pathological features from DICOM lung images [21]. Some multimodal approaches such as  
88 ClinPhen [22] and Exomiser [23] have successfully combined multiple levels of information with

89 both phenotype information (HPO terms) and genetic information (variants) to rank candidate  
90 genes in Mendelian diseases. Other tools such as INTEGRO [24] have been developed to  
91 automatically data-mine disease-gene associations for a specific input disease from multiple  
92 curated sources of knowledge.

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

96 In this study, we present IMPatientT (**I**ntegrated digital **M**ultimodal **P**ATIENT **d**a**T**a), a free and  
97 open-source web application that aims to be an integrated tool to digitize, process and explore  
98 multimodal patient data. IMPatientT is a turnkey solution that aims to aggregate patient data and  
99 provides simple tools and interfaces for a clinician to extract information from multimodal  
100 patient data in a single endpoint. Using a modular architecture, we developed four components  
101 to: (i) create a standard vocabulary describing a domain of interest, (ii) digitize and process free-  
102 text records by automatically mapping them to a set of standard terms, (iii) annotate and  
103 segment images with standard vocabulary, and (iv) generate a dashboard with automatic  
104 visualizations to explore the patient data and perform automatic diagnosis suggestions.

105 Finally, we demonstrate the usefulness of IMPatientT on a set of congenital myopathy (CM) cases.  
106 CM are a family of rare genetic diseases, including multiple distinct subtypes, that still lack proper  
107 diagnosis with more than 50% of patients without a genetic cause identified[25]. We exploited  
108 IMPatientT to create a list of standard muscle-histology terms that were then used to process  
109 patient histological records and annotate biopsy images. Finally, multiple exploratory  
110 visualizations were automatically generated.

# INPUT

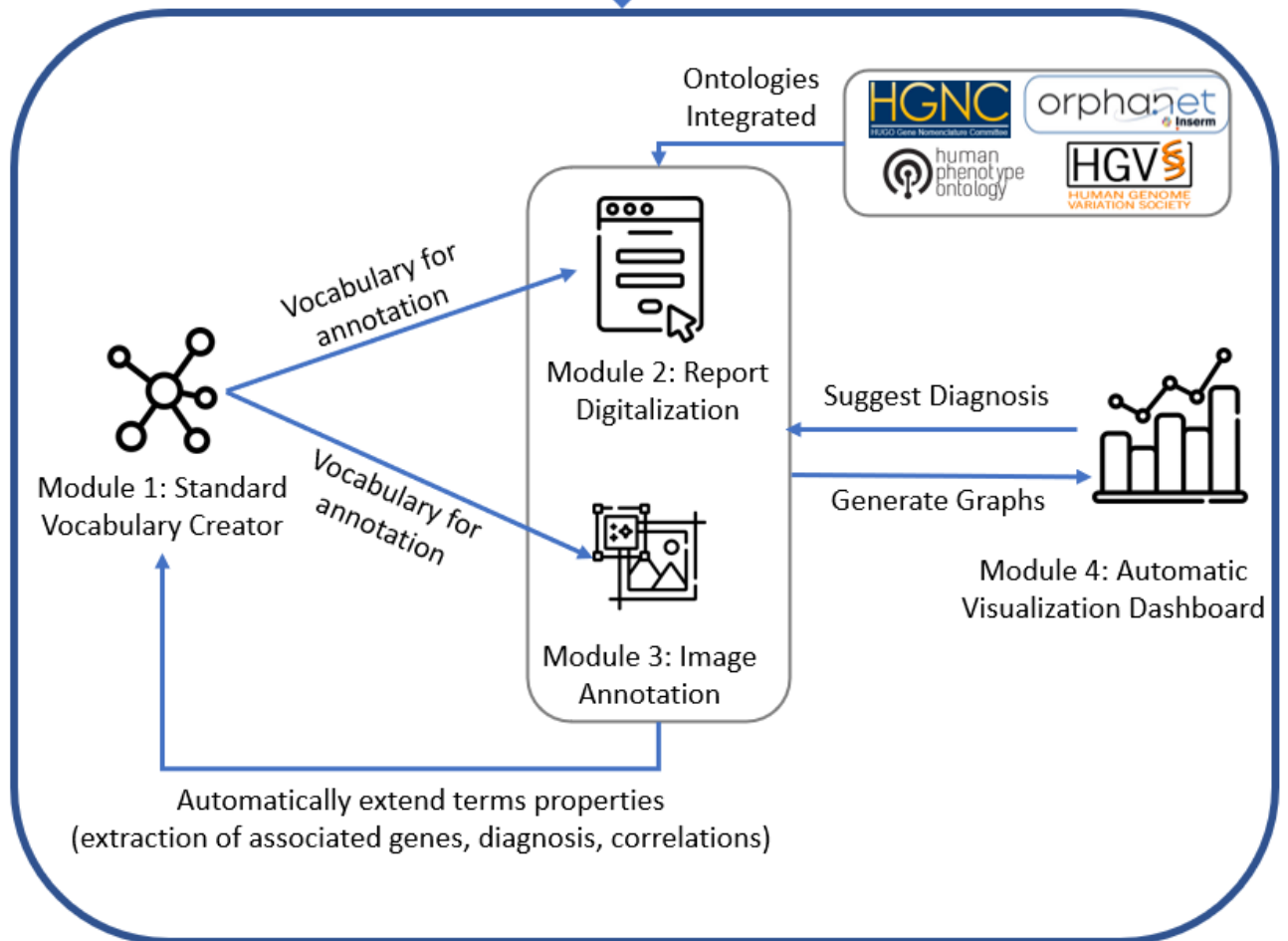


User with unprocessed patient data (free-text medical reports and unannotated images)

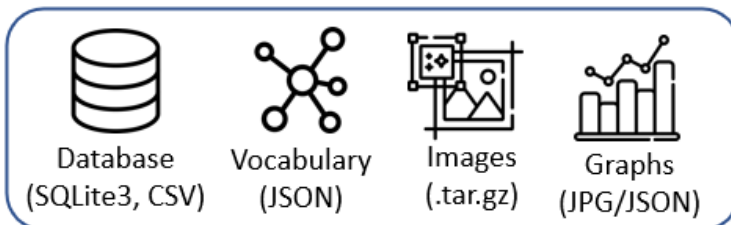
**Step 1:** Creating a standard vocabulary (**module 1**)

**Step 2:** Filling reports and image forms to register patients (**modules 2 and 3**)

# WEB-APPLICATION



# PROCESSED DATA



111

112 **Figure 1: IMPatientT web application organization**

## 113 **MATERIALS AND METHODS**

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

### 120 **Module 1: Standard Vocabulary Creator**

121 The standard vocabulary creator module allows to create and modify a hierarchical list of  
122 vocabulary terms with rich definitions that can be used as an image annotation class, for text  
123 reports processing, or suggestion of diagnosis. The standard vocabulary is an essential module of  
124 IMPatientT as it interacts with all subsequent modules.

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

129 The tree is generated and rendered with the JavaScript library JSTree (version 3.3.12). Each node  
130 (term) can have only one parent. For each created node (vocabulary terms), the user can assign a  
131 name and organize the tree structure (hierarchy) through the drag and drop interface. Each term  
132 in the tree is associated with nine optional properties. Four properties are defined by the user:  
133 description, list of synonyms, translation in another language, show the term as annotation class.  
134 Two properties are automatically generated: the term's unique identifier (ID) and the  
135 hexadecimal color associated with the term (for image annotation). Additional term properties



136 (associated diagnosis/disease class, associated genes, list of positively correlating terms [*i.e.* co-  
137 occurring terms in reports]) are extracted from patient records registered in the database.

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

## (a) Standard Vocabulary Tree

[Download Vocabulary \(JSON\)](#)

Search

- ▶ ATPase staining
- ▶ Electron Microscopy (EM)
- ▶ Enzymology Miscellaneous
- ▶ HE and TG staining
- ▶ Oxidative staining
  - ▶ Activity Repartition
    - ▶ Core
      - ▶ Cores over the whole Section
      - ▶ Cores with Blurred Boundaries
      - ▶ Cores with Net Limits
      - ▶ Minicores
      - ▶ **Peripheral Cores**
      - ▶ Single Central Core
    - ▶ NADH
    - ▶ Oxy activity: Fibre aspect
    - ▶ Structural Reorganisation

## (b) Vocabulary Properties

Vocabulary ID  
MHO:000124

Vocabulary Name  
Peripheral Cores

Alternative Language  
Core Périphériques x

Synonyms  
Synonyms

Show as Image Annotation Class

Associated HPO Terms (Extracted from reports)

Associated Genes (Extracted from reports)  
HGNC:10483 RYR1 HGNC:1052 BIN1 HGNC:12403 TTN HGNC:129 ACTA1 HGNC:7577 MYH7

Associated Disease (Extracted from reports)  
ORPHA:172976 Congenital myopathy with cores UNCLEAR

Positively Correlates with (Extracted from reports ; >0.5)  
MHO:000124 Peripheral Cores MHO:000125 Single Central Core

Description  
"Peripheral core" refers to areas of reduced oxidative and glycolytic enzymatic activity along the longitudinal axis of skeletal muscle fibers, as seen on enzymatic stains such as NADH

142

143 **Figure 2: Screenshot of the Standard Vocabulary Creator module (module 1).** (a) The  
144 hierarchical structure viewer and editor tool that supports drag and drop modification and  
145 creation/deletion/modification using the mouse. (b) The properties of the selected term node  
146 with its unique ID, display name, alternative language translation, synonyms, description,

147 associated genes and diseases and correlating terms extracted from the application instance  
148 database.

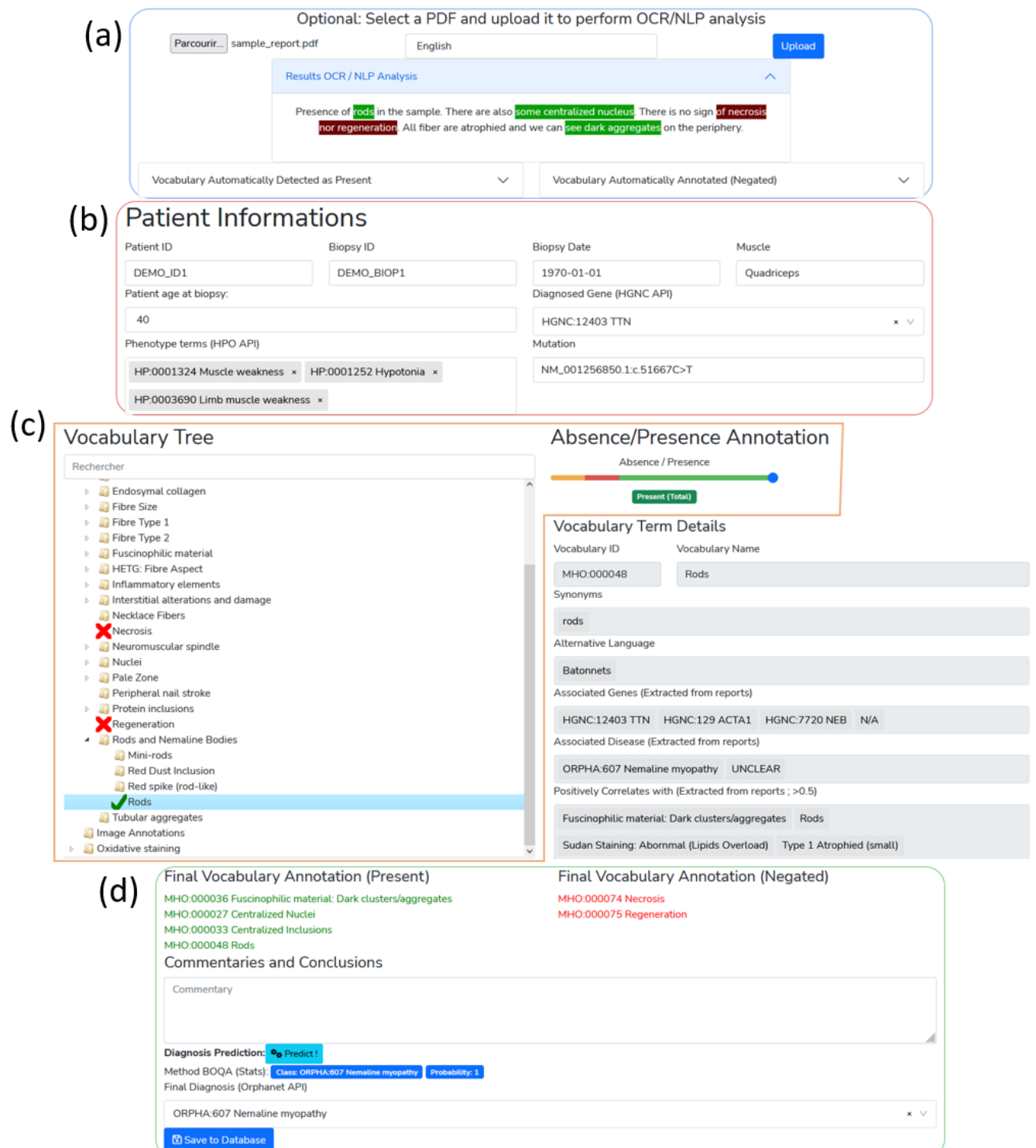
## 149 **Module 2: Report Digitization**

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

154 In the first part of the digitization form (Fig 3a), a PDF file of the free-text report can be uploaded  
155 for natural language processing (NLP) of the content. The text of the PDF report is automatically  
156 extracted and processed with NLP. The NLP method is only used to detect histological terms  
157 defined in the standard vocabulary. Detected standard vocabulary terms are highlighted (see  
158 corresponding section below “Optical Character Recognition and Vocabulary Terms Detection”).  
159 Highlighted terms allow to easily identify what standard vocabulary terms were detected as  
160 present or in negative form. This is useful for quantitative performance assessment.

161 The second part (Fig 3b) of the digitization form contains patient informations, such as patient ID,  
162 document ID, age of the patient. This section also allows the user to input patient information  
163 that are not defined in by the standard vocabulary and thus, not processed in the NLP section.  
164 For example, IMPatientT exploits well-established ontologies to normalize the genetic diagnosis  
165 and phenotypes (Fig 4). For example, in the gene field, when the user input characters, gene  
166 symbols are retrieved from the HUGO Gene Nomenclature Committee (HGNC) and  
167 suggested.[26] Mutation notations are formatted according to the Human Genome Variation  
168 Society (HGVS) sequence variant nomenclature[27]. Phenotypes are retrieved and suggested  
169 using the HPO ontology. None of these fields contain patient-identifying data and are optional.

170



171

172

173

174

175

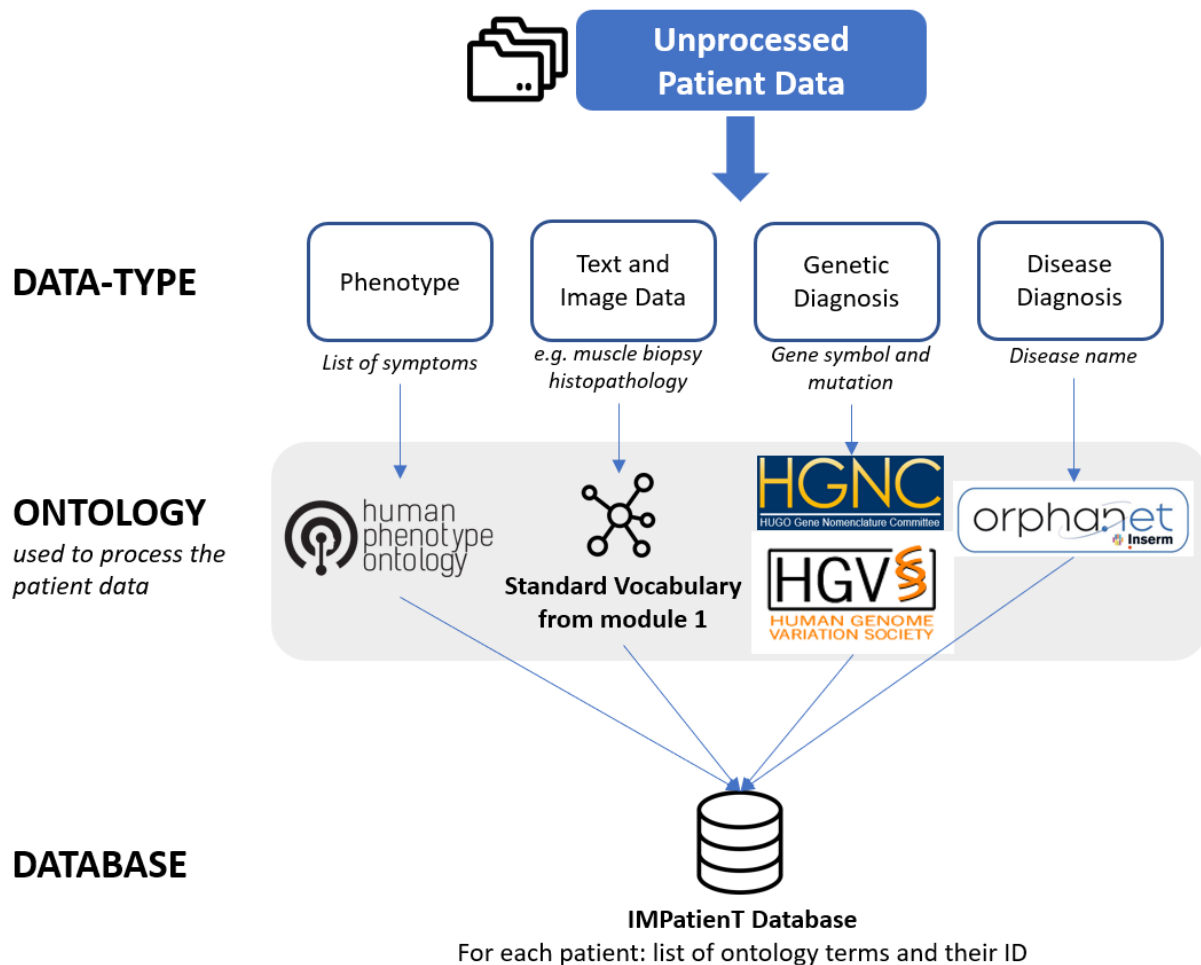
176

177

178

179

**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.



180

For each patient: list of ontology terms and their ID

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

183

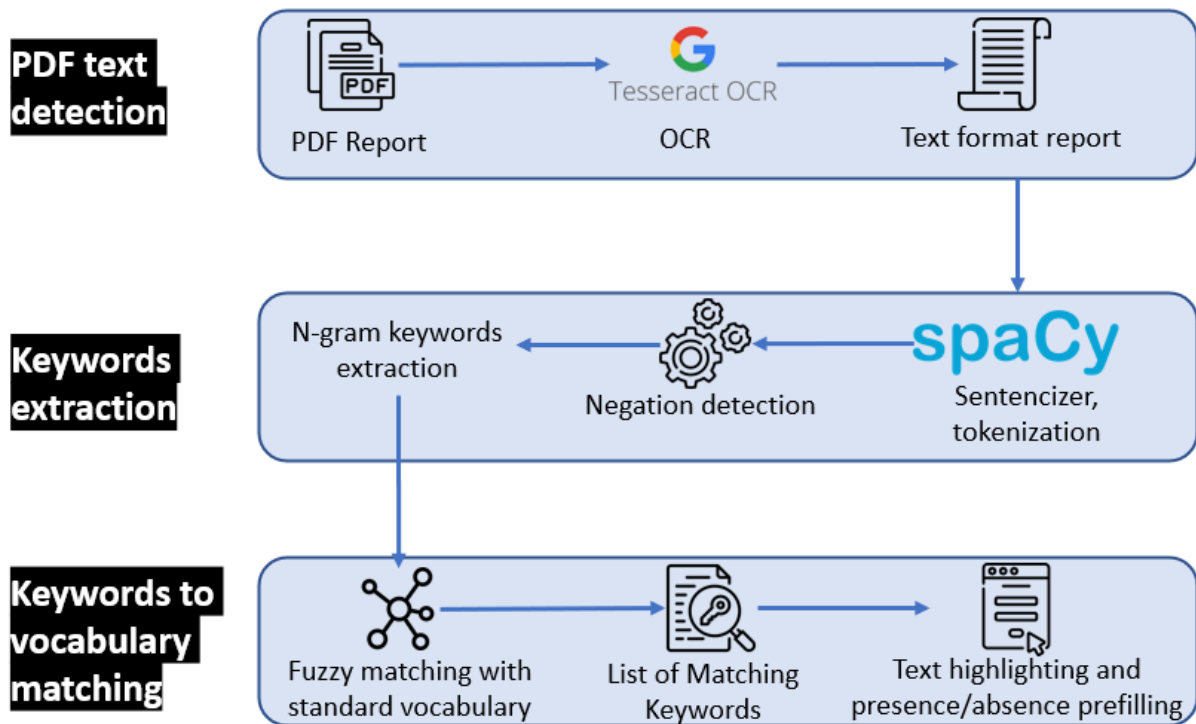
184 The third part of the digitization form (Fig 3c) contains the standard vocabulary tree viewer with  
 185 an absence/presence slider. This section allows the user to correct the automatic detection of the  
 186 NLP method or to add new observations. Each vocabulary term can be marked as present,  
 187 absent or no information. For terms marked as present, the slider is used to indicate a notion of  
 188 quantity or certainty of the term. For example, the statement "There is a small number of fibers  
 189 containing rods" can be annotated by hand by setting the vocabulary "Rods" to the value  
 190 "Present" with a low quantity value. For terms that have been automatically detected, this slider  
 191 value is automatically set to 0 (present in a negated sentence) or 1 (present).

192 Finally, the fourth part (Fig 3d) of the form allow the user to input comments and a final diagnosis  
193 for the patient, disease name are suggested from the Orphanet [28] knowledge base. It also  
194 includes an automatic suggestion of the diagnosis based on already registered patients using  
195 BOQA [20] (see the corresponding section below "Patient Disease Suggestions Method").

## 196 **Optical Character Recognition and Vocabulary Term Detection**

197 The patient report digitization in module 2 is facilitated by the automatic text recognition and  
198 keyword detection method. The user uploads a PDF version of the text reports to perform Optical  
199 Character Recognition (OCR), followed by Natural Language Processing (NLP) to automatically  
200 detect terms from the standard vocabulary in the report. The NLP method is only match the raw  
201 text to the standard vocabulary defined in Standard Vocabulary Module 1. Figure 5 describes the  
202 workflow of the vocabulary terms detection method. First the PDF file is converted to plain text  
203 using the Tesseract OCR (implemented in python as pyTesseract). Then, the text is processed  
204 with Spacy, an NLP python library, by splitting the text into sentences and then into individual  
205 words. The resulting list of sentences is then processed to detect negation using a simple  
206 implementation of the concept of NegEx [29]. An  $n$ -gram (monograms, digrams, and trigrams)  
207 procedure is applied

208 to the list of words to identify contiguous words in the context of all the sentences of the report.  
209 The  $n$ -grams are then mapped against the user-created standard vocabulary using fuzzy partial  
210 matching (using Levenshtein distance) with a score threshold of 0.8. Matched keywords are kept  
211 and shown on the interface with a green or red highlight of the detected text using Mark.JS  
212 JavaScript library (green indicates the presence of the keyword, red indicates the presence in a  
213 negated sentence). Keywords are also automatically marked as present or absent (negated) in  
214 the vocabulary tree.



215

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

218

## 219 Disease Suggestions

220 The report digitization module 2 contains a disease recommendation algorithm inspired by the  
 221 BOQA algorithm described by Bauer *et al.* [20]. Basically, the algorithm computes the similarity  
 222 between a list of input vocabulary terms annotated as “present” for a patient (the query) and a  
 223 simulated patient profile for each disease class (model report) that is generated based on the  
 224 data from already registered patients.

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

230 generated model report for disease A will have a probability=0.9 of containing vocabulary term B  
231 and a probability=0.1 of containing vocabulary term C.

232 Due to the stochastic nature of the generation of the model report, for any given prediction, the  
233 generation and computation of the similarity with the query is repeated 50 times. For each  
234 repetition, if a disease has a prediction probability>0.5, it is considered to be the best prediction,  
235 otherwise the prediction is “no prediction”. Finally, of the 50 repetitions, the prediction with the  
236 highest occurrence is taken as the final prediction.

### 237 **Module 3: AI-Assisted Image Annotation Using Automatic Segmentation**

238 To process patient image data, we developed the image annotation module (module 3) to  
239 upload, annotate and perform image segmentation with standard vocabulary terms. This module  
240 is based on the *“interactive image segmentation with Dash and Scikit-image”* demonstration  
241 application [30–32]. The original source code was modified to be compatible with the standard  
242 vocabulary tree and the database.

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

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



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

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



261  
262 **Figure 6: Screenshot of the image annotation module. (a)** Image viewer used to navigate,  
263 zoom and annotate the histology image. **(b)** Menu interface to select the annotation label, brush  
264 width and segmentation parameters.

## 265 **Module 4: Automatic Visualization Dashboard**

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

## 269 **Interaction Between the Modules**

270 IMPatientT is divided into four modules that are interconnected. The standard vocabulary module  
271 provides the vocabulary used for the image annotation module and for the NLP method used for  
272 the (histologic) standard vocabulary terms detection in the report digitization module. Any  
273 modification in the vocabulary is automatically propagated to these modules, updating the form  
274 templates and triggering the recalculation of all visualizations with the latest vocabulary  
275 information. Any modification to the standard vocabulary also updates all patients in the  
276 database to the latest version of the vocabulary, meaning that term names and definitions will be  
277 updated, and deleted terms will be marked as outdated. Adding patient information in the  
278 database, whether they are text reports (module 2) or images data (module 3), will automatically  
279 update the visualization dashboard with the latest patient information in the database. The term  
280 frequency statistics calculated by the visualization dashboard and used by the disease suggestion  
281 algorithm are automatically updated as well, providing live performances increase. The  
282 visualization dashboard is also directly linked to the standard vocabulary and during the  
283 generation of the visualizations, the rich definition of the standard terms is updated with newly  
284 associated genes, diagnosis and positively correlating terms.

## 285 **Application Security and Personal Data**

286 IMPatientT is developed as a free and open-source project meaning that the code can be audited  
287 by anyone in the GitHub code repository. The code is regularly scanned for known issues and  
288 outdated libraries to mitigate security issues. There is no patient-identifying data kept in the  
289 database, only a custom identifier and age. No name or date of birth are required or stored.

290 Additionally, access to all modules and data entered via the web application is restricted by a  
291 login-page and user accounts can only be created by the administrator of the platform. No user  
292 information is stored except for the username, email and salted and hashed passwords.

## 293 **RESULTS**

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

### 297 **IMPatientT Main Functionalities**

298 Table 1 shows the main functionalities of IMPatientT compared to other similar tools used in the  
299 community. IMPatientT integrates tools that are simple, portable, easy to implement and similar  
300 to multiple state-of-the-art solutions but in a single platform. Out of 18 selected features,  
301 IMPatientT integrates 14 of them versus a mean of 4.4 for other software with the best one being  
302 SAMS and PhenoStore integrating 6 features each. However, software such as SAMS, PhenoStore,  
303 Phenotips and Cytomine each integrates features that are not yet present in IMPatientT.

304 IMPatientT implements novel functionalities to process and exploit patient data. For example,  
305 IMPatientT is compatible with any domain of research thanks to its standard vocabulary builder.  
306 Also, with the OCR/NLP method, IMPatientT can process histologic text reports, allowing the user  
307 to exploit scanned documents. Finally, IMPatientT also provides useful utilities to exploit patient  
308 data with the various visualizations, the term, frequency table, correlation matrix and the  
309 automatic enrichment of the vocabulary terms definition (associated genes and diseases).

310 **Table 1: Comparison of functionalities from IMPatientT compared to common state-of-the-art tools.**

| Group                               | Functionalities                                   | IMPatientT | Phenotips | PhenoStore | SAMS | Protégé | Doc2HPO | Cytomine | Ilastik | INTEGRO |
|-------------------------------------|---|------------|-----------|------------|------|---------|---------|----------|---------|---------|
| General Application Characteristics | Web application                                   | ✓          | ✓         | ✓          | ✓    |         | ✓       | ✓        |         |         |
|                                     | Patient database                                  | ✓          | ✓         | ✓          | ✓    |         |         |          |         |         |
|                                     | Free to use and open-source                       | ✓          |           |            | ✓    | ✓       | ✓       | ✓        | ✓       | ✓       |
|                                     | Support multimodal data                           | ✓          |           |            |      |         |         |          |         |         |
|                                     | Support for patient pedigree data                 |            | ✓         | ✓          |      |         |         |          |         |         |
| Standard Vocabulary                 | Vocabulary Builder                                | ✓          |           |            |      | ✓       |         | ✓        |         |         |
|                                     | Advanced vocabulary terms definition              | ✓          |           |            |      | ✓       |         |          |         |         |
|                                     | Full-featured ontology builder                    |            |           |            |      | ✓       |         |          |         |         |
| Report digitization                 | Integrates reference ontologies (HPO, Orphanet)   | ✓          | ✓         |            | ✓    |         | ✓       |          |         | ✓       |
|                                     | Form for patient medical report digitization      | ✓          |           | ✓          | ✓    |         |         |          |         |         |
|                                     | Text recognition with OCR                         | ✓          |           |            |      |         |         |          |         |         |
|                                     | Text processing with NLP                          | ✓          |           |            |      |         | ✓       |          |         |         |
|                                     | Export data to Phenopacket format                 |            |           | ✓          | ✓    |         |         |          |         |         |
| Image annotation                    | Image annotation and segmentation with AI         | ✓          |           |            |      |         |         | ✓        | ✓       |         |
|                                     | Support for DICOM and whole slide images          |            |           |            |      |         |         | ✓        |         |         |
| Patient data exploitation           | Automatic visualization dashboard                 | ✓          |           | ✓          |      |         |         |          |         |         |
|                                     | Diagnosis prediction system                       | ✓          | ✓         |            |      |         |         |          |         |         |
|                                     | Data mining of information for specific diagnosis | ✓          |           |            |      |         |         |          |         | ✓       |

311  
312

## 313 **IMPatient Usage**

314 Figure 1 shows how the user can interact with the web application to digitize, process, and  
315 explore patient data. In IMPatient, modules can be used independently, allowing users to only  
316 use the tools they need. For example, a user might only have text report data, in this case they  
317 would be able to use the standard vocabulary creator, the report digitization tools and the  
318 visualization dashboard to process and explore their data. In another scenario, a user could only  
319 be interested in annotating an image dataset using a shared standard vocabulary that can be  
320 modified and updated collaboratively. In this use case, they would be able to only use the  
321 standard vocabulary creator and the image annotation module. However, the main strength of  
322 IMPatient lies in the multimodal approach it provides and the module interactions.

323 For the complete multimodal approach, the first step is to create a standard vocabulary using the  
324 Standard Vocabulary Creator interface (module 1). The user only needs to create a few terms  
325 (nodes) to begin using the web application. Defining the properties of the terms (definition,  
326 synonyms...) is optional, and organizing them in a hierarchical structure is also optional.

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

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

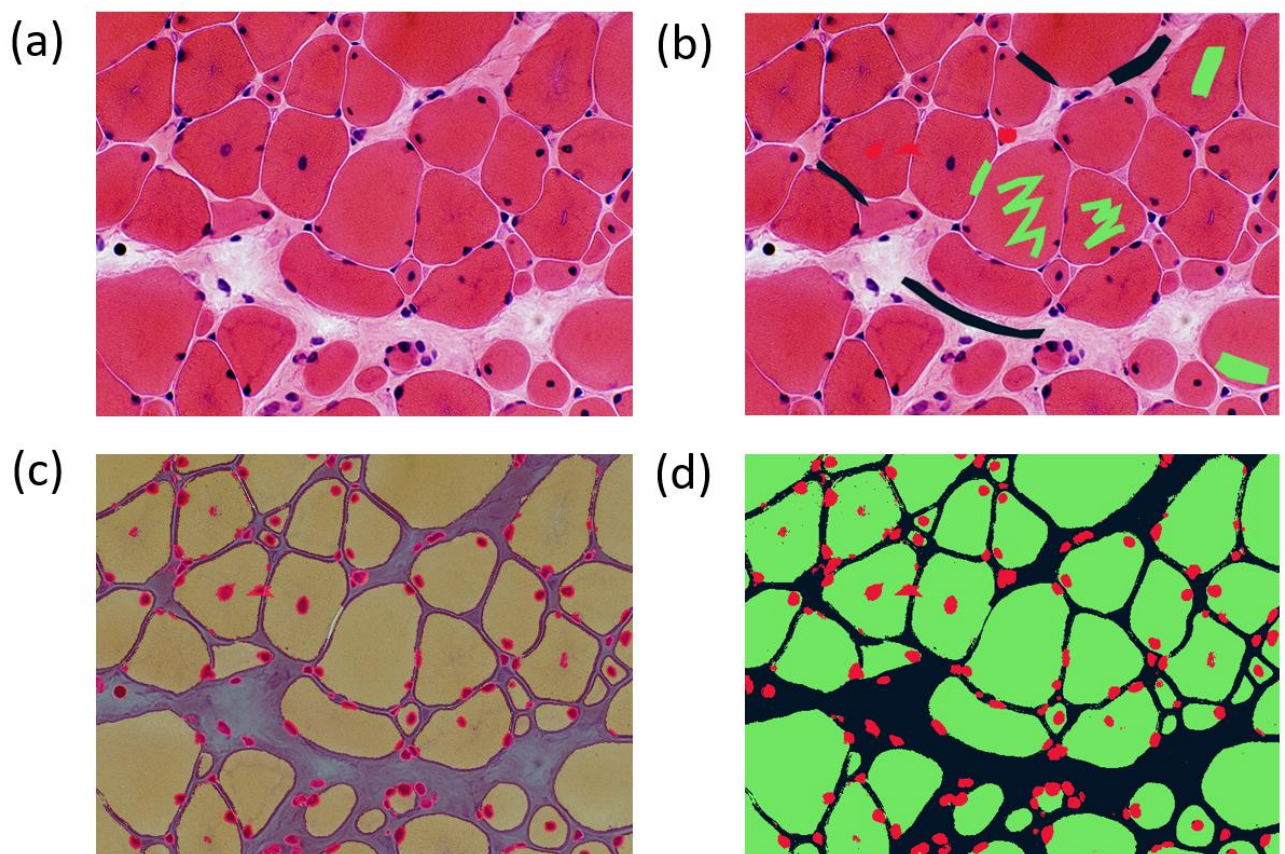
## 338 **Use Case: Congenital Myopathy Histology Reports**

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

346 For text data, Supplementary Figure S1 show the results of the automatic NLP method applied to  
347 an artificial muscle histology report. Twenty-two keywords were detected and match to the  
348 standard vocabulary and seven of them were detected in negated sentences (red highlight).  
349 Among the 22 vocabulary terms detected. Out of the twenty-two keywords, eighteen were  
350 correctly detected and one was detected in the wrong state of negation: “abnormal fiber  
351 differentiation” is highlighted as negated while it is present is a non-negated sentence part. Three  
352 keywords (fiber type, internalized nuclei, centralized nuclei) were detected as matching for  
353 multiple keywords from the vocabulary at the same time due to high similarity. For example, the  
354 keyword “internalized nuclei” and “centralized nuclei” have a similarity score of 86 using the  
355 Levenstein distance. Two keywords defined in the standard vocabulary were missed and not  
356 highlighted: “biopsy looks abnormal” (“abnormal biopsy” in the vocabulary) and “purplish shade”  
357 (“purplish aspect” in the vocabulary).

358 For the image data, figure 7 shows an example of the segmentation of a biopsy image, where we  
359 annotated the cytoplasm of the cells (green), intercellular spaces (black) and cell nuclei (red). The  
360 raw image (Fig 7a) is annotated with free-shape areas associated with standard vocabulary terms  
361 (Fig 7b). Then, the whole image is automatically segmented based on the annotations, producing  
362 the segmentation mask where each pixel is associated with a class (Fig 7c 7d).

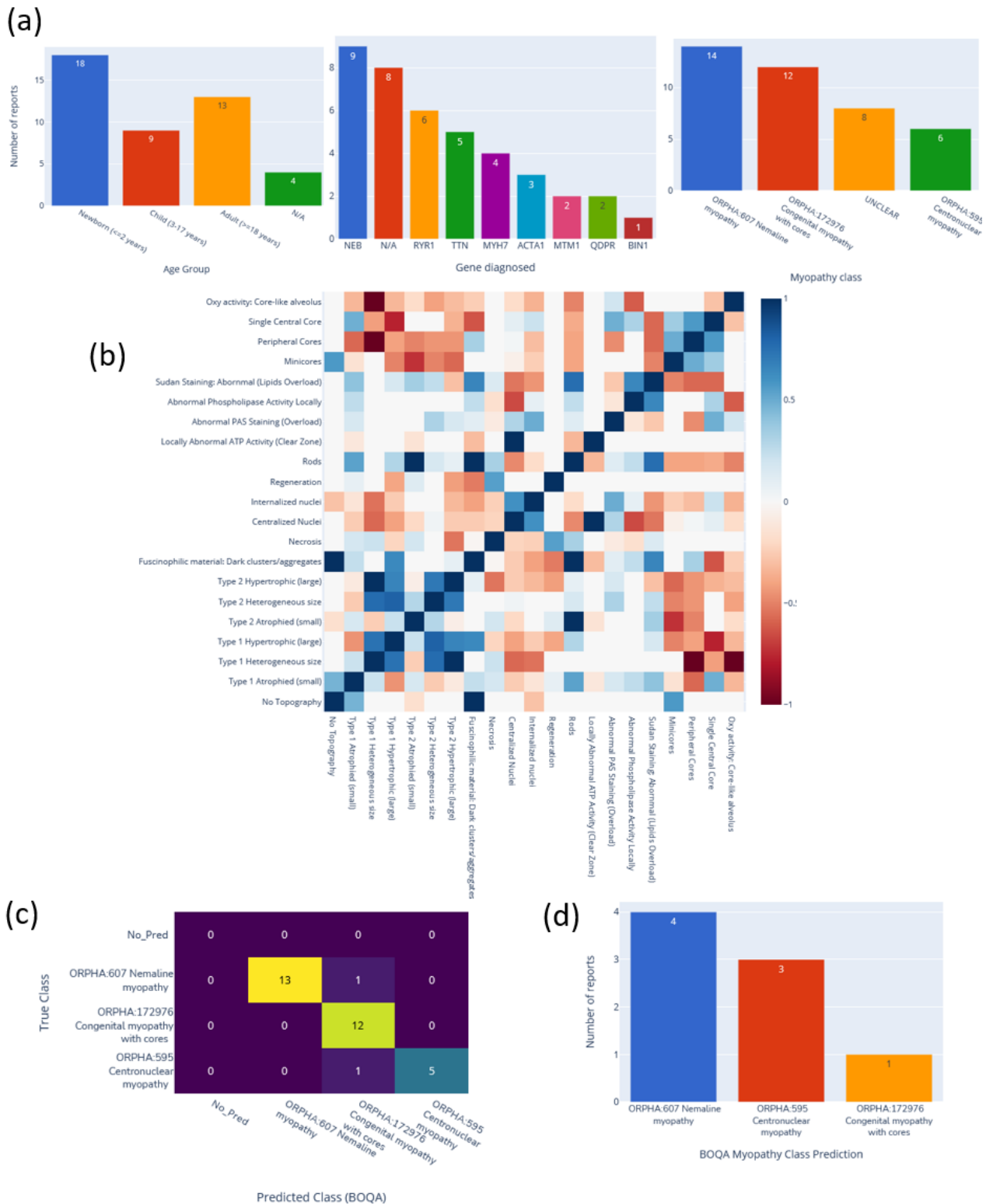
363 The automatic visualization dashboard was used to generate the six visualizations provided in  
364 figure 8. These visualizations include a breakdown of the patients in the database by age, genes,  
365 or diagnosis (Fig 8a). A correlation matrix (using Pearson correlation coefficient) between the  
366 occurrence of standard vocabulary terms is generated (Fig 8b), which can serve as a starting  
367 point for exploration of co-occurrence of features in patients. The confusion matrix of the final  
368 diagnosis of patients versus the suggested diagnosis with BOQA (Fig 8c) allows the user to  
369 monitor the accuracy of the disease suggestion function. In addition, a histogram showing the  
370 classification of patients without a final diagnosis is provided to indicate possible prognosis of  
371 undiagnosed patients (Fig 8d). Finally, the frequency of each standard vocabulary term by gene  
372 and by disease is automatically calculated and shown in two tables (see supplementary tables S2  
373 and S3).



375 **Figure 7: Image segmentation process in the image segmentation module. (a)** Raw image  
376 input before annotation. **(b)** Image with limited manual annotation of cytoplasm (green), cell  
377 nucleus (red) and intercellular space (black). **(c)** Blended image of the raw image and segmented

378 image after automated segmentation with a random-forest classifier. **(d)** Segmented image mask  
379 alone.





380

381 **Figure 8: Automatic visualization of 40 generated congenital myopathy reports.** (a)  
 382 Histogram of the number of reports by age group, by diagnosed gene (top 9) or by congenital  
 383 myopathy class. (b) Correlation matrix of standard vocabulary terms after annotation for all  
 384 reports. (c) Confusion matrix of BOQA algorithm performance for suggestion of the three main  
 385 congenital myopathy classes (NM, COM, CNM, n=32). Colors indicate the number of reports for  
 386 each cell of the matrix, the lighter the color the more reports. (d) Histogram of the reclassification  
 387 by BOQA of reports without a final diagnostic (n=8).

## 388 **DISCUSSION**

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

395 The semi-automatic approach implemented in IMPatientT offers faster digitization processes  
396 while ensuring accuracy through manual review. This is achieved by analyzing text data using  
397 OCR and NLP to automatically match the text to the standard vocabulary, followed by manual  
398 correction. For image data, the user first provides sparse annotations on the image, which are  
399 then used to compute an automatic segmentation of the whole image. For data exploration,  
400 IMPatientT uses a fully automatic approach including various visualizations as well as diagnosis  
401 suggestions, while allowing the user to extract the processed data in a standard format for  
402 further analysis (database, images, frequency tables).

403 IMPatientT aims to integrate multiple approaches in a unified platform with two main objectives:  
404 universality (*i.e.* not restricted to a specific domain) and multimodality (*i.e.* integration of multiple  
405 data types). To our knowledge, other tools similar to IMPatientT do not fulfill both objectives.

406 We performed a comparison of the main functionalities of IMPatientT with other tools used in the  
407 community. Phenotips, SAMS and PhenoStore are similar to IMPatientT as they are designed as a  
408 patient information database. However, they are restricted to processing patient phenotype data  
409 by using HPO and do not integrate multimodal data. IMPatientT goes further by allowing for  
410 custom observations with the vocabulary builder and with automatic digitization with OCR/NLP  
411 as well as integrating tools to exploit image data.

412 Other tools are similar to one or two modules only of IMPatientT. For example, Doc2HPO is a tool  
413 that also uses a semi-automatic approach to digitize clinical text according to a list of HPO terms,  
414 based on NLP methods and negation detection. However, as Doc2HPO is also restricted to HPO,  
415 it does not provide custom vocabulary tree facilities. In contrast IMPatientT is suitable for  
416 digitization of text data from any domain of interest.

417 For image data, software such as Cytomine and Ilastik are widely used and perform well on  
418 biological data, but they do not allow the user to take into consideration the multimodal aspects  
419 of patient data by keeping the raw image and the expert interpretation (histological report) in a  
420 single database along with a collaborative and rich-defined custom ontology.

421 Finally, in IMPatientT we reimplemented the diagnosis suggestion algorithm called BOQA that is  
422 also used in Phenomizer, a tool to rank a list of the top matching diseases based on a list of input  
423 HPO terms. We modified the algorithm to consider frequencies of terms by disease to have  
424 meaningful predictions. However, BOQA uses binary states for terms (terms are marked as  
425 present or absent) and is not compatible with numeric features. In the future, it will be necessary  
426 to implement a more complex system such as explainable AI with learning classifier systems [33].  
427 This should improve accuracy, explainability, and handling of quantitative values, although at the  
428 cost of computational power.

429 IMPatientT still lacks some feature compared to other tools, such as a pedigree editor, support for  
430 DICOM and gigapixel images and phenotypic data export to the Phenopacket format. In the  
431 future, we plan to further develop IMPatientT by adding these features to the interface. We also  
432 want to explore the automatization of the standard vocabulary creation with the analysis of a  
433 complete corpus of text. For text analysis, we wish to implement additional context  
434 comprehension, *i.e.* not only negation but also hypothetical statements, uncertainty and family  
435 context as well as better text-vocabulary terms matching. Finally, we plan to expand the scope of

436 the OCR/NLP method by integrating existing NLP tools to automatically detect HPO terms, gene  
437 symbols and disease name the report text.

## 438 **CONCLUSIONS**

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

## 447 **Source-code and Data Availability**

448 The source-code for IMPatientT v1.5.0 is available in its GitHub repository  
449 (<https://github.com/lambda-science/IMPatientT>). The datasets generated and analyzed during the  
450 current study are also available in the same repository.

## 451 **Conflicts of Interest**

452 The authors declare that they have no conflict of interest.

## 453 **Funding Statement**

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

## 456 Acknowledgements

457 We thank the BiGEst-ICube platform for their assistance.

## 458 Supplementary Materials

- 459 • **Figure S1** - nlp\_qualitative\_results.pptx - **Qualitative assessment of the**  
460 **performances of the NLP method matching text to the standard vocabulary. (a)**  
461 Raw muscle histology report text with detected keywords highlighted in green and red. A  
462 red highlight indicated that the keyword is in a negated sentence. **(b)** Table of some  
463 highlighted keywords and the details of the match (matching vocabulary ID and term,  
464 position in the raw text, matching n-gram [raw text] and the similarity score of the  
465 comparison). Green and red colors correspond to keywords detected as present and  
466 present in negated sentence respectively.
- 467 • **Table S2** - table\_frequencies\_per\_gene.csv - **Table of frequencies of standard**  
468 **vocabulary per genes.** This CSV file contains all frequencies of standard vocabulary  
469 terms for each gene with the total number of reports per gene and the number of  
470 occurrences of each term if not 0.
- 471 • **Table S3** - table\_frequencies\_per\_diag.csv - **Table of frequencies of standard**  
472 **vocabulary per diagnosis.** This CSV file contains all frequencies of standard vocabulary  
473 terms for each diagnosis with the total number of reports per diagnosis and the number  
474 of occurrences of each term if not 0.

## 475 References

- 476 [1] Kerr WT, Hwang ES, Raman KR, Barritt SE, Patel AB, Le JM, et al. Multimodal diagnosis of  
477 epilepsy using conditional dependence and multiple imputation. 2014 Int. Workshop  
478 Pattern Recognit. Neuroimaging, 2014, p. 1–4. <https://doi.org/10.1109/PRNI.2014.6858526>.
- 479 [2] Yan R, Ren F, Rao X, Shi B, Xiang T, Zhang L, et al. Integration of Multimodal Data for Breast  
480 Cancer Classification Using a Hybrid Deep Learning Method. In: Huang D-S, Bevilacqua V,

- 481 Premaratne P, editors. *Intell. Comput. Theor. Appl.*, Cham: Springer International Publishing;  
482 2019, p. 460–9. [https://doi.org/10.1007/978-3-030-26763-6\\_44](https://doi.org/10.1007/978-3-030-26763-6_44).
- 483 [3] Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial  
484 intelligence for diagnosis and prognosis of early stages of Alzheimer’s disease. *Transl Res J*  
485 *Lab Clin Med* 2018;194:56–67. <https://doi.org/10.1016/j.trsl.2018.01.001>.
- 486 [4] Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for  
487 early detection of Alzheimer’s disease stage. *Sci Rep* 2021;11:3254.  
488 <https://doi.org/10.1038/s41598-020-74399-w>.
- 489 [5] North KN, Wang CH, Clarke N, Jungbluth H, Vainzof M, Dowling JJ, et al. Approach to the  
490 diagnosis of congenital myopathies. *Neuromuscul Disord NMD* 2014;24:97–116.  
491 <https://doi.org/10.1016/j.nmd.2013.11.003>.
- 492 [6] Cassandrini D, Trovato R, Rubegni A, Lenzi S, Fiorillo C, Baldacci J, et al. Congenital  
493 myopathies: clinical phenotypes and new diagnostic tools. *Ital J Pediatr* 2017;43:101.  
494 <https://doi.org/10.1186/s13052-017-0419-z>.
- 495 [7] Böhm J, Vasli N, Malfatti E, Le Gras S, Feger C, Jost B, et al. An integrated diagnosis strategy  
496 for congenital myopathies. *PloS One* 2013;8:e67527.  
497 <https://doi.org/10.1371/journal.pone.0067527>.
- 498 [8] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical  
499 terminology. *Nucleic Acids Res* 2004;32:D267–70. <https://doi.org/10.1093/nar/gkh061>.
- 500 [9] Köhler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The  
501 Human Phenotype Ontology in 2021. *Nucleic Acids Res* 2021;49:D1207–17.  
502 <https://doi.org/10.1093/nar/gkaa1043>.
- 503 [10] Musen MA. The Protégé Project: A Look Back and a Look Forward. *AI Matters* 2015;1:4–12.  
504 <https://doi.org/10.1145/2757001.2757003>.
- 505 [11] Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient  
506 and accurate HPO concept curation. *Nucleic Acids Res* 2019;47:W566–70.  
507 <https://doi.org/10.1093/nar/gkz386>.
- 508 [12] Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: Patient  
509 Phenotyping Software for Clinical and Research Use. *Hum Mutat* 2013;34:1057–65.  
510 <https://doi.org/10.1002/humu.22347>.
- 511 [13] Steinhaus R, Proft S, Seelow E, Schalau T, Robinson PN, Seelow D. Deep phenotyping:  
512 symptom annotation made simple with SAMS. *Nucleic Acids Res* 2022:gkac329.  
513 <https://doi.org/10.1093/nar/gkac329>.
- 514 [14] Laurie S, Piscia D, Matalonga L, Corvó A, Fernández-Callejo M, Garcia-Linares C, et al. The  
515 RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and  
516 gene discovery for rare diseases. *Hum Mutat* 2022;43:717–33.  
517 <https://doi.org/10.1002/humu.24353>.
- 518 [15] Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis  
519 of multi-gigapixel imaging data using Cytomine. *Bioinformatics* 2016;32:1395–401.  
520 <https://doi.org/10.1093/bioinformatics/btw013>.
- 521 [16] Aubreville M, Bertram C, Klopffleisch R, Maier A. SlideRunner - A Tool for Massive Cell  
522 Annotations in Whole Slide Images. *ArXiv180202347 Cs* 2018:309–14.  
523 [https://doi.org/10.1007/978-3-662-56537-7\\_81](https://doi.org/10.1007/978-3-662-56537-7_81).

- 524 [17] Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, et al. ilastik: interactive  
525 machine learning for (bio)image analysis. *Nat Methods* 2019;16:1226–32.  
526 <https://doi.org/10.1038/s41592-019-0582-9>.
- 527 [18] Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate  
528 genes for human diseases. *Nat Methods* 2015;12:841–3.  
529 <https://doi.org/10.1038/nmeth.3484>.
- 530 [19] Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human  
531 genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 2009;85:457–64.  
532 <https://doi.org/10.1016/j.ajhg.2009.09.003>.
- 533 [20] Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and  
534 noise-tolerant semantic searches. *Bioinformatics* 2012;28:2502–8.  
535 <https://doi.org/10.1093/bioinformatics/bts471>.
- 536 [21] Cinaglia P, Tradigo G, Cascini GL, Zumpano E, Veltri P. A framework for the decomposition  
537 and features extraction from lung DICOM images. *Proc. 22nd Int. Database Eng. Appl.*  
538 *Symp., New York, NY, USA: Association for Computing Machinery; 2018, p. 31–6.*  
539 <https://doi.org/10.1145/3216122.3216127>.
- 540 [22] Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen  
541 extracts and prioritizes patient phenotypes directly from medical records to expedite  
542 genetic disease diagnosis. *Genet Med* 2019;21:1585–93. [https://doi.org/10.1038/s41436-](https://doi.org/10.1038/s41436-018-0381-1)  
543 [018-0381-1](https://doi.org/10.1038/s41436-018-0381-1).
- 544 [23] Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-  
545 generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*  
546 2015;10:2004–15. <https://doi.org/10.1038/nprot.2015.124>.
- 547 [24] Cinaglia P, Guzzi PH, Veltri P. INTEGRO: an algorithm for data-integration and disease-gene  
548 association. 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM, 2018, p. 2076–81.  
549 <https://doi.org/10.1109/BIBM.2018.8621193>.
- 550 [25] H J, S T, F Z, A S, J O, C S, et al. Congenital myopathies: disorders of excitation-contraction  
551 coupling and muscle contraction. *Nat Rev Neurol* 2018;14.  
552 <https://doi.org/10.1038/nrneurol.2017.191>.
- 553 [26] Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC  
554 and VGNC resources in 2021. *Nucleic Acids Res* 2021;49:D939–46.  
555 <https://doi.org/10.1093/nar/gkaa980>.
- 556 [27] den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al.  
557 HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum*  
558 *Mutat* 2016;37:564–9. <https://doi.org/10.1002/humu.22981>.
- 559 [28] INSERM. Orphanet: an online database of rare diseases and orphan drugs 1997.  
560 <http://www.orpha.net> (accessed February 13, 2022).
- 561 [29] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for  
562 Identifying Negated Findings and Diseases in Discharge Summaries. *J Biomed Inform*  
563 2001;34:301–10. <https://doi.org/10.1006/jbin.2001.1029>.
- 564 [30] Gouillart E. Interactive Machine Learning - Image segmentation. GitHub 2020.  
565 <https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation>  
566 (accessed November 23, 2021).

- 567 [31] Walt S van der, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-  
568 image: image processing in Python. PeerJ 2014;2:e453. <https://doi.org/10.7717/peerj.453>.
- 569 [32] Hossain S. Visualization of Bioinformatics Data with Dash Bio. Proc 18th Python Sci Conf  
570 2019:126–33. <https://doi.org/10.25080/Majora-7ddc1dd1-012>.
- 571 [33] Urbanowicz RJ, Moore JH. ExSTraCS 2.0: Description and Evaluation of a Scalable Learning  
572 Classifier System. Evol Intell 2015;8:89–116. <https://doi.org/10.1007/s12065-015-0128-8>.

573