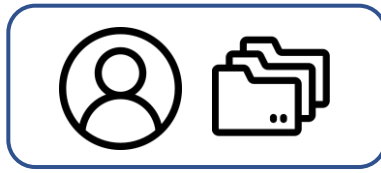


INPUT



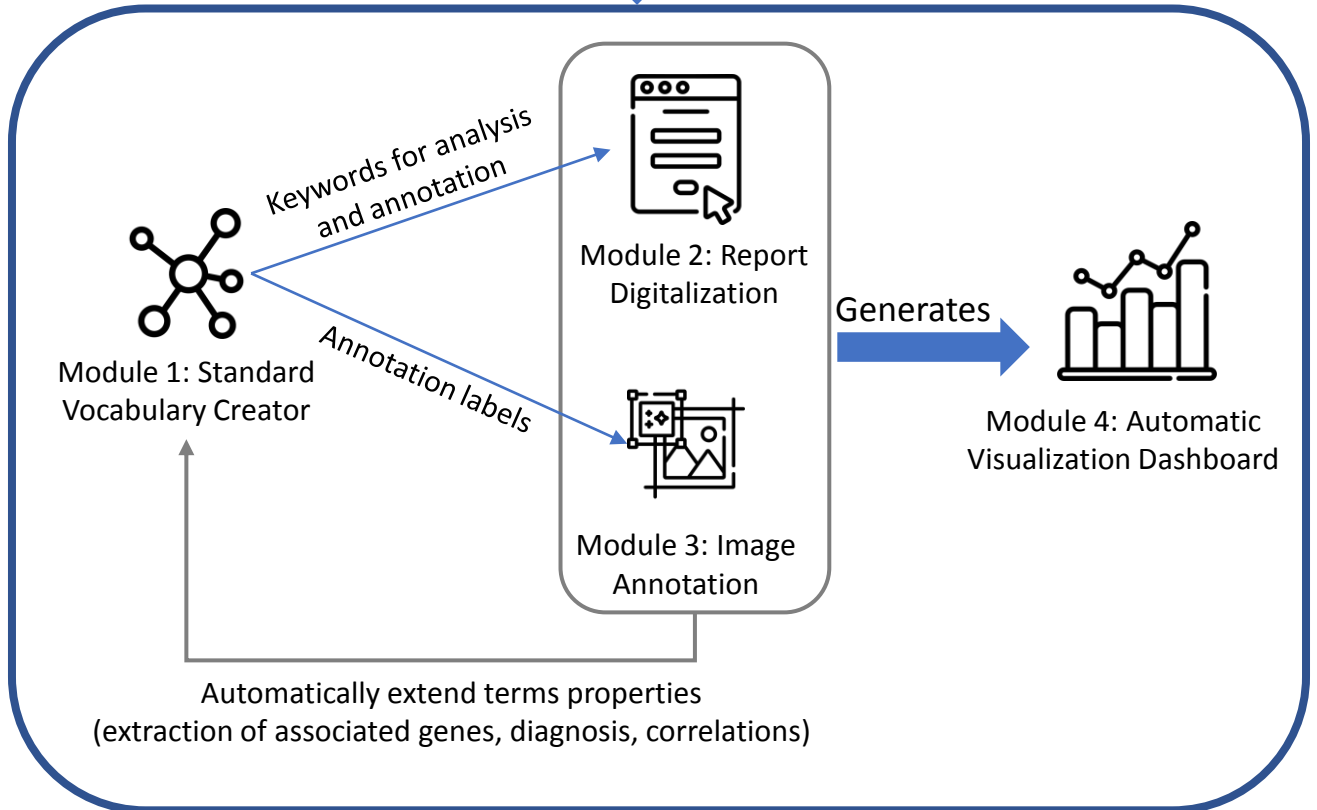
User with unprocessed patient data (free-text medical reports and unannotated images)



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

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

WEB-APPLICATION



OUTPUT

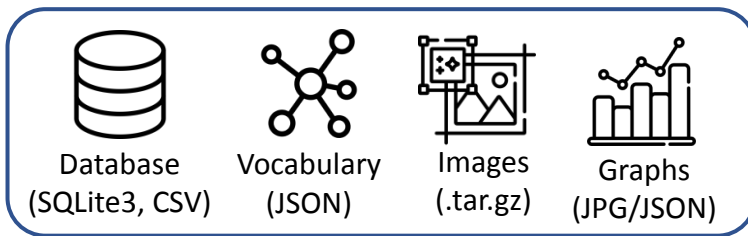


Figure 1: IMPatientT web application workflow and organisation

(A) Standard Vocabulary Tree

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

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

(B) Vocabulary Properties

Vocabulary ID
MHO:000124

Vocabulary Name
Peripheral Cores

Alternative Language
Core Périphériques x

Synonyms
Synonyms

Show as Image Annotation Class

Associated HPO Terms (Extracted from reports)

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

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

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

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

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

- (A) The hierarchical structure viewer and editor tool that supports drag and drop modification and creation/deletion/modification using the mouse.
- (B) The properties of the selected term node with its unique ID, display name, alternative language translation, synonyms, description, associated genes and diseases and correlating terms extracted from the application instance database.

(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 aggregates** on the periphery.

Vocabulary Automatically Detected as Present Vocabulary Automatically Annotated (Negated)

(B) Patient Informations

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

Patient age at biopsy: 40 Diagnosed Gene (HGNC API): 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

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: Abnormal (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 !

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

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

Save to Database

Figure 3: Screenshot of the report digitalization module.

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

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

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

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

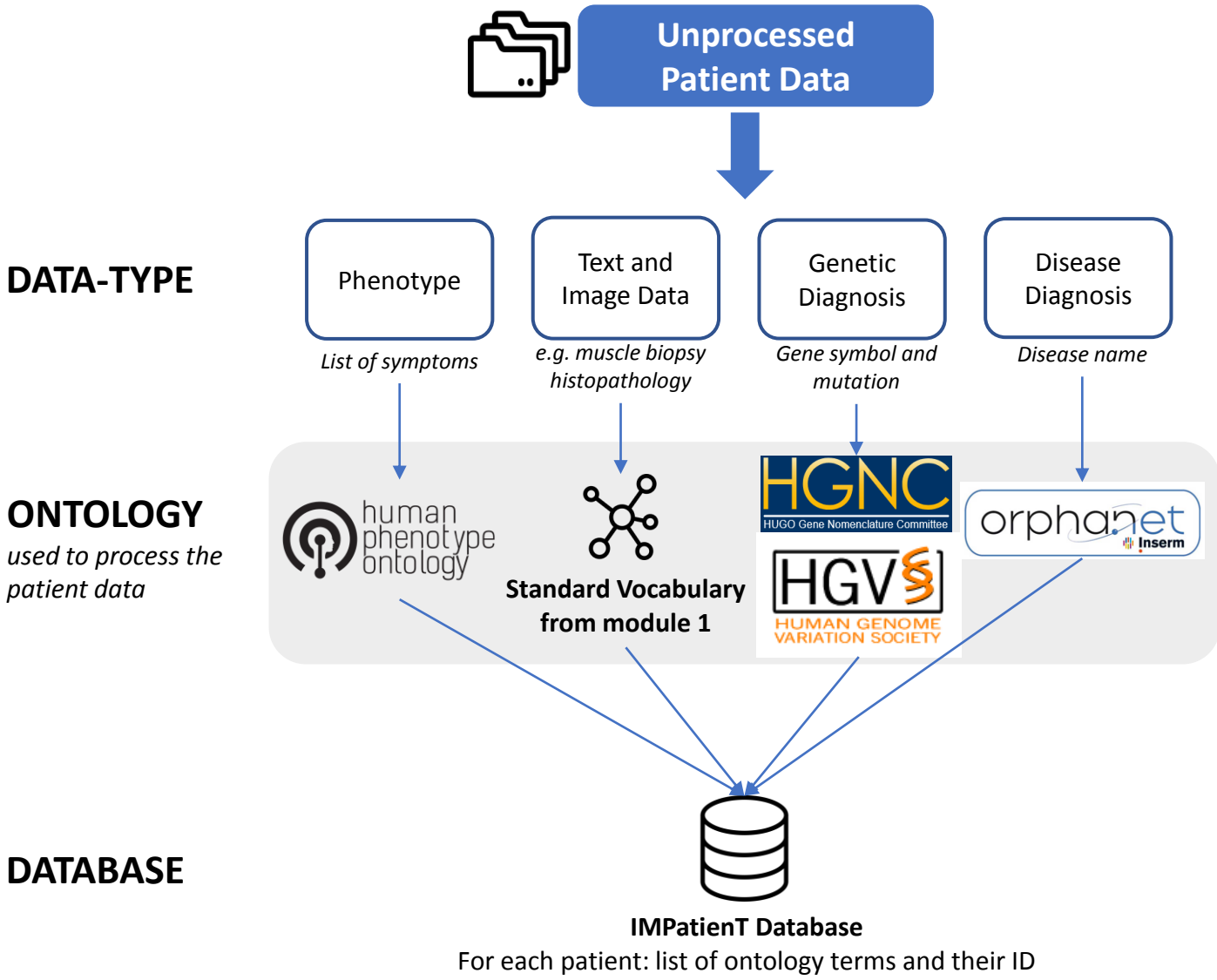
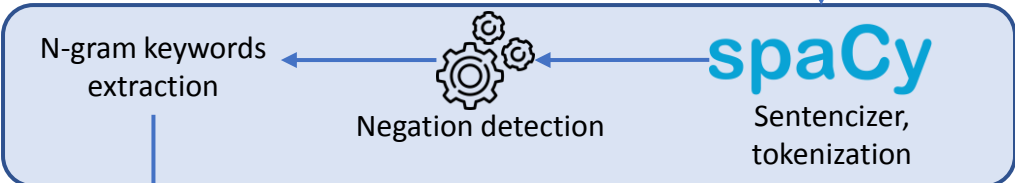


Figure 4: Overview of the ontologies used by IMPatientT to process patient data in the report digitization module (module 2).

PDF text detection



Keywords extraction



Keywords to vocabulary matching

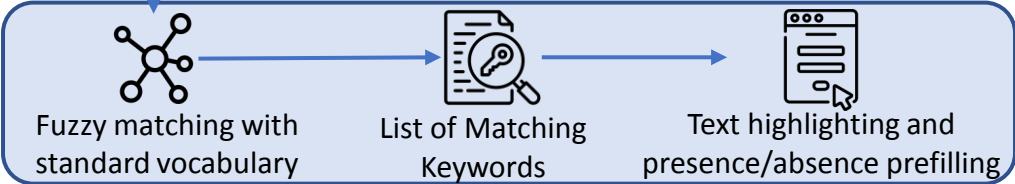
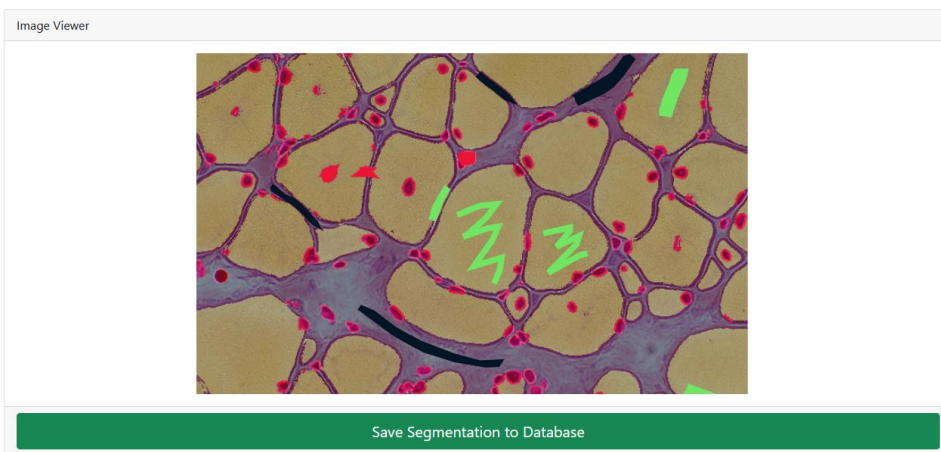


Figure 5: Optical character recognition and vocabulary term detection method used in the report digitization module (module 2) to automatically analyse free-text reports.

(A)



(B)

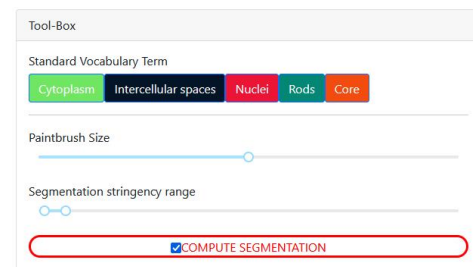


Figure 6: Screenshot of the image annotation module.

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

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

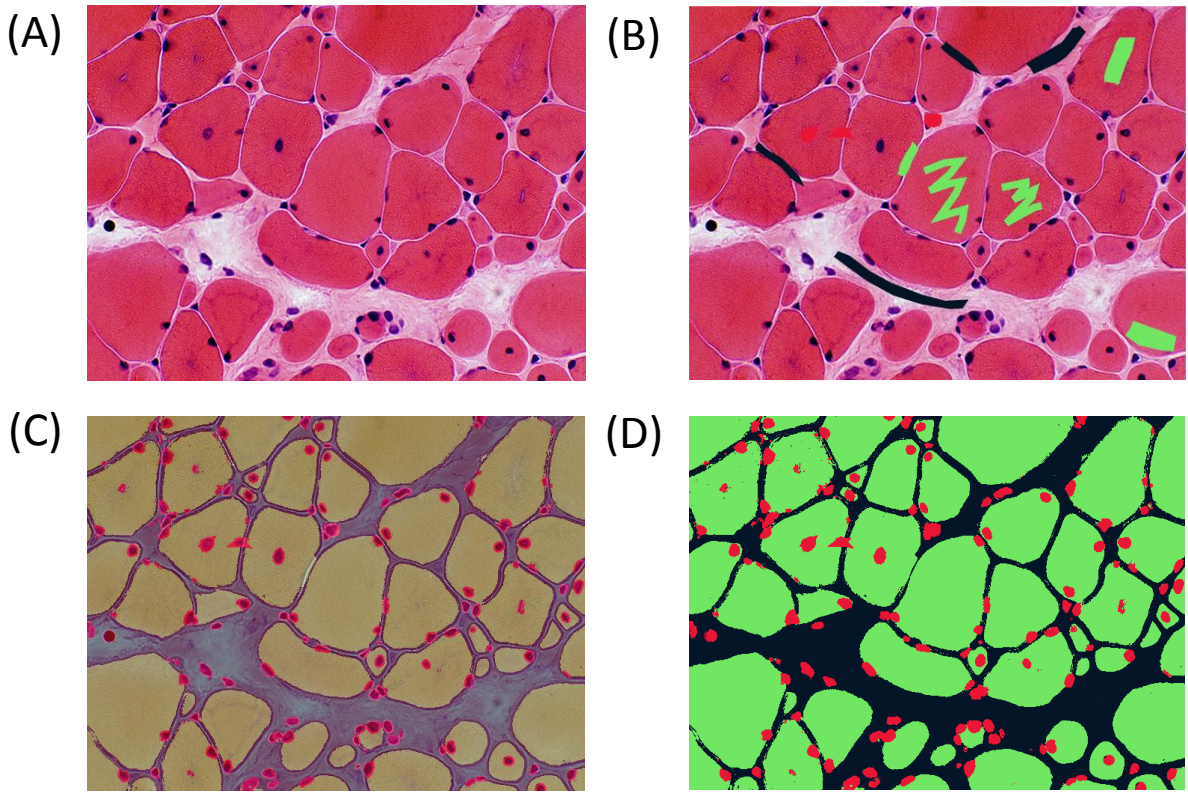


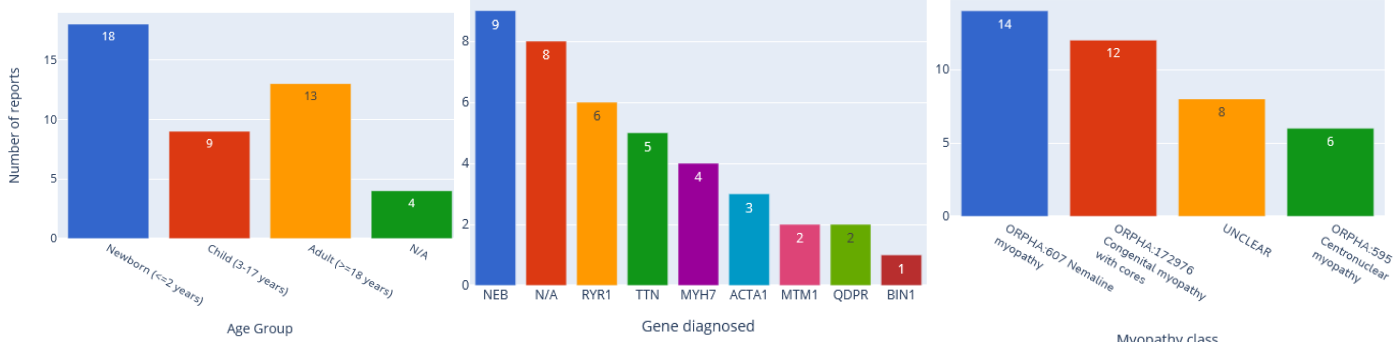
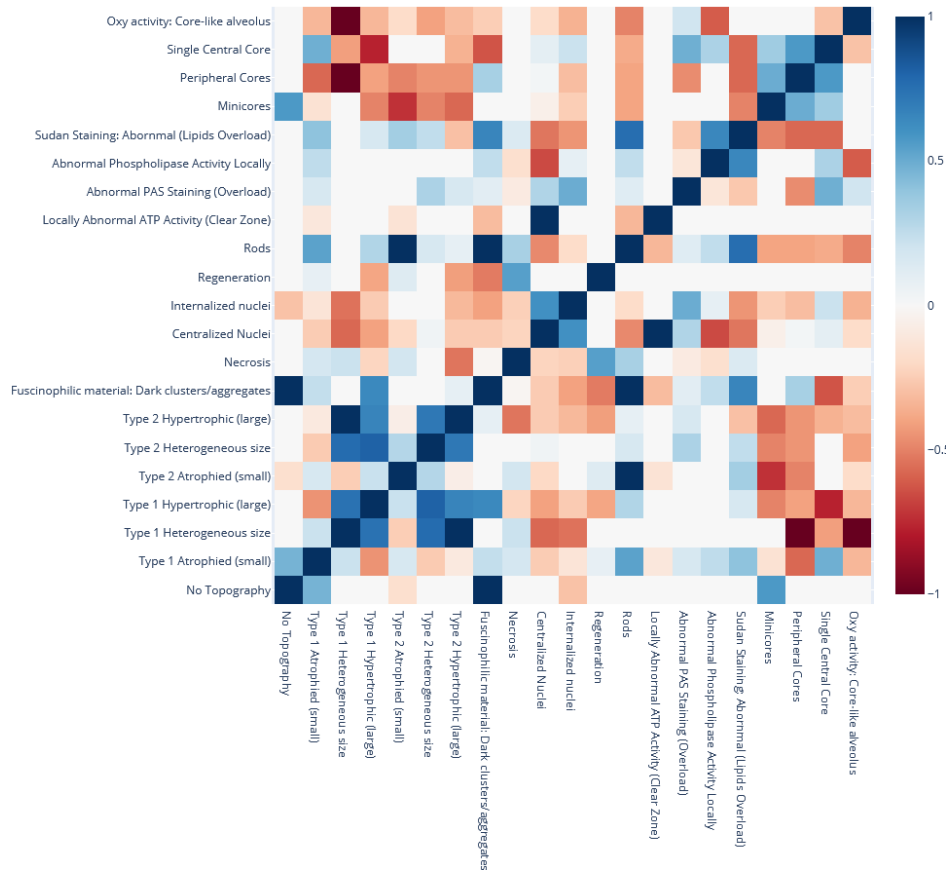
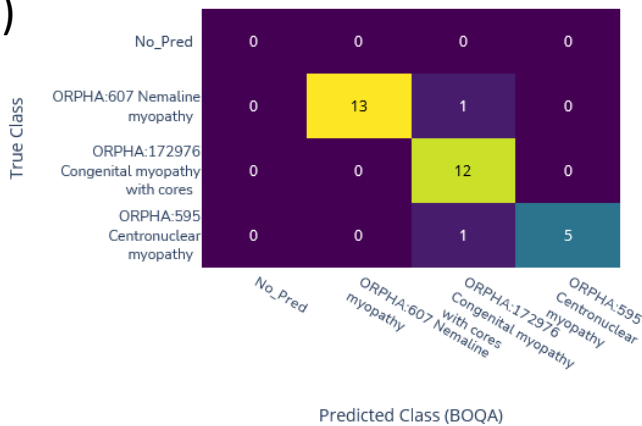
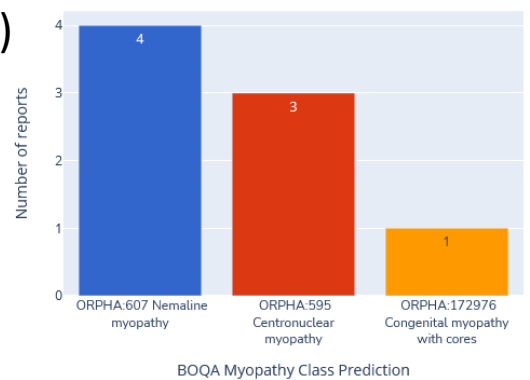
Figure 7: Image segmentation process in the image segmentation module.

(A) Raw image input before annotation.

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

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

(D) Segmented image mask alone.

(A)**(B)****(C)****(D)****Figure 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).