



HAL
open science

IMPatientT: 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.. IMPatientT: an Integrated web application to digitize, process and explore Multimodal PATIENT daTa. *Journal of Neuromuscular Diseases*, 2024, Online ahead of print. 10.3233/JND-230085 . hal-03635350v3

HAL Id: hal-03635350

<https://hal.science/hal-03635350v3>

Submitted on 17 Jun 2024

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 - NonCommercial 4.0 International License

Research Article

IMPatientT: An Integrated Web Application to Digitize, Process and Explore Multimodal PATIENT data

Corentin Meyer^a, Norma Beatriz Romero^b, Teresinha Evangelista^b, Brunot Cadot^a, Jocelyn Laporte^d, Anne Jeannin-Girardon^a, Pierre Collet^a, Ali Ayadi^a, Kirsley Chennen^a and Olivier Poch^a

^a*Complex Systems and Translational Bioinformatics (CSTB), ICube Laboratory, UMR 7357, University of Strasbourg, 1 rue Eugène Boeckel, Strasbourg, France*

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

^c*Sorbonne Université, INSERM, Center for Research in Myology, Myology Institute, GHU Pitié-Salpêtrière, Paris, France*

^d*Department Translational Medicine, IGBMC, CNRS UMR 7104, 1 rue Laurent Fries, Illkirch, France*

Accepted 23 March 2024

Abstract. Medical acts, such as imaging, lead to the production of various medical text reports that describe the relevant findings. This induces multimodality in patient data by combining image data with free-text and consequently, multimodal data have become central to drive research and improve diagnoses. However, the exploitation of patient data is problematic as the ecosystem of analysis tools is fragmented according to the type of data (images, text, genetics), the task (processing, exploration) and domain of interest (clinical phenotype, histology). To address the challenges, we developed IMPatientT (Integrated digital Multimodal PATIENT data), a simple, flexible and open-source web application to digitize, process and explore multimodal patient data. IMPatientT has a modular architecture allowing to: (i) create a standard vocabulary for a domain, (ii) digitize and process free-text data, (iii) annotate images and perform image segmentation, (iv) generate a visualization dashboard and provide diagnosis decision support. To demonstrate the advantages of IMPatientT, we present a use case on a corpus of 40 simulated muscle biopsy reports of congenital myopathy patients. As IMPatientT provides users with the ability to design their own vocabulary, it can be adapted to any research domain and can be used as a patient registry for exploratory data analysis. A demo instance of the application is available at <https://impatient.lbgi.fr/>.

Keywords: Muscular diseases, histology, image processing, computer-assisted, diagnosis, computer-assisted, electronic health records, artificial intelligence

INTRODUCTION

Patient data now incorporates the results of numerous technologies, including imaging, next-generation sequencing and more recently wearable devices. Furthermore, medical acts such as echography, radiology or histology, produce imaging data that are gener-

*Correspondence to: Olivier Poch, CSTB – ICube UMR 7357, CRBS, 1 rue Eugène Boeckel, 67000 Strasbourg, Tel.: +33 3 3 68 85 32 95; E-mail: olivier.poch@unistra.fr.

ally combined with medical reports to describe the relevant findings. Thus, multimodality is induced in patient data, as imaging data is inherently linked to free-text reports. The link between image and report data is crucial as raw images can be re-interpreted during the patient's medical journey with new domain knowledge or by different experts leading to complementary reports. 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 (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. First, multimodal patient data needs to be processed in an integrated way to preserve this link in a single database or data warehouse. Second, useful tools to process and explore multimodal data are essential to drive research and improve diagnosis.

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

Clinical reports are generally written using free-text, and therefore 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, including Prote'ge' [10]. Text mining processes have been developed that exploit these standard vocabularies to automatically detect important keywords in free-text data. For example, Doc2HPO [11] can extract a list of HPO terms from free-text medical records. Other software packages, *e.g.* Phenotips [12], have been developed to centralize and process general patient information, including demographics, pedigree, common measurements, phenotypes and genetic results. SAMS [13] and RD-Connect PhenoStore [14] are further

examples of web applications that aim to perform deep phenotyping of patients by building a single database of standardized patient data using well-established ontologies such as HPO.

A number of tools have been developed to analyze and explore patient data, based on a list of HPO terms describing a patient's specific phenotypic profile. For example, 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].

Finally, for imaging data, software to process and annotate gigapixel scale microscopy images are widely used, including Cytomine [18], SlideRunner [19] and Ilastik [20]. While Cytomine incorporates an ontology builder and complex image processing tools, it is restricted to image data only. For exploitation of patient images, guidelines and frameworks have been proposed to standardize the measurement of pathological features for example from DICOM lung images [21]. Some multimodal approaches such as ClinPhen [22] and Exomiser [23] have successfully combined multiple levels of information with both phenotype information (HPO terms) and genetic information (variants) to rank candidate genes in Mendelian diseases. Other tools such as INTEGRO [24] have been developed to automatically mine disease-gene associations for a specific input disease from multiple curated sources of knowledge.

This large ecosystem of tools highlights the need for an integrated tool that can: (i) 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 (Integrated digital Multimodal PATIENT daTa)**, a free and open-source web application designed to provide an integrated tool to digitize, process and explore multimodal patient data. **IMPatient** is a turnkey solution that can aggregate patient data and provides simple tools and interfaces allowing clinicians to easily extract information from multimodal patient data. **IMPatient** is based on a modular architecture, and currently incorporates 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)

138 annotate and segment images with standard vocab-
139 ulary, and (iv) generate a dashboard with automatic
140 visualizations to explore the patient data and perform
141 automatic diagnosis suggestions.

142 We demonstrate the usefulness of IMPatienT on
143 a set of congenital myopathy (CM) cases. CM are
144 a family of rare genetic diseases, including multi-
145 ple distinct subtypes, that still lack proper diagnosis
146 with more than 50% of patients without a genetic
147 cause identified [25]. We exploited IMPatienT to cre-
148 ate a vocabulary of standard muscle-histology terms
149 that were then used to process patient histological
150 records and annotate biopsy images. Finally, multi-
151 ple exploratory visualizations were automatically
152 generated.

153 MATERIALS AND METHODS

154 IMPatienT is a web application developed with
155 the Flask micro-framework, which is a Python-based
156 web framework. Figure 1 illustrates the global orga-
157 nization of the web application, currently composed
158 of four modules: (i) Standard Vocabulary Creator,
159 (ii) Report Digitization, (iii) Image Annotation, and
160 (iv) Automatic Visualization Dashboard. All modules
161 incorporate free, open-source and well-maintained
162 libraries that are described in detail in the correspond-
163 ing sections.

164 *Module 1: standard vocabulary creator*

165 The standard vocabulary creator module allows
166 to create and modify a hierarchical list of vocabu-
167 lary terms with rich definitions that can be used
168 as image annotation classes, for processing of text
169 reports, or diagnosis decision support. Alternatively,
170 users can also import existing vocabularies in OBO
171 format instead. The standard vocabulary creator is an
172 essential module as it interacts with all subsequent
173 modules. This module also has export function to
174 download the current standard vocabulary created in
175 IMPatienT in JSON and OBO format (Open Biolog-
176 ical and Biomedical Ontologies file format).

177 Figure 2 shows a screenshot of the page used to
178 create/import, manage and export the standard vocabu-
179 lary tree. The tree is generated and rendered with
180 the JavaScript library JSTree (version 3.3.12). Each
181 node in the tree represents a vocabulary term, and
182 each term can have only one parent. The ergonomic
183 drag and drop system use the graphical user interface
184 (GUI) and allows the user to intuitively and quickly
185 edit and reorganize the vocabulary by adding new

186 terms or modifying existing ones. For each created
187 node (vocabulary term), the user can assign a name
188 and organize the tree structure (hierarchy).

189 Each term in the tree is associated with nine
190 optional properties, available *via* the vocabulary term
191 (node) detailed form. Four properties are defined by
192 the user: description, list of synonyms, translation in
193 another language, and use as annotation class. Two
194 properties are automatically generated: the unique
195 identifier (ID) and the hexadecimal color associated
196 with the term (for image annotation). Additional term
197 properties, including associated diagnosis/disease
198 class, associated genes, and the list of positively
199 correlated terms (*i.e.* co-occurring terms in reports),
200 are extracted from patient records registered in the
201 database.

202 Finally, if the user defines an alternative translation
203 for terms, there is an “invert vocabulary language”
204 button to conveniently switch between languages.
205 For instance, the user can create a vocabulary in any
206 language and define the translation in English, then
207 switch between the two display modes easily.

208 *Module 2: report digitization*

209 The standard vocabulary terms are used to pro-
210 cess documents that are in a free-text format. Module
211 2 uses a semi-automatic approach for digitization
212 and processing of free-text reports that combines fast
213 automatic detection of terms with manual reviewing
214 of the detection. The interface of Module 2 is a form
215 divided into four parts (Fig. 3).

216 In the first part of the digitization form (Fig. 3a),
217 a PDF file of the free-text report can be uploaded
218 for natural language processing (NLP) of the con-
219 tent. The text of the PDF report is automatically
220 extracted and processed with NLP. The NLP method
221 is only used to detect histological terms defined in
222 the standard vocabulary. Detected standard vocabu-
223 lary terms are highlighted (see corresponding section
224 below “Optical Character Recognition and Vocabu-
225 lary Term Detection”). Highlighted terms allow to
226 easily identify which standard vocabulary terms were
227 detected as present in positive or in negative form.
228 This is useful for quantitative performance assess-
229 ment.

230 The second part (Fig. 3b) of the digitization form
231 contains patient information, such as patient ID, doc-
232 ument ID, patient age. This section also allows the
233 user to input patient information that is not defined
234 by the standard vocabulary and thus, not processed
235 in the NLP section. For example, IMPatienT exploits

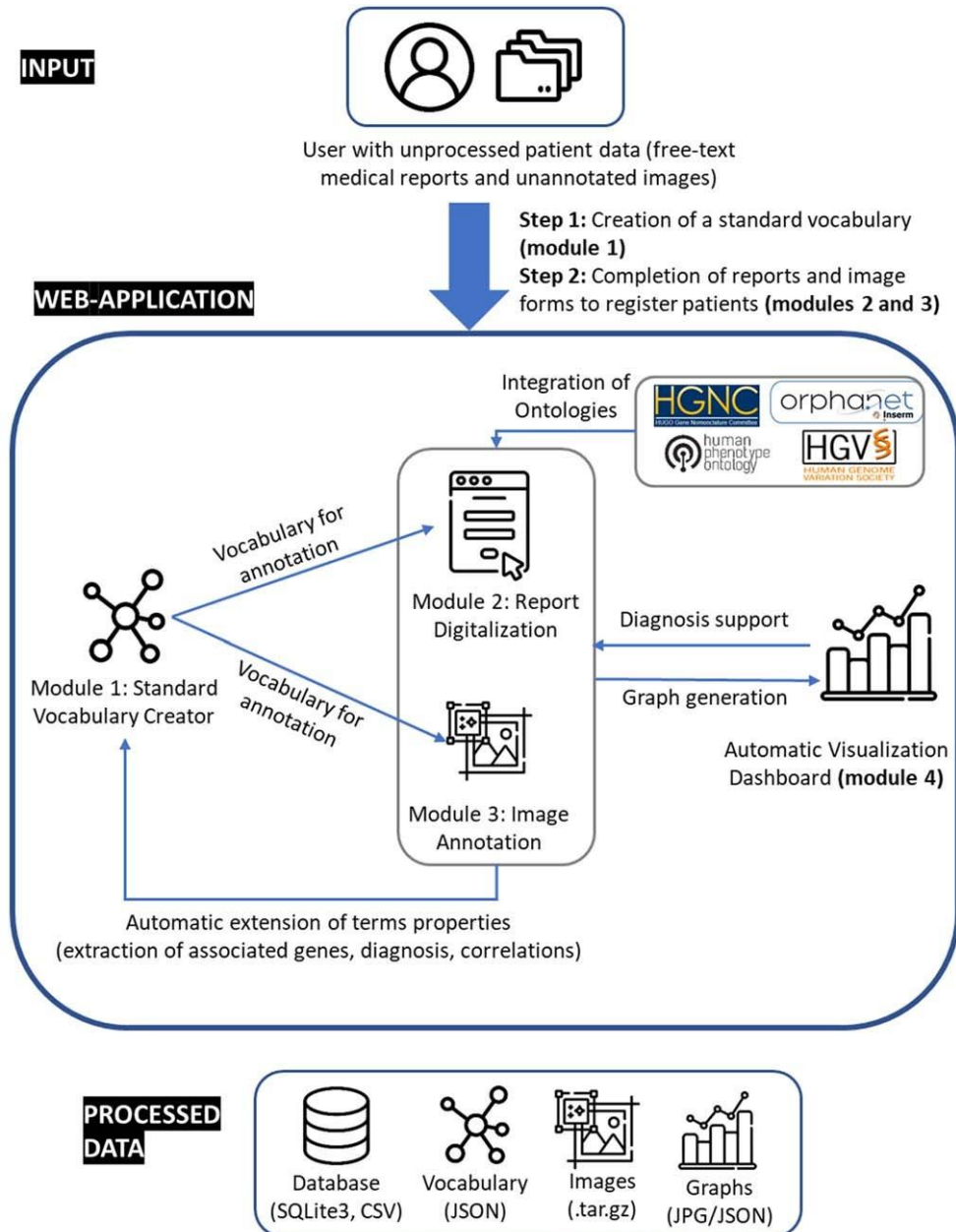


Fig. 1. Organization of IMPatientT web application.

236 well-established ontologies to normalize the genetic
 237 diagnosis and phenotypes (Fig. 4). For example,
 238 in the gene field, when the user inputs a char-
 239 acter string, gene symbols are retrieved from the
 240 HUGO Gene Nomenclature Committee (HGNC) and
 241 suggested [26]. Mutation notations are formatted
 242 according to the Human Genome Variation Society
 243 (HGVS) sequence variant nomenclature [27]. Phen-
 244 otypes are retrieved and suggested using the HPO.

245 These fields do not contain patient-identifying data
 246 and are optional.

247 The third part of the digitization form (Fig. 3c)
 248 contains the standard vocabulary tree viewer with an
 249 absence/presence slider. This section allows the user
 250 to correct the automatic detection of the NLP method
 251 or to add new observations. Each vocabulary term can
 252 be marked as present, absent or no information. For
 253 terms marked as present, the slider is used to indi-

(a) Standard Vocabulary Tree

[Download Vocabulary \(.JSON\)](#)
[Download Vocabulary \(.OBO\)](#)
[Upload Vocab \(Experimental\)](#)

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
 - ▶ Fibre aspect
 - ▶ NADH
 - ▶ Structural Reorganisation

[Save Tree](#)
[Invert Vocabulary Language](#)

(b) Vocabulary Properties

Vocabulary ID

MHO:000124

Vocabulary Name

Peripheral Cores

Alternative Language

Core Périphériques

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

Fig. 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 identifier (ID), name, alternative language translation, synonyms, description, associated genes and diseases and correlated terms extracted from the application instance database.

254 cate a notion of quantity or certainty of the term. For
 255 example, the statement “There are a small number of
 256 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 automat-

257
 258
 259

(a) Optional: Select a PDF and upload it to perform OCR/NLP analysis

Parcourir... sample_report.pdf English Upload

Results OCR / NLP Analysis

Presence of **rods** in the sample. There are also **some centralized nucleus**. There is no sign **of necrosis nor regeneration**. All fiber are atrophied and we can **see dark aggregate** on the periphery.

Vocabulary Automatically Detected as Present (dropdown) Vocabulary Automatically Annotated (Negated) (dropdown)

(b) Patient Information

Patient ID: DEMO_ID1 Biopsy ID: DEMO_BIOP1 Biopsy Date: 1970-01-01 Muscle: Quadriceps

Patient age at biopsy: 40 Diagnosed Gene (HGNC AFI): HGNC:12403 TTN

Phenotype terms (HPO API): HP:0001324 Muscle weakness, HP:0001252 Hypotonia, HP:0003690 Limb muscle weakness

Mutation: NM_001256850.1:c.51667C>T

(c) Vocabulary Tree

Rechercher

- Endosomal collagen
- Fibre Size
- Fibre Type 1
- Fibre Type 2
- Fuscinophilic material
- HETG: Fibre Aspect
- Inflammatory elements
- Interstitial alterations and damage
- Necklace Fibers
- Necrosis
- Neuromuscular spindle
- Nuclei
- Pale Zone
- Peripheral nail stroke
- Protein inclusions
- Regeneration
- Rods and Nemaline Bodies
 - Mini-rods
 - Red Dust Inclusion
 - Red spike (rod-like)
 - Rods**
 - Tubular aggregates
- Image Annotations
- Oxidative staining

Absence/Presence Annotation

Absence / Presence (slider)

Present (Total)

Vocabulary Term Details

Vocabulary ID: MHO:000048 Vocabulary Name: Rods

Synonyms: rods

Alternative Language: Batonnets

Associated Genes (Extracted from reports): HGNC:12403 TTN, HGNC:129 ACTA1, HGNC:7720 NEB, N/A

Associated Disease (Extracted from reports): ORPHA:607 Nemaline myopathy, UNCLEAR

Positively Correlates with (Extracted from reports: >0.5): Fuscinophilic material: Dark clusters/aggregates, Rods

Sudan Staining: Abormal (Lipids Overload), Type 1 Atrophied (small)

(d) Final Vocabulary Annotation (Present)

MHO:000036 Fuscinophilic material, Dark clusters/aggregates
 MHO:000027 Centralized Nuclei
 MHO:000033 Centralized Inclusions
 MHO:000048 Rods

Final Vocabulary Annotation (Negated)

MHO:000074 Necrosis
 MHO:000075 Regeneration

Commentaries and Conclusions

Commentary

Diagnosis Prediction: Predict 1

Method BOQA (Stats): Class: ORPHA:607 Nemaline myopathy Probability: 1

Final Diagnosis (Orphanet API): ORPHA:607 Nemaline myopathy

Save to Database

Fig. 3. Screenshot of the interface for 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) Overview section of all annotated terms, diagnosis selection and commentaries with automatic diagnosis support using the BOQA algorithm.

260 ically set to 0 (present in a negated sentence) or 1
 261 (present).

262 The fourth part (Fig. 3d) of the form allows the
 263 user to input comments and a final diagnosis for

264 the patient. Disease names are extracted and sug-
 265 gested from the Orphanet [28] knowledge base. It also
 266 includes an automatic disease suggestion based on
 267 already registered patients and the BOQA algorithm

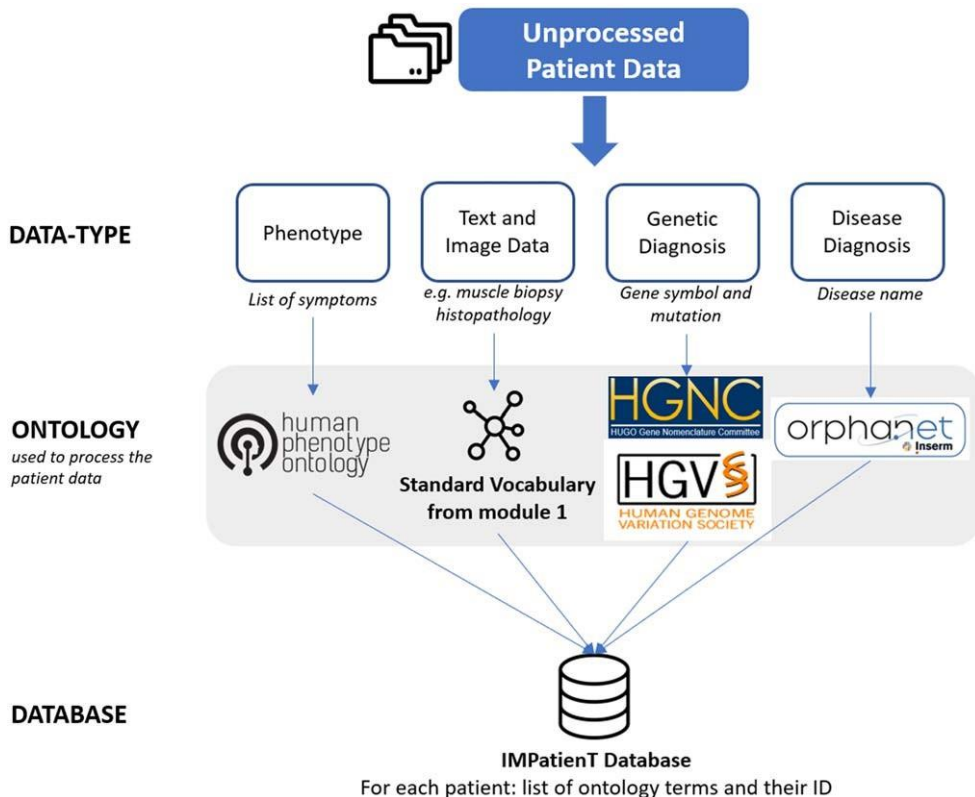


Fig. 4. Overview of the ontologies used by IMPatientT to process patient data in the report digitization module (module 2).

268 [17] (see the corresponding section below “Method for Patient Disease Suggestions”).
269

270 *Optical character recognition and vocabulary*
271 *term detection*

272 The patient report digitization in module 2 is facil-
273 itated by an automatic text recognition and keyword
274 detection method. The user uploads a PDF version
275 of the text reports to perform Optical Character
276 Recognition (OCR), followed by Natural Language
277 Processing (NLP) to automatically detect terms from
278 the standard vocabulary in the report. The NLP
279 method only matches the raw text to the standard
280 vocabulary defined in Standard Vocabulary Module
281 1. Figure 5 shows the workflow of the vocabu-
282 lary terms detection method. First, the PDF file is
283 converted to plain text using the Tesseract OCR
284 (implemented in python as pyTesseract). Then, the
285 text is processed with Spacy, an NLP python library,
286 by splitting the text into sentences and then into
287 individual words. The resulting list of sentences is
288 processed to detect negation using a simple imple-

289 mentation of the concept of NegEx [29]. An n -gram
290 (monograms, digrams, and trigrams) procedure is
291 applied to the list of words to identify contiguous
292 words in the context of all the sentences of the report.
293 The n -grams are then mapped against the user-created
294 standard vocabulary using fuzzy partial matching
295 (based on Levenshtein distance) with a score thresh-
296 old of 0.8. Matched keywords are kept and shown
297 on the interface by green or red highlighting of the
298 detected text using the Mark.JS JavaScript library
299 (green indicates the presence of the keyword, red indi-
300 cates the presence in a negated sentence). Keywords
301 are also automatically marked as present or absent
302 (negated) in the vocabulary tree.

303 *Disease suggestions*

304 The report digitization module 2 contains an im-
305 plication of the BOQA algorithm described by
306 Bauer et al. [17], for disease recommendation. Basi-
307 cally, the implemented BOQA algorithm computes
308 the similarity between a list of input vocabulary terms
309 annotated as “present” for a patient (the query) and
309

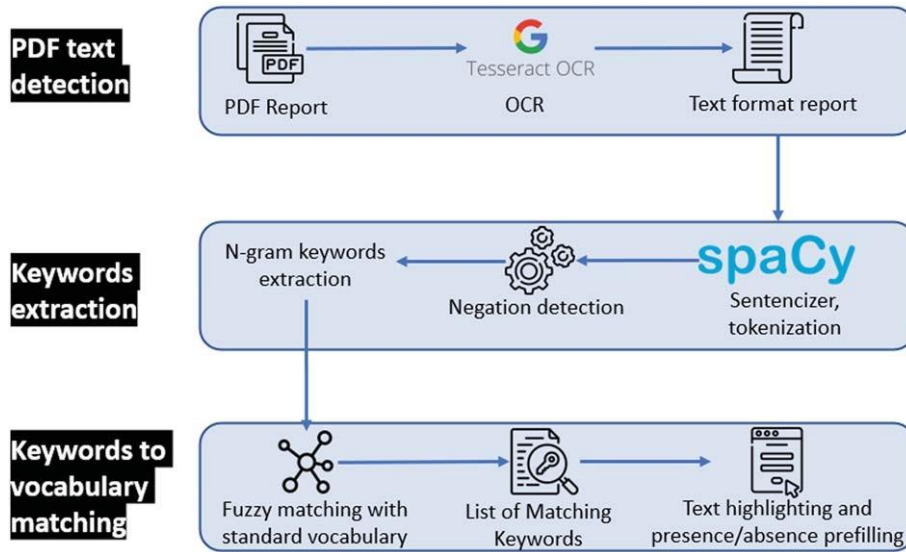


Fig. 5. Optical character recognition and vocabulary term detection method used in the report digitization module (module 2) to automatically analyze free-text reports.

a simulated patient profile for each disease class (model report) that is generated based on the data from already registered patients.

We implemented the BOQA algorithm in python and 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 computation 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 [30–32]. 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 Figs. 6a and 6b. The interface allows the user 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 behav-

339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369

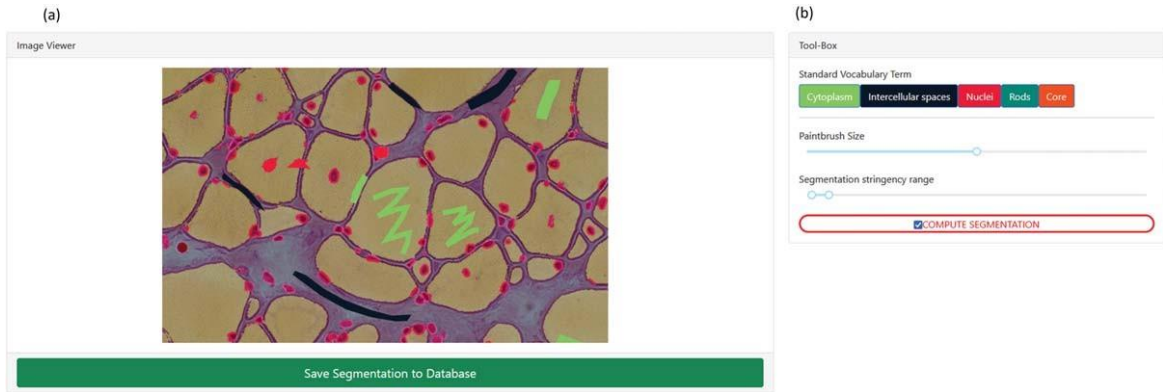


Fig. 6. **Screenshot of the image annotation module.** (a) Image viewer used to navigate, zoom and annotate the histology image. (b) Menu interface to select the annotation label, brush width and segmentation parameters.

ior 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 format), 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 the creation of interactive graphs.

Interaction between the modules

IMPatientT is divided into four modules that are interconnected. The standard vocabulary module provides the vocabulary used for the image annotation module and for the NLP method used for the (histologic) standard vocabulary term detection in the report digitization module. Any modification in the vocabulary is automatically propagated to these modules, updating the form templates and triggering the recalculation of all visualizations with the latest vocabulary information. Any modification to the standard vocabulary also updates all patients in the database to the latest version of the vocabulary, meaning that term names and definitions will be updated, and deleted terms will be marked as outdated. Adding patient information in the database, whether they are text reports (module 2) or image data (module 3), will automatically update the visualization dashboard

with the latest patient information in the database. The term frequency statistics calculated by the visualization dashboard and used by the disease suggestion algorithm are automatically updated as well, increasing live performances. The visualization dashboard is also directly linked to the standard vocabulary and during the generation of the visualizations, the rich definition of the standard terms is updated with newly associated genes, diagnosis and positively correlated terms.

Application security and personal data

IMPatientT is developed as a free and open-source project meaning that the code can be audited by anyone in the GitHub code repository (<https://github.com/lambda-science/IMPatientT>).

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. The synthetic dataset generated and analyzed during the current study is also available in the same repository. No name or date of birth is 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

IMPatientT is an interactive and user-friendly web application that integrates a semi-automatic approach for text and image data digitization, processing, and

432 exploration. Due to its modular architecture and its
 433 standard vocabulary creator, it has a wide range of
 434 potential uses.

435 *IMPatientT main functionalities*

436 Table 1 shows the main functionalities of IMPa-
 437 tientT compared to other similar tools used in the
 438 community. IMPatientT aims to be a one-stop plat-
 439 form that integrates tools for an end-to-end analysis
 440 workflow of multi-modal patient’s data. Out of 18
 441 selected features, IMPatientT integrates 14 of them
 442 versus a mean of 4.4 for other software with the best
 443 ones being SAMS and PhenoStore integrating 6 fea-
 444 tures each. Nevertheless, software such as SAMS,
 445 PhenoStore, Phenotips and Cytomine each integrate
 446 features that are not yet present in IMPatientT.

447 IMPatientT implements novel functionalities to
 448 process and exploit patient data. For example, IMPa-
 449 tientT is compatible with any research domain thanks
 450 to its standard vocabulary builder. Also, with the
 451 OCR/NLP method, IMPatientT can process histologic
 452 text reports, allowing the user to exploit scanned doc-
 453 uments. Finally, IMPatientT provides useful tools to
 454 exploit patient data with the various visualizations,
 455 the term, frequency table, correlation matrix and the
 456 automatic enrichment of the vocabulary term defini-
 457 tions (associated genes and diseases).

458 *IMPatientT usage*

459 Figure 1 shows how the user can interact with
 460 the web application to digitize, process, and explore
 461 patient data. In IMPatientT, modules can be used
 462 independently, allowing users to only use the tools
 463 they need. For example, a user might only have text
 464 report data, in this case they would be able to use
 465 the standard vocabulary creator, the report digitiza-
 466 tion tools and the visualization dashboard to process
 467 and explore their data. In another scenario, a user
 468 could only be interested in annotating an image
 469 dataset using a shared standard vocabulary that can
 470 be modified and updated collaboratively. In this use
 471 case, they would be able to only use the standard
 472 vocabulary creator and the image annotation mod-
 473 ule. However, the main strength of IMPatientT lies in
 474 the multimodal approach it provides and the strong
 475 interactions between modules.

476 For the complete multimodal approach, the first
 477 step is to create a standard vocabulary using the Stan-
 478 dard Vocabulary Creator interface (module 1). The
 479 user only needs to create a few terms (nodes) to begin

480 using the web application. Defining the properties
 481 of the terms (definition, synonyms, etc.) is optional,
 482 and organizing them in a hierarchical structure is also
 483 optional.

484 In the second step, the user can start digitizing
 485 patient reports using module 2. This can be done
 486 either manually by filling out the form in module 2
 487 and checking terms as present or absent in a given
 488 report, or automatically using the Vocabulary Term
 489 Matching method to process a PDF version of the
 490 report. Using module 3, the user can also upload,
 491 annotate, and segment image data.

492 Finally, the user can explore multiple visual-
 493 izations (histograms, correlation matrix, confusion
 494 matrix, frequency tables) that are automatically
 495 generated in module 4. All data entered *via* the
 496 web application are retrievable in standard formats,
 497 including the whole database of reports as a single
 498 SQLite3 file or CSV files, the images and their seg-
 499 mentation models and masks as a GZIP archive, the
 500 standard vocabulary with annotation as a JSON file
 501 and various graphs and tables as JSON or PNG files.

502 *Use case: congenital myopathy histology reports*

503 As a use case of IMPatientT, we focused on con-
 504 genital myopathies (CM). We used the standard
 505 vocabulary creator to create a sample muscle his-
 506 tology standard vocabulary based on common terms
 507 used in muscle biopsy reports from the Paris Institute
 508 of Myology. Then, we inserted 40 digitally gener-
 509 ated patients in the database with random sampling
 510 of standard vocabulary terms and associated a gene
 511 and disease class from a list of common CM genes
 512 and three recurring CM subtypes (nemaline myopa-
 513 thy, core myopathy and centronuclear myopathy). All
 514 these data are available on the demo instance of IMPa-
 515 tientT (<https://impatient.lbgi.fr/>).

516 For text data, Supplementary Figure S1 shows
 517 the results of the automatic NLP method applied
 518 to an artificial muscle histology report. Twenty-two
 519 keywords were detected that match to the stan-
 520 dard vocabulary and seven of them were detected in
 521 negated sentences (red highlight). Out of the twenty-
 522 two keywords, eighteen were correctly detected and
 523 one was detected in the wrong state of negation:
 524 “abnormal fiber differentiation” is highlighted as
 525 negated although it is present in a non-negated sen-
 526 tence part. Three keywords (fiber type, internalized
 527 nuclei, centralized nuclei) were detected as matching
 528 for multiple keywords from the vocabulary due to
 529 high similarity. For example, the keywords “internal-

Table 1
Functionalities of IMPatient compared to common state-of-the-art tools

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

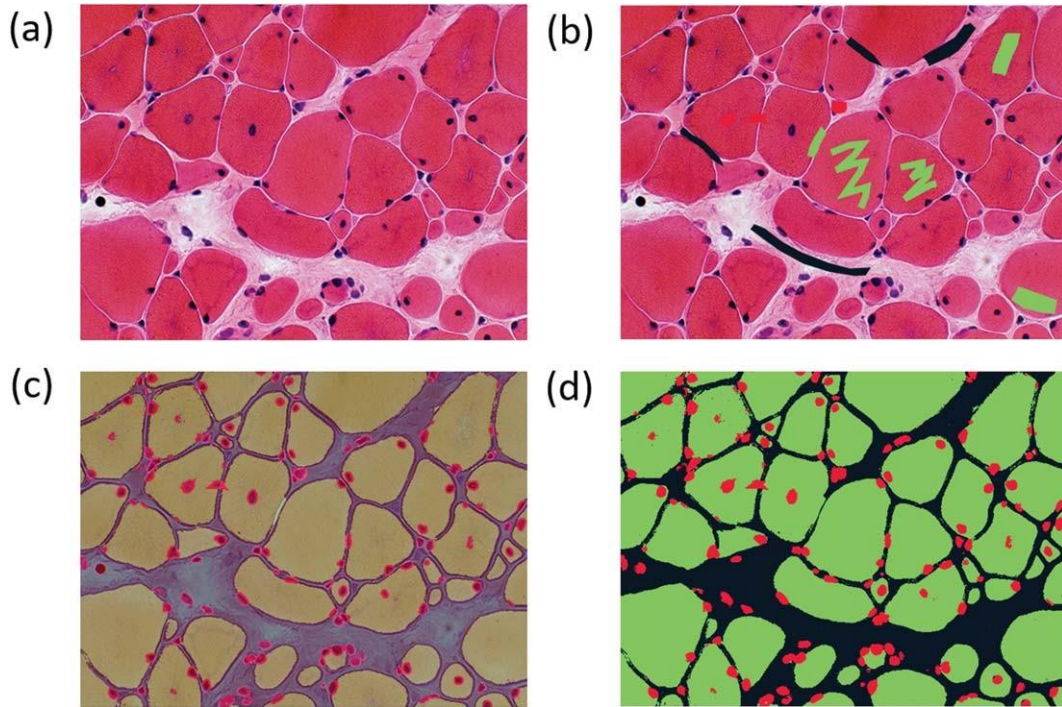


Fig. 7. **Image segmentation process in the image segmentation module.** (a) Raw image input before annotation. (b) Image with limited manual annotation of cytoplasm (green), cell nucleus (red) and intercellular space (black). (c) Blended image of the raw image and segmented image after automated segmentation with a random-forest classifier. (d) Segmented image mask alone.

530 ized nuclei” and “centralized nuclei” have a similarity
 531 score of 86 using the Levenstein distance. Two
 532 keywords defined in the standard vocabulary were
 533 missed and not highlighted: “biopsy looks abnormal”
 534 (“abnormal biopsy” in the vocabulary) and “purplish
 535 shade” (“purplish aspect” in the vocabulary).

536 For the image data, Fig. 7 shows an example of
 537 the segmentation of a biopsy image, where we annotated
 538 the cytoplasm of the cells (green), intercellular
 539 spaces (black) and cell nuclei (red). The raw image
 540 (Fig. 7a) is annotated with free-shape areas associated
 541 with standard vocabulary terms (Fig. 7b). Then, the
 542 whole image is automatically segmented based on the
 543 annotations, producing the segmentation mask where
 544 each pixel is associated with a class (Figs. 7c, 7d).

545 The automatic visualization dashboard was used
 546 to generate the six visualizations provided in Fig. 8.
 547 These visualizations include a breakdown of the
 548 patients in the database by age, genes, or diagnosis
 549 (Fig. 8a). A correlation matrix (using Pearson
 550 correlation coefficient) between the occurrences of
 551 standard vocabulary terms is generated (Fig. 8b),
 552 which can serve as a starting point for exploration of
 553 co-occurrence of features in patients. The confusion
 554 matrix of the final diagnosis of patients versus the

555 suggested disease name with BOQA (Fig. 8c) allows
 556 the user to monitor the accuracy of the disease sug-
 557 gession function. In addition, a histogram showing
 558 the classification of patients without a final diagnosis
 559 is provided to indicate possible prognosis of undi-
 560 agnosed patients (Fig. 8d). Finally, the frequency of
 561 each standard vocabulary term by gene and by disease
 562 is automatically calculated and shown in two tables
 563 (Supplementary Tables S1 and S2).

564 DISCUSSION

565 IMPatientT is a one-stop platform that simplifies
 566 the digitization, processing, and exploration of both
 567 textual and image patient data. The web application
 568 is centered around the concept of a standard vocabu-
 569 lary tree (that is easy to create or to import existing
 570 ones) and used to process text and image data. This
 571 allows IMPatientT to work with patient data from
 572 domains that still lack a consensus ontology and rely
 573 on well-established ontologies for patient data, such
 574 as HPO for phenotypes, Orphanet for disease names
 575 or HGCN/HGVS for genetic diagnoses.

576 The semi-automatic approach implemented in
 577 IMPatientT offers faster digitization processes while

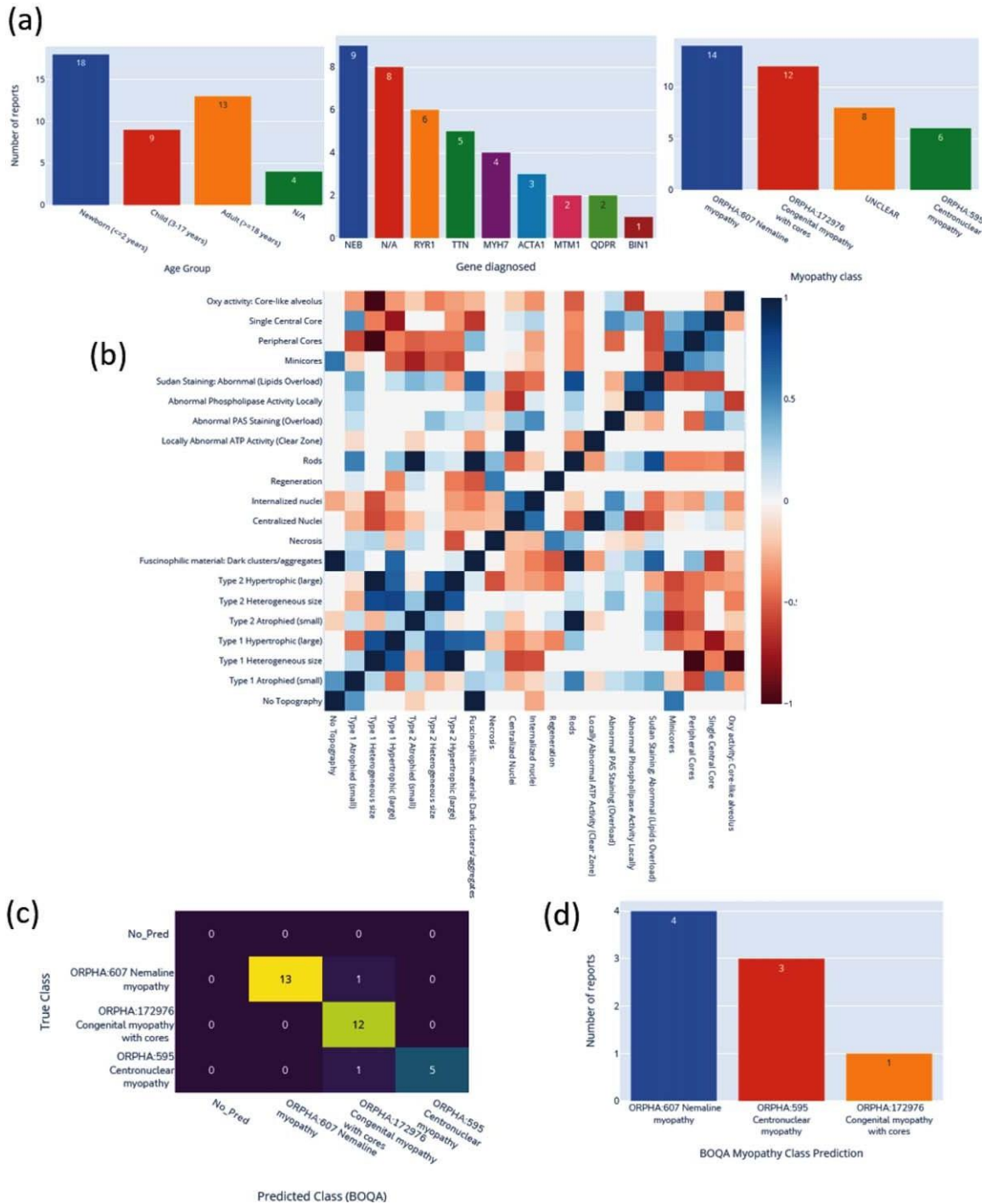


Fig. 8. Automatic visualization of 40 generated congenital myopathy reports. (a) Histogram of the number of reports by age group, by diagnosed gene (top 9) or by congenital myopathy class. (b) Correlation matrix of standard vocabulary terms after annotation for all reports. (c) Confusion matrix of BOQA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

578 ensuring accuracy through manual review. This is
579 achieved by analyzing text data using OCR and NLP
580 to automatically match the text to the standard vocab-
581 ulary, followed by manual correction. For image
582 data, the user first provides sparse annotations on the
583 image, which are then used to compute an automatic
584 segmentation of the whole image. For data explo-
585 ration, IMPatientT uses a fully automatic approach
586 including various visualizations as well as diagno-
587 sis suggestions, while allowing the user to extract
588 the processed data in a standard format for further
589 analysis (database, images, frequency tables).

590 IMPatientT aims to integrate multiple approaches
591 in a unified platform with two main objectives: uni-
592 versality (*i.e.* not restricted to a specific domain) and
593 multimodality (*i.e.* integration of multiple data types).
594 To our knowledge, other tools similar to IMPatientT
595 do not fulfill both objectives.

596 We performed a comparison of the main func-
597 tionalities of IMPatientT with other tools used in the
598 community. Workflows such as Phenotips, SAMS
599 and PhenoStore are similar to IMPatientT as they are
600 designed as a patient information database. However,
601 they are restricted to processing patient phenotype
602 data using HPO and do not integrate multimodal data.
603 IMPatientT goes further by allowing custom observa-
604 tions with the vocabulary builder for fields of study
605 currently without ontology (e.g. histology) and with
606 automatic digitization with OCR/NLP as well as inte-
607 grating tools to exploit image data. In contrast to a
608 tool like Prote^{ge} that is the reference tool to build
609 an ontology, the module 1 of IMPatientT provides a
610 user-friendly interface where non-ontology experts
611 can start building a vocabulary in a tree representa-
612 tion way. To improve the interoperability, IMPatientT
613 allows the export in OBO format, which is compatible
614 with Prote^{ge} for further improvement of the created
615 vocabulary based on ontological best practices.

616 Other tools are available that implement the func-
617 tionality of one or two IMPatientT modules. For
618 example, Doc2HPO is a tool that also uses a semi-
619 automatic approach to digitize clinical text according
620 to a list of HPO terms, based on NLP methods
621 and negation detection. However, as Doc2HPO is
622 restricted to HPO, it does not provide custom vocab-
623 ulary tree facilities. In contrast IMPatientT is suitable
624 for digitization of text data from any domain of inter-
625 est that can be covered with its vocabulary creation
626 module when no ontology is currently available.

627 For image data, software such as Cytomine and
628 Ilastik are widely used and perform well on biolog-
629 ical data. In contrast to these standalone tools that

630 can process big raw image data locally, the mod-
631 ule 3 of IMPatientT is currently a web application
632 that can process lightweight images (in Mega-octet).
633 However, lightweight images (e.g. JPEG) are the
634 most common form images available in patients data
635 for histopathologies in myopathies or eye fundus
636 in retinopathies for instance. Moreover, IMPatientT
637 allow the user to take into consideration the mul-
638 timodal aspects of patient data by keeping the raw
639 image and the expert interpretation (histological
640 report) in a single database along with a collaborative
641 and rich user-defined vocabulary.

642 Finally, in IMPatientT we reimplemented the diag-
643 nosis support algorithm called BOQA that is also
644 used in Phenomizer, a tool to rank a list of the top
645 matching diseases based on a list of input HPO terms.
646 We modified the algorithm to consider frequencies of
647 vocabulary terms by disease to have meaningful pre-
648 dictions. In contrast, BOQA uses binary states for
649 terms (terms are marked as present or absent) and is
650 not compatible with numeric features. In the future,
651 it will be interesting to implement a more complex
652 system such as explainable AI with learning classifier
653 systems [33]. This should improve accuracy, explain-
654 ability, and handling of quantitative values, although
655 at the cost of computational power.

656 IMPatientT still lacks some features compared to
657 other tools, such as a pedigree editor, support for
658 DICOM and gigapixel images and phenotypic data
659 export to the Phenopacket format. In the future, we
660 plan to further develop IMPatientT by adding these
661 features to the interface. We also plan to add a vocab-
662 ulary search engine and sharing feature to encourage
663 reuse or to explore the automatization of the standard
664 vocabulary creation with the analysis of a complete
665 corpus of text. For text analysis, we intend to imple-
666 ment additional context comprehension, *i.e.* not only
667 negation but also hypothetical statements, uncertainty
668 and family context as well as better text-vocabulary
669 terms matching. Finally, we plan to expand the scope
670 of the OCR/NLP method by integrating existing NLP
671 tools that will automatically detect and align to HPO
672 terms, gene symbols and disease names in the report
673 text.

674 ACKNOWLEDGMENTS

675 We thank the BiGEst-ICube platform for their
676 assistance. We thank the Agence Nationale de la
677 Recherche (ANR), 80 | Prime CNRS (MYO-xIA
678 Project), the University of Strasbourg and INSERM
679 for funding this work.

CONFLICTS OF INTEREST

The authors have no conflict of interest to report.

SUPPLEMENTARY MATERIALS

The supplementary material is available in the electronic version of this article: <https://dx.doi.org/10.3233/JND-230085>.

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. 2014 Int. Workshop Pattern Recognit. Neuroimaging, 2014, p. 1-4. <https://doi.org/10.1109/PRNI.2014.6858526>
- [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. *Intell. Comput. Theor. Appl.*, Cham: Springer International Publishing; 2019, p. 460-9. https://doi.org/10.1007/978-3-030-26763-6_44
- [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;194:56-67. <https://doi.org/10.1016/j.trsl.2018.01.001>
- [4] Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci Rep* 2021;11:3254. <https://doi.org/10.1038/s41598-020-74399-w>
- [5] North KN, Wang CH, Clarke N, Jungbluth H, Vainzof M, Dowling JJ, et al. Approach to the diagnosis of congenital myopathies. *Neuromuscul Disord* 2014;24:97-116. <https://doi.org/10.1016/j.nmd.2013.11.003>
- [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;43:101. <https://doi.org/10.1186/s13052-017-0419-z>
- [7] Bo'hm 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:e67527. <https://doi.org/10.1371/journal.pone.0067527>
- [8] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32:D267-70. <https://doi.org/10.1093/nar/gkh061>
- [9] Ko'hler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res* 2021;49:D1207-17. <https://doi.org/10.1093/nar/gkaa1043>
- [10] Musen MA. The Prote'ge' Project: A Look Back and a Look Forward. *AI Matters* 2015;1:4-12. <https://doi.org/10.1145/2757001.2757003>
- [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;47:W566-70. <https://doi.org/10.1093/nar/gkz386>
- [12] Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Che'nier S, et al. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. *Hum Mutat* 2013;34:1057-65. <https://doi.org/10.1002/humu.22347>
- [13] Steinhilber R, Proft S, Seelow E, Schalau T, Robinson PN, Seelow D. Deep phenotyping: symptom annotation made simple with SAMS. *Nucleic Acids Res* 2022;50(W1):W677-81. <https://doi.org/10.1093/nar/gkac329>
- [14] Laurie S, Piscia D, Matalonga L, Corvo' A, Ferna'ndez-Callejo M, Garcia-Linares C, et al. The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. *Hum Mutat* 2022;43:717-33. <https://doi.org/10.1002/humu.24353>
- [15] Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods* 2015;12:841-3. <https://doi.org/10.1038/nmeth.3484>
- [16] Ko'hler S, Schulz MH, Krawitz P, Bauer S, Do'iken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 2009;85:457-64. <https://doi.org/10.1016/j.ajhg.2009.09.003>
- [17] Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics* 2012;28:2502-8. <https://doi.org/10.1093/bioinformatics/bts471>
- [18] Mare'e R, Rollus L, Ste'vens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics* 2016;32:1395-401. <https://doi.org/10.1093/bioinformatics/btw013>
- [19] Aubreville M, Bertram C, Klopffleisch R, Maier A. SlideRunner - A Tool for Massive Cell Annotations in Whole Slide Images. *ArXiv180202347 Cs* 2018:309-14. https://doi.org/10.1007/978-3-662-56537-7_81
- [20] Berg S, Kutra D, Kroeger T, Strachle CN, Kausler BX, Haubold C, et al. ilastik: interactive machine learning for (bio)image analysis. *Nat Methods* 2019;16:1226-32. <https://doi.org/10.1038/s41592-019-0582-9>
- [21] Cinaglia P, Tradigo G, Cascini GL, Zumpano E, Veltri P. A framework for the decomposition and features extraction from lung DICOM images. *Proc. 22nd Int. Database Eng. Appl. Symp., New York, NY, USA: Association for Computing Machinery*; 2018, p. 31-6. <https://doi.org/10.1145/3216122.3216127>
- [22] 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;21:1585-93. <https://doi.org/10.1038/s41436-018-0381-1>
- [23] Smedley D, Jacobsen JOB, Ja'ger M, Ko'hler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc* 2015;10:2004-15. <https://doi.org/10.1038/nprot.2015.124>
- [24] Cinaglia P, Guzzi PH, Veltri P. INTEGRO: an algorithm for data-integration and disease-gene association. 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM, 2018, p. 2076-81. <https://doi.org/10.1109/BIBM.2018.8621193>
- [25] Jungbluth H, Treves S, Zorzato F, Sarkozy A, Ochala J, Sewry C, et al. Congenital myopathies: disorders of excitation-contraction coupling and muscle contraction. *Nat Rev Neurol* 2018;14:151-67. <https://doi.org/10.1038/nrneuro.2017.191>
- [26] 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;49:D939-46. <https://doi.org/10.1093/nar/gkaa980>
- [27] 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;37:564-9. <https://doi.org/10.1002/humu.22981>
- [28] INSERM. Orphanet: an online database of rare diseases and orphan drugs 1997. <http://www.orpha.net> (accessed February 13, 2022).
- [29] 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;34:301-10. <https://doi.org/10.1006/jbin.2001.1029>
- [30] Gouillart E. Interactive Machine Learning – Image segmentation. GitHub 2020. [https://github.com/plotly/dash-](https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation) 818
819
820
821
822
823
824
825
826
827
828
829
830
- [sample-apps/tree/main/apps/dash-image-segmentation](https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation) (accessed November 23, 2021).
- [31] Walt S van der, Scho'nberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. *PeerJ* 2014;2:e453. <https://doi.org/10.7717/peerj.453>
- [32] Hossain S. Visualization of Bioinformatics Data with Dash Bio. *Proc 18th Python Sci Conf* 2019:126-33. <https://doi.org/10.25080/Majora-7ddc1dd1-012>
- [33] Urbanowicz RJ, Moore JH. ExSTraCS 2.0: Description and Evaluation of a Scalable Learning Classifier System. *Evol Intell* 2015;8:89-116. <https://doi.org/10.1007/s12065-015-0128-8>