



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.. 2022. hal-03635350v1

HAL Id: hal-03635350

<https://hal.science/hal-03635350v1>

Preprint submitted on 8 Apr 2022 (v1), 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.

1 TITLE

2 IMPatientT: an integrated web application to digitize, process and explore
3 multimodal patient data.

4 Authors

5 Corentin Meyer¹, Norma Beatriz Romero², Teresinha Evangelista², Brunot Cadot³, Jocelyn
6 Laporte⁴, Anne Jeannin-Girardon¹, Pierre Collet¹, Kirsley Chennen¹, Olivier Poch^{1*}

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

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

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

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

15 * Correspondence: olivier.poch@unistra.fr

16 ABSTRACT

17 Background

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

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

27 **Results**

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

36 **Conclusions**

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

41 **KEYWORDS**

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

44 BACKGROUND

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

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 / histological analysis (imaging information) with medical records and sequencing data is
57 essential for differential diagnosis between CM subtypes (5–7). Centralization of multimodal data
58 using dedicated 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). Several
66 tools have been developed to easily manage and extend these standard vocabularies, such as
67 OBOEdit(10). Text mining processes have been developed based on these standard vocabularies,

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

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

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

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

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

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

100 **IMPLEMENTATION**

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

107 **Module 1: Standard Vocabulary Creator**

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

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

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

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

129 **Module 2: Report Digitization**

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

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

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

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

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

157 **Optical Character Recognition and Vocabulary Term Detection**

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

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

174 **Disease Suggestions**

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

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

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

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

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

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

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

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

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

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

215 **Module 4: Automatic Visualization Dashboard**

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

219 **Application Security and Personal Data**

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

227 **RESULTS**

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

231 **Web Application Workflow**

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

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

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

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

248 **Use Case: Congenital Myopathy Histology Reports**

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

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

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

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

271 **DISCUSSION**

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

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

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

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

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

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

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

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

310 CONCLUSIONS

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

319 Availability and requirements

320 **Project name:** IMPatientT v1.4.3

321 **Project home page:** <https://github.com/lambda-science/IMPatientT>

322 **Project Version:** v1.4.3 <https://github.com/lambda-science/IMPatientT/releases/tag/v1.4.3>

323 **Operating system(s):** Platform independent

324 **Programming language:** Python 3.9, HTML5, CSS3, JS

325 **Other requirements:** Flask 2.0, Spacy 3.2, Dash 2, Plotly 5, Scikit-Learn 1.0, Scikit-Image 0.19,
326 Tesseract 4, JSTree 3.3, MarkJS 8.11, JQuery 3.6, Bootstrap 5

327 **License:** GNU AGPL

328 **Any restrictions to use by non-academics:** /

329 **Figures titles and legends**

330 **Figure 1:** IMPatientT web application workflow and organisation

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

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

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

337 **Figure 3:** Screenshot of the report digitalization module.

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

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

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

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

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

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

350 **Figure 6:** Screenshot of the image annotation module.

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

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

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

354 (A) Raw image input before annotation.

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

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

359 (D) Segmented image mask alone.

360 **Additional files**

361 **Additional File 1:**

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

368 **Additional File 2:**

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

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

375 **List of abbreviations**

376 AI: Artificial Intelligence

377 BOQA: Bayesian Ontology Query Algorithm

378 CM: Congenital Myopathies

379 EDA: Exploratory Data Analysis

380 GUI: Graphical User Interface

381 HGNC: HUGO Gene Nomenclature Committee

382 HGVS: Human Genome Variation Society

383 HPO: Human Phenotype Ontology

384 NLP: Natural Language Processing

385 OCR: Optical Character Recognition

386 **Declarations**

387 **Ethics approval and consent to participate**

388 Not applicable

389 **Consent for publication**

390 Not applicable

391 **Availability of data and materials**

392 The datasets generated and analyzed during the current study are available in the IMPatient
393 repository, <https://github.com/lambda-science/IMPatient>.

394 **Competing interests**

395 The authors declare that they have no competing interests.

396 **Funding**

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

399 **Authors' Contributions**

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

404 **Acknowledgements**

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

406 **References**

- 407 1. Kerr WT, Hwang ES, Raman KR, Barritt SE, Patel AB, Le JM, et al. Multimodal diagnosis of
408 epilepsy using conditional dependence and multiple imputation. In: 2014 International
409 Workshop on Pattern Recognition in Neuroimaging. 2014. p. 1–4.
- 410 2. Yan R, Ren F, Rao X, Shi B, Xiang T, Zhang L, et al. Integration of Multimodal Data for Breast
411 Cancer Classification Using a Hybrid Deep Learning Method. In: Huang D-S, Bevilacqua V,
412 Premaratne P, editors. Intelligent Computing Theories and Application. Cham: Springer
413 International Publishing; 2019. p. 460–9. (Lecture Notes in Computer Science).
- 414 3. Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial
415 intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Transl Res J*
416 *Lab Clin Med.* 2018 Apr;194:56–67.

- 417 4. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for
418 early detection of Alzheimer's disease stage. *Sci Rep*. 2021 Feb 5;11(1):3254.
- 419 5. North KN, Wang CH, Clarke N, Jungbluth H, Vainzof M, Dowling JJ, et al. Approach to the
420 diagnosis of congenital myopathies. *Neuromuscul Disord NMD*. 2014 Feb;24(2):97–116.
- 421 6. Cassandrini D, Trovato R, Rubegni A, Lenzi S, Fiorillo C, Baldacci J, et al. Congenital
422 myopathies: clinical phenotypes and new diagnostic tools. *Ital J Pediatr*. 2017 Nov
423 15;43(1):101.
- 424 7. Böhm J, Vasli N, Malfatti E, Le Gras S, Feger C, Jost B, et al. An integrated diagnosis strategy for
425 congenital myopathies. *PLoS One*. 2013;8(6):e67527.
- 426 8. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical
427 terminology. *Nucleic Acids Res*. 2004 Jan 1;32(Database issue):D267–70.
- 428 9. Köhler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The
429 Human Phenotype Ontology in 2021. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D1207–17.
- 430 10. Day-Richter J, Harris MA, Haendel M, The Gene Ontology OBO-Edit Working Group, Lewis S.
431 OBO-Edit—an ontology editor for biologists. *Bioinformatics*. 2007 Aug 15;23(16):2198–200.
- 432 11. Liu C, Peres Kury FS, Li Z, Ta C, Wang K, Weng C. Doc2Hpo: a web application for efficient and
433 accurate HPO concept curation. *Nucleic Acids Res*. 2019 Jul 2;47(W1):W566–70.
- 434 12. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, et al. PhenoTips: Patient
435 Phenotyping Software for Clinical and Research Use. *Hum Mutat*. 2013;34(8):1057–65.
- 436 13. Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of
437 multi-gigapixel imaging data using Cytomine. *Bioinformatics*. 2016 May 1;32(9):1395–401.
- 438 14. Aubreville M, Bertram C, Klopffleisch R, Maier A. SlideRunner - A Tool for Massive Cell
439 Annotations in Whole Slide Images. *ArXiv180202347 Cs*. 2018;309–14.
- 440 15. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate
441 genes for human diseases. *Nat Methods*. 2015 Sep;12(9):841–3.
- 442 16. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human
443 genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009
444 Oct;85(4):457–64.
- 445 17. Bauer S, Kohler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and
446 noise-tolerant semantic searches. *Bioinformatics*. 2012 Oct 1;28(19):2502–8.
- 447 18. Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen
448 extracts and prioritizes patient phenotypes directly from medical records to expedite genetic
449 disease diagnosis. *Genet Med*. 2019 Jul;21(7):1585–93.
- 450 19. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation
451 diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015
452 Dec;10(12):2004–15.

- 453 20. H J, S T, F Z, A S, J O, C S, et al. Congenital myopathies: disorders of excitation-contraction
454 coupling and muscle contraction. *Nat Rev Neurol* [Internet]. 2018 Mar [cited 2022 Mar
455 16];14(3). Available from: <https://pubmed.ncbi.nlm.nih.gov/29391587/>
- 456 21. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC
457 and VGNC resources in 2021. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D939–46.
- 458 22. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al.
459 HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*.
460 2016;37(6):564–9.
- 461 23. INSERM. Orphanet: an online database of rare diseases and orphan drugs [Internet]. 1997
462 [cited 2022 Feb 13]. Available from: <http://www.orpha.net>
- 463 24. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for
464 Identifying Negated Findings and Diseases in Discharge Summaries. *J Biomed Inform*. 2001
465 Oct 1;34(5):301–10.
- 466 25. Gouillart E. Interactive Machine Learning - Image segmentation [Internet]. GitHub. 2020 [cited
467 2021 Nov 23]. Available from: [https://github.com/plotly/dash-sample-](https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation)
468 [apps/tree/main/apps/dash-image-segmentation](https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-image-segmentation)
- 469 26. Walt S van der, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-
470 image: image processing in Python. *PeerJ*. 2014 Jun 19;2:e453.
- 471 27. Hossain S. Visualization of Bioinformatics Data with Dash Bio. *Proc 18th Python Sci Conf*.
472 2019;126–33.
- 473 28. Urbanowicz RJ, Moore JH. ExSTraCS 2.0: Description and Evaluation of a Scalable Learning
474 Classifier System. *Evol Intell*. 2015 Sep;8(2):89–116.
- 475