Common Infrastructure for National Cohorts in Europe, Canada, and Africa
- CINECA -

Deliverable D5.4 - Curation-support application for care planning

Work Package:	WP5 - Healthcare Interoperability and Clinical Applications
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leader(s):	Andrew Stubbs (EMC), Patrick Ruch (HESSO)
Contributing Partner(s):	
Contractual Delivery Date:	30th April 2023
Actual Delivery Date:	30th April 2023
Authors of this Deliverable:	Patrick Ruch, Nona Naderi, Luc Mottin, Emilie Pasche, Anais Mottaz
Contributors:	Nona Naderi, Emilie Pasche, Anaïs Mottaz, Luc Mottin, Patrick Ruch
Reviewed by:	Dylan Spalding
Approved by:	Thomas Keane
Dissemination Level:	Public
Type of Deliverable:	Report
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	54 months

Table of contents:

1. Executive Summary

This deliverable aims at designing and evaluating a curation support system to help clinicians to decide which treatments should be considered given a set of somatic/germline mutations and some particular clinical conditions (e.g. diagnosis). It builds on top of previous deliverables from WP5, in particular D5.1 [1], as well as on the services delivered by WP1-3, such as the WP1 dataset search services and WP3 content normalization services. In the use-case scenario described in D5.1, the end-user was first invited to select some datasets. User queries are then distributed across a federated Beacon network, comprising data from UKBB (EMBL-EBI), CoLaus (Switzerland), H3 (South Africa) and CHILD (Canada). Beyond that, the search pipeline now integrates subtask T5.3.2 (Scoring service to assess pathogenicity scales of variants) directly as an information displayed to the user, but also as a feature of the treatment recommendation function. The scoring and care planning services are evaluated based on various scoring functions (e.g., cohort distributions, SIFT, Polyphen, literature counts). The evaluation shows a moderate association between the pathogenetic scoring proposed by T5.3.2 and the baseline scores selected for comparison. We also report on the efficiency of the pipeline.

2. Introduction and objectives

The main objective of WP5 within the CINECA project is to develop clinical application scenarii to leverage CINECA's services. D5.4 in particular combines WP1 (data retrieval) and WP3 (content normalization) services to support treatment recommendation based on a list of clinical variants. The pipeline applies to both germline and somatic variants to deliver some personalized treatment recommendations based on variant pathogenicity scales and direct look up to latest biomedical literature (see Fig. 1 for the connection of the different services of WP5). Relying on Variomes, a variant search engine developed by the SIB to support the curation of genomic variants [2], the T5.3.2 service is able to assess pathogenicity scales of variants. On top of this, and by screening the literature with filters such as the disease, the age or the gender of the patient, T5.4 is able to provide a customized treatment plan that takes into account the characteristics of each individual. The typical end user of the integrated services could be a bioinformatician, embedded into an oncology or rare disease healthcare department, who would be preparing reports for tumor board discussions or genetic counseling, or more directly a biocurator at CiViC or UniProt [18,19]. Fig. 2 is a diagram representing, among others, the database architecture in Beacon V1.1. In D5.4, we use a similar structure to link prescription data with the variant information.

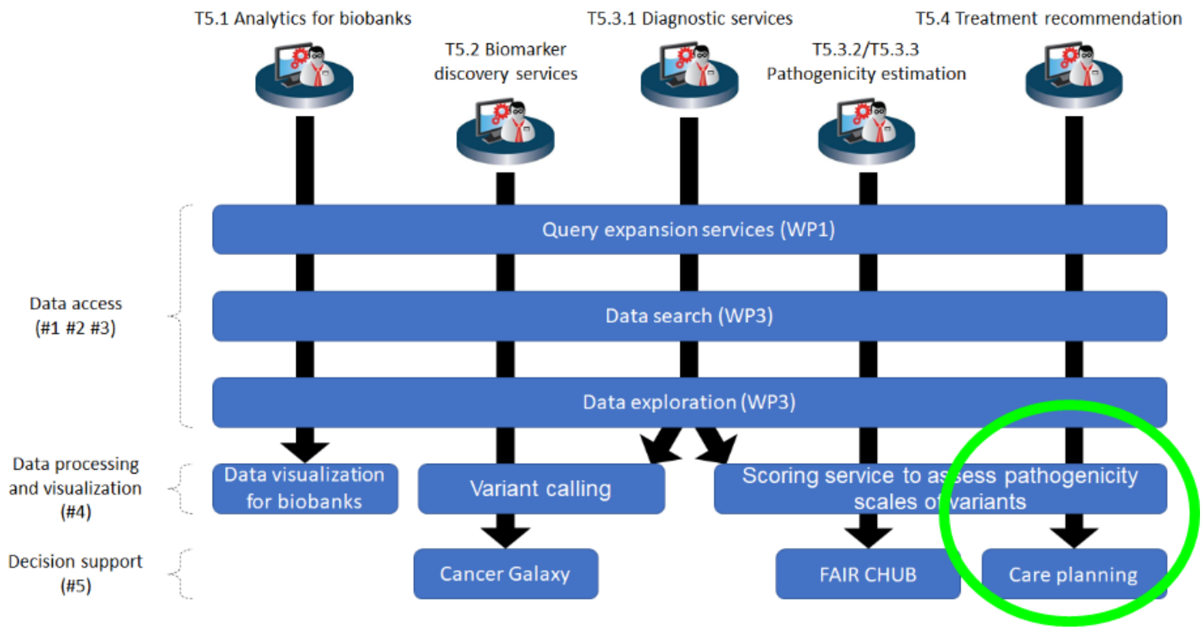

T5.1 Analytics for biobanks
T5.2 Biomarker discovery services
T5.3.1 Diagnostic services
T5.3.2/T5.3.3 Pathogenicity estimation
T5.4 Treatment recommendation
Data access (#1 #2 #3)
Query expansion services (WP1)
Data search (WP3)
Data exploration (WP3)
Data processing and visualization (#4)
Data visualization for biobanks
Variant calling
Scoring service to assess pathogenicity scales of variants
Decision support (#5)
Cancer Galaxy
FAIR CHUB
Care planning

Figure 1: Interoperability of the services developed by WP5.

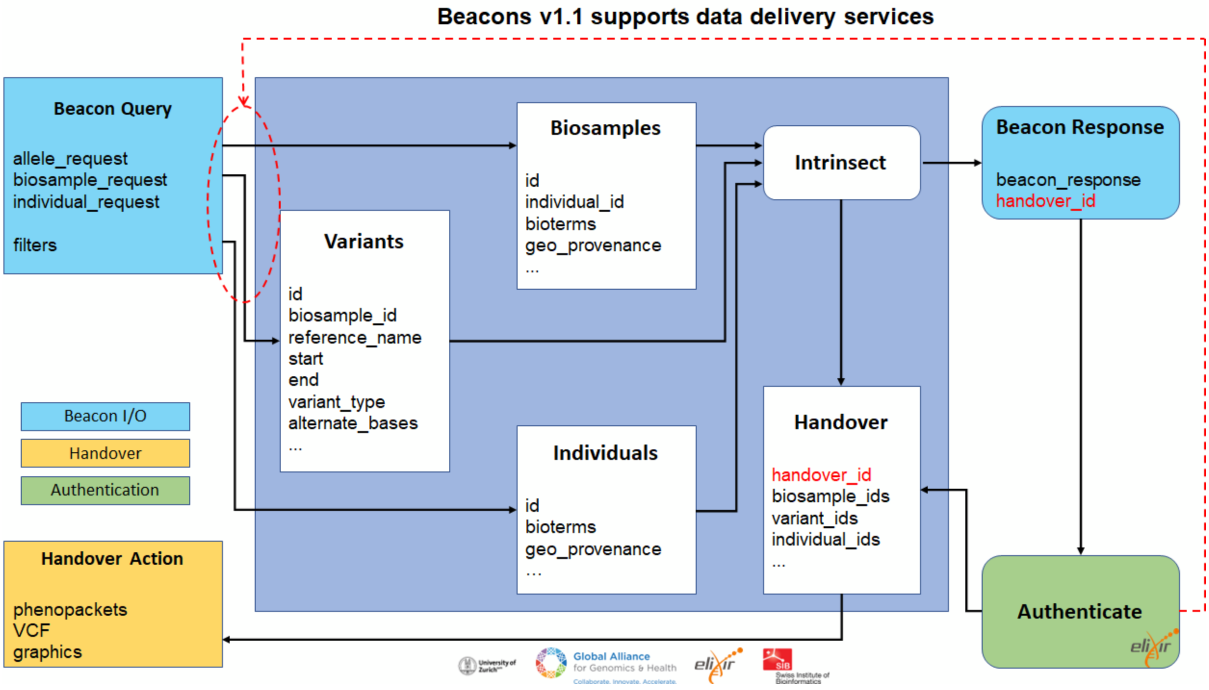


Figure 2: Beacon’s architecture with handover_id, which are needed to support the linking of data beyond variants and samples. D5.4 uses such cross-references to integrate prescription data.

3. Use-cases

Two main scenarios are explored in this report: 1) on request curation (interactive or not), which associate a pathogenicity scale and a list of possible treatments to a particular variant;

2) prioritization of a list of variant, as obtained from a subset of VCF files: ranking per pathogenicity scales and then association with the top recommended treatment.

The first scenario could be interactive or not, while the second is asynchronous due to time-consuming processing of large VCF contents. It is worth observing that the implementation of the interactive model remains relatively challenging since the screening of the literature, which provides the opportunity to evaluate the WP1's query expansion services, may require up to two minutes, for highly frequent variants, in the current settings. Hopefully, the search is usually faster as the majority of variants, as found in resources such as ClinVar, are classified as VUS (Variants of Uncertain Significance), i.e. variants which have not been properly characterized or intensively studied, and are infrequently mentioned in scientific literature with an average of one document per variant within the entire PubMed Central (PMC) database.

3.1. Characterization of the pathogenicity

One of the purposes of our system is to provide characterisation of the pathogenicity of variants, i.e. the potential impact of a genetic variant on an organism's ability to cause disease. Understanding the pathogenicity of a genetic variant will guide the subsequent treatment recommendation service.

Our estimation of pathogenicity is based on two standards named SIFT and PolyPhen [7,8]. To simplify matters, a SIFT score predicts whether an amino acid substitution affects protein function while the PolyPhen score (aka. PolyPhen-2 score) predicts the possible impact of an amino acid substitution on the structure and function of a human protein. PolyPhen-2 and SIFT scores use the same range, 0.0 to 1.0, but with opposite meanings.

This function of characterizing the pathogenicity of a variant can be operated independently of the exploratory research process as described in D5.1. However, its primary use is as a subsequent step to the selection of datasets of interest in the full workflow. If the user has not yet obtained access to one or more datasets (DAC contact step), he can land on this panel to continue his study. As a curation support service, it aims at providing information such as the literature related to his research, as well as the occurrences of his input variant in the data available on the network (simulated here only with the 4 CINECA synthetic datasets).

For the services described below, some optional demographic or clinical information can be supplied in the input including age or gender. These advanced query fields are filters to pre-select a normal-population cohort and to obtain prevalence of mutations.

S1 - On request variant pathogenicity estimate

The first use-case explores an “on request” curation task as intended for task T5.3.2. In this basic workflow, we seek to provide the researcher with information of interest for a particular variant. Thus, the user just needs to fill in variables such as the gene and the variant, and obtain in return a pathogenicity scale based on specific criteria (such as ASCO, CAP, AMP… see details in paragraph 4.1) [4]. In connection with the dataset exploration described in D5.1.

S2 - Ranking of a set of variants

Similarly to S1, this second use-case is intended to extend the user's search to a larger list and to sort the given variants according to their pathogenicity scores (SIFT/PolyPhen).

3.2. Recommendation of treatments

The other major objective of D5.4 is to provide the medical expert with a list of relevant treatments for a given variant. Thus, we built a second set of services based on the characterization of pathogenicity described above.

S3 - Literature curation

By screening the literature, or any cohort containing prescription data, the researcher can expect to find a list of treatments of interest for a given actionable variant. Using the same kind of parameters as for the services of pathogenicity characterization, S3 therefore systematically searches through existing articles, trials, and other literature in order to identify evidence-based recommendations that can be used to inform the researcher's decision-making.

S4 - Curation based on the ranking of variants

This service is an extension of S3. It aims at providing a list of treatments extracted from the literature, from a list of variants provided as input by the user.

4. Data description

4.1 Pathogenicity scales

The ASCO/CAP/AMP guidelines aim to improve the accuracy and consistency of biomarker testing for cancer, and to provide clinicians with the information they need to make informed treatment decisions for their patients. These guidelines are regularly updated to reflect advances in the field of molecular testing and to ensure that they remain relevant and useful for clinicians. Concretely, the guidelines provide recommendations for the molecular testing of tumors, including the selection of appropriate tests, interpretation of test results, and reporting of findings to guide treatment decisions.

To characterize a variant with respect to pathogenicity, we rely on the ASCO/CAP/AMP guidelines, see below Figure 3 for the variant categorization. Besides, the literature search is based on the SIBiLS API, a comprehensive library, aggregating PMC, MEDLINE and ClinicalTrials.gov contents, which have been annotated with various common biomedical onto-terminologies [3].

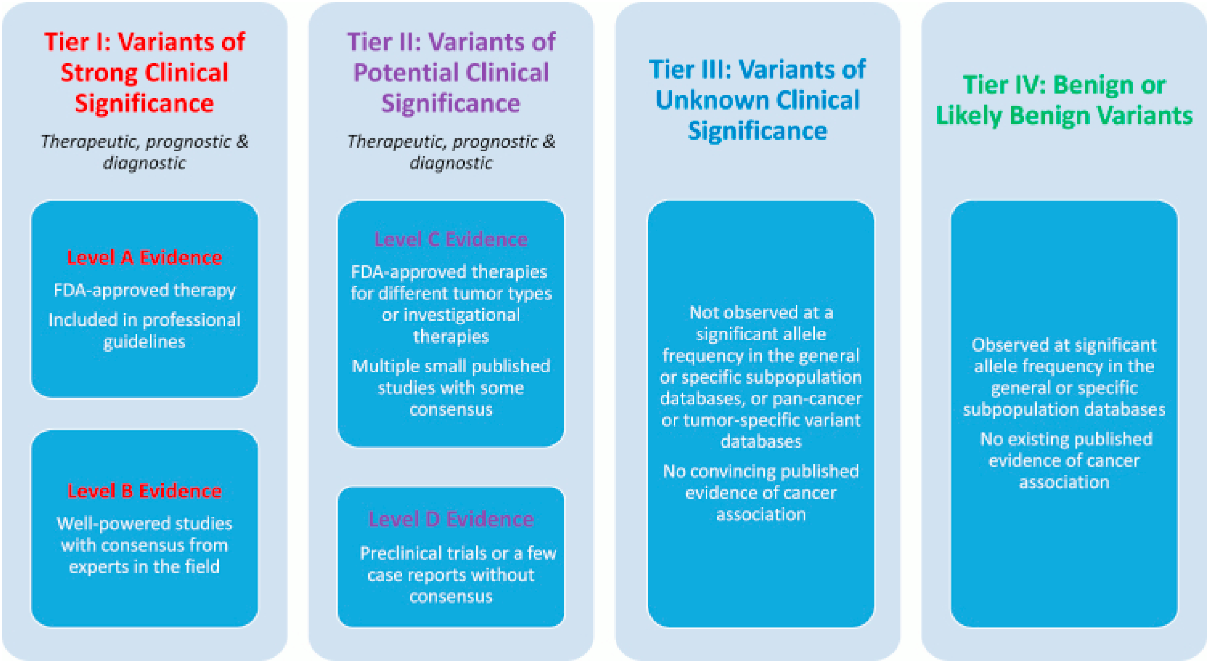

Tier I: Variants of Strong Clinical Significance
Therapeutic, prognostic & diagnostic
Level A Evidence
FDA-approved therapy
Included in professional guidelines
Level B Evidence
Well-powered studies with consensus from experts in the field
Tier II: Variants of Potential Clinical Significance
Therapeutic, prognostic & diagnostic
Level C Evidence
FDA-approved therapies for different tumor types or investigational therapies
Multiple small published studies with some consensus
Level D Evidence
Preclinical trials or a few case reports without consensus
Tier III: Variants of Unknown Clinical Significance
Not observed at a significant allele frequency in the general or specific subpopulation databases, or pan-cancer or tumor-specific variant databases
No convincing published evidence of cancer association
Tier IV: Benign or Likely Benign Variants
Observed at significant allele frequency in the general or specific subpopulation databases
No existing published evidence of cancer association

Figure 3. Evidence-based variant categorization from ASCO/CAP/AMP guidelines.

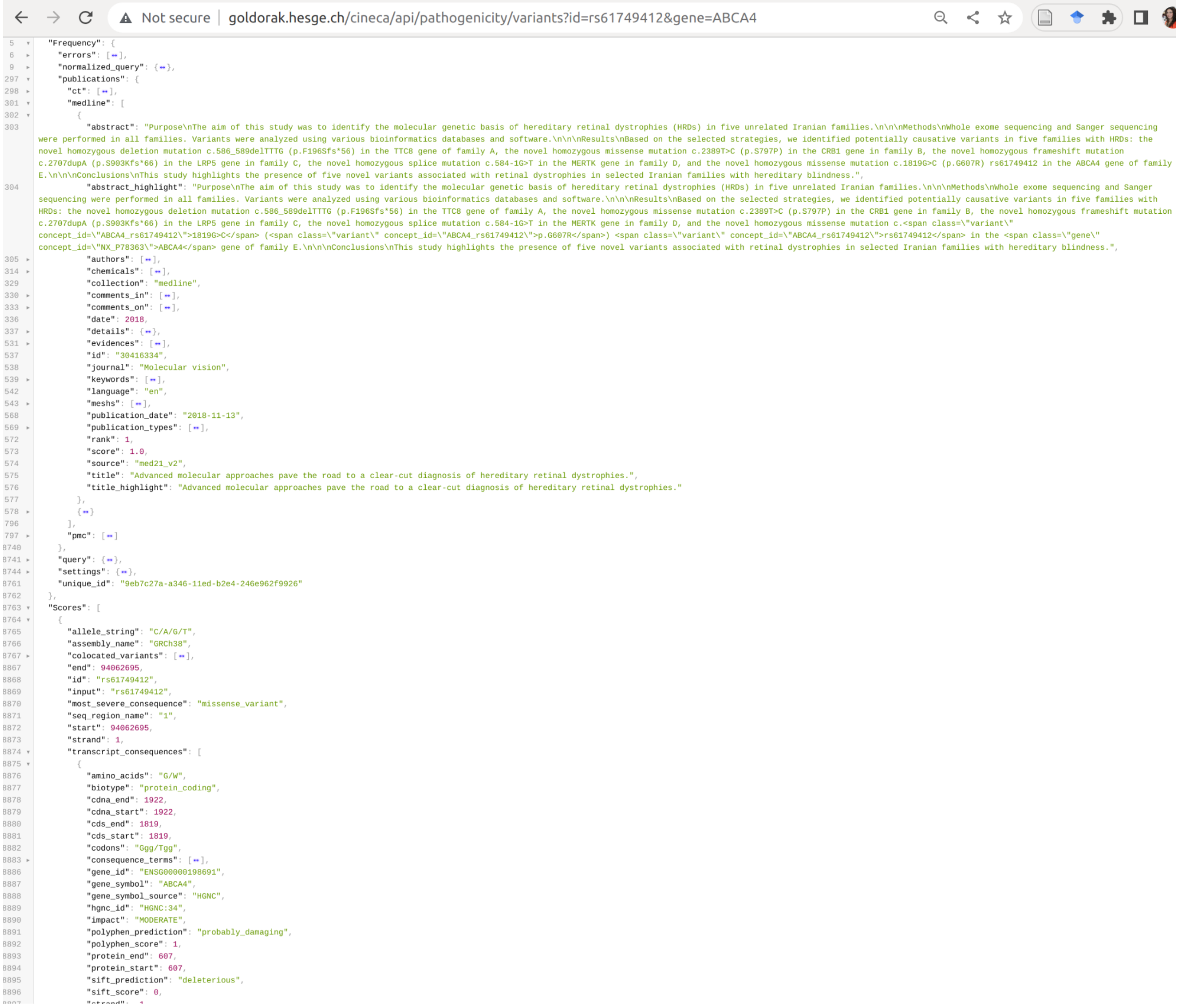
4.2 Methodology for pathogenicity assessment

In order to estimate the pathogenicity of the variants, we rely on the following approach:
1) we count the frequency of the variant in the selected cohorts,
2) we compute the frequency of the variants in the biomedical literature, including in the Medline, PubMed Central (PMC), Clinical Trials (CT), and in supplementary materials (SUPP) of the publications using Variomes API[1] [2]
3) we retrieve the SIFT [7] and PolyPhen-2 (Polymorphism Phenotyping V2) [8] scores from Ensembl Variant Effect Predictor (VEP) using the Ensembl API[2] [6].

In case no response was received from VEP, we call the WP1's SynVar service[3] [11] to get the other formats of variants and then query VEP again. The pathogenicity service is available at http://goldorak.hesge.ch/cineca/api/pathogenicity/about.

An example of pathogenicity scores from this service is shown in Figure 4.

[1] https://candy.hesge.ch/Variomes/
[2] http://www.ensembl.org/index.html
[3] https://goldorak.hesge.ch/synvar/


Not secure
goldorak.hesge.ch/cineca/api/pathogenicity/variants?id=rs61749412&gene=ABCA4

```
  "Frequency": {
    "errors": [↔],
    "normalized_query": {↔},
    "publications": {
      "ct": [↔],
      "medline": [
        {
          "abstract": "Purpose\nThe aim of this study was to identify the molecular genetic basis of hereditary retinal dystrophies (HRDs) in five unrelated Iranian families.\n\n\nMethods\nWhole exome sequencing and Sanger sequencing were performed in all families. Variants were analyzed using various bioinformatics databases and software.\n\n\nResults\nBased on the selected strategies, we identified potentially causative variants in five families with HRDs: the novel homozygous deletion mutation c.586_589delTTTG (p.F196Sfs*56) in the TTC8 gene of family A, the novel homozygous missense mutation c.2389T>C (p.S797P) in the CRB1 gene in family B, the novel homozygous frameshift mutation c.2707dupA (p.S903Kfs*66) in the LRP5 gene in family C, the novel homozygous splice mutation c.584-1G>T in the MERTK gene in family D, and the novel homozygous missense mutation c.1819G>C (p.G607R) rs61749412 in the ABCA4 gene of family E.\n\n\nConclusions\nThis study highlights the presence of five novel variants associated with retinal dystrophies in selected Iranian families with hereditary blindness.",
          "abstract_highlight": "Purpose\nThe aim of this study was to identify the molecular genetic basis of hereditary retinal dystrophies (HRDs) in five unrelated Iranian families.\n\n\nMethods\nWhole exome sequencing and Sanger sequencing were performed in all families. Variants were analyzed using various bioinformatics databases and software.\n\n\nResults\nBased on the selected strategies, we identified potentially causative variants in five families with HRDs: the novel homozygous deletion mutation c.586_589delTTTG (p.F196Sfs*56) in the TTC8 gene of family A, the novel homozygous missense mutation c.2389T>C (p.S797P) in the CRB1 gene in family B, the novel homozygous frameshift mutation c.2707dupA (p.S903Kfs*66) in the LRP5 gene in family C, the novel homozygous splice mutation c.584-1G>T in the MERTK gene in family D, and the novel homozygous missense mutation c.<span class=\"variant\" concept_id=\"ABCA4_rs61749412\">1819G>C</span> (<span class=\"variant\" concept_id=\"ABCA4_rs61749412\">p.G607R</span>) <span class=\"variant\" concept_id=\"ABCA4_rs61749412\">rs61749412</span> in the <span class=\"gene\" concept_id=\"NX_P78363\">ABCA4</span> gene of family E.\n\n\nConclusions\nThis study highlights the presence of five novel variants associated with retinal dystrophies in selected Iranian families with hereditary blindness.",
          "authors": [↔],
          "chemicals": [↔],
          "collection": "medline",
          "comments_in": [↔],
          "comments_on": [↔],
          "date": 2018,
          "details": {↔},
          "evidences": [↔],
          "id": "30416334",
          "journal": "Molecular vision",
          "keywords": [↔],
          "language": "en",
          "meshs": [↔],
          "publication_date": "2018-11-13",
          "publication_types": [↔],
          "rank": 1,
          "score": 1.0,
          "source": "med21_v2",
          "title": "Advanced molecular approaches pave the road to a clear-cut diagnosis of hereditary retinal dystrophies.",
          "title_highlight": "Advanced molecular approaches pave the road to a clear-cut diagnosis of hereditary retinal dystrophies."
        },
        {↔}
      ],
      "pmc": [↔]
    },
    "query": {↔},
    "settings": {↔},
    "unique_id": "9eb7c27a-a346-11ed-b2e4-246e962f9926"
  },
  "Scores": [
    {
      "allele_string": "C/A/G/T",
      "assembly_name": "GRCh38",
      "colocated_variants": [↔],
      "end": 94062695,
      "id": "rs61749412",
      "input": "rs61749412",
      "most_severe_consequence": "missense_variant",
      "seq_region_name": "1",
      "start": 94062695,
      "strand": 1,
      "transcript_consequences": [
        {
          "amino_acids": "G/W",
          "biotype": "protein_coding",
          "cdna_end": 1922,
          "cdna_start": 1922,
          "cds_end": 1819,
          "cds_start": 1819,
          "codons": "Ggg/Tgg",
          "consequence_terms": [↔],
          "gene_id": "ENSG00000198691",
          "gene_symbol": "ABCA4",
          "gene_symbol_source": "HGNC",
          "hgnc_id": "HGNC:34",
          "impact": "MODERATE",
          "polyphen_prediction": "probably_damaging",
          "polyphen_score": 1,
          "protein_end": 607,
          "protein_start": 607,
          "sift_prediction": "deleterious",
          "sift_score": 0,
```

Figure 4: An example of pathogenicity scores for variant rs61749412 and gene ABCA4 from pathogenicity service.

4.3 Treatment recommendation data

Treatment related data are extracted from the literature - or any cohort containing some prescription data - using two ontologies, ATC and Drugbank, to ensure consistent annotations as well as the standardization of the information [12,13].

5. Results

5.1. System architecture

On the front-end, the T5.4 demonstrator is provided with a graphical user interface inspired by D5.1 (cf. screenshots in [1] paragraph 4.1). To limit the burden of processing several knowledge bases (especially for S1 and S2), the search of several variants is carried out in parallel and the results are displayed progressively in the output table. On the back-end, all the services are developed as REST APIs and provide a JSON output. For real non-synthetic data, all services will run over HTTPS protocol. S1 and S2 services are intensive computing as they have to communicate with external databases such as Variomes and Ensembl to obtain relevant SIFT/PolyPhen scores and literature counts. However, both also have a built-in cache so that not all information is reprocessed. S3 and

S4, meanwhile, rely on the frequency of occurrence of drug treatment in the scientific literature associated with a variant. The user/system interaction is described below, using the Unified Modeling Language (UML). In Figure 5, we visualize the user, the exploration system and the services used to estimate the pathogenicity of a variant and to rank a list of relevant treatments.

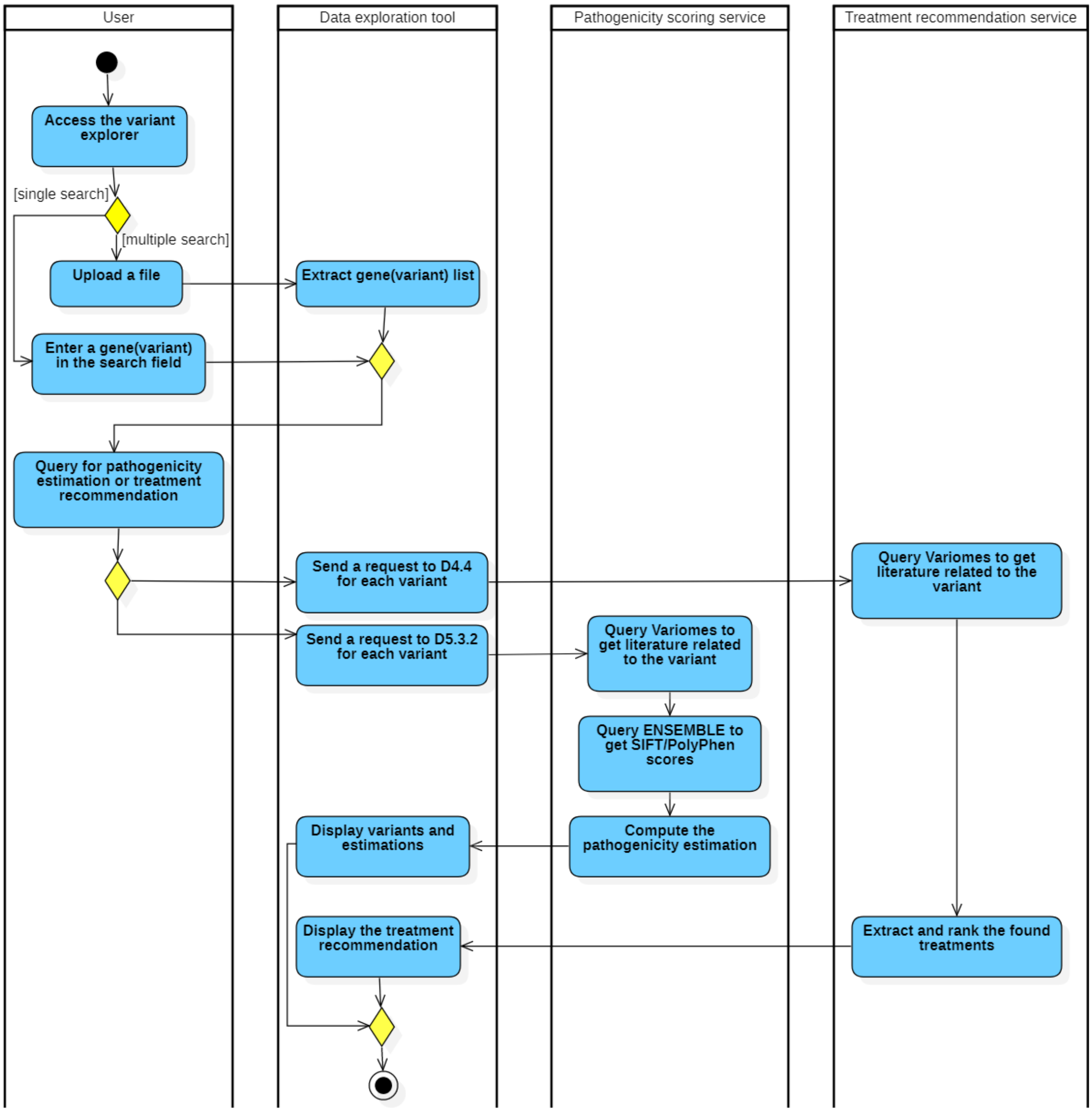
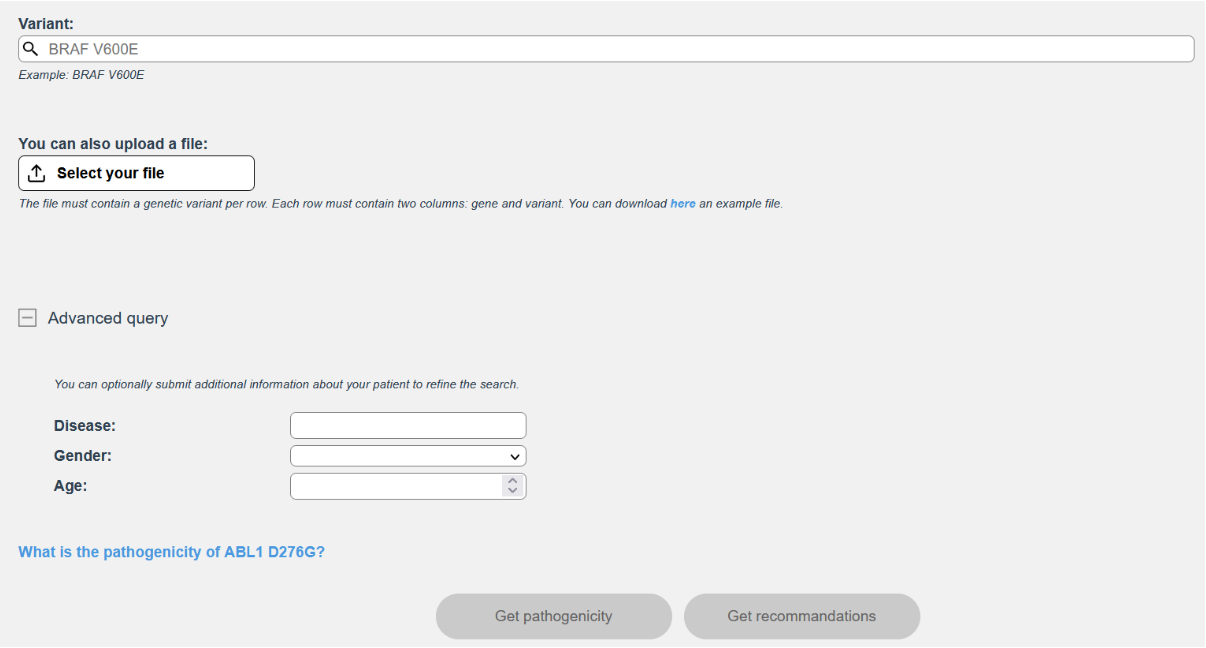

User
Data exploration tool
Pathogenicity scoring service
Treatment recommendation service
Access the variant explorer
[single search]
[multiple search]
Upload a file
Extract gene(variant) list
Enter a gene(variant) in the search field
Query for pathogenicity estimation or treatment recommendation
Send a request to D4.4 for each variant
Query Variomes to get literature related to the variant
Send a request to D5.3.2 for each variant
Query Variomes to get literature related to the variant
Query ENSEMBLE to get SIFT/PolyPhen scores
Display variants and estimations
Compute the pathogenicity estimation
Display the treatment recommendation
Extract and rank the found treatments

Figure 5 : Activity diagram of the two services: pathogenicity estimation and treatment recommendation.

5.2. User interfaces

When landing on the D5.4 demonstrator[4], the user can perform a basic search by typing the query of his choice in the “Variant” search field (Figure 6). Optionally, he can use the fields available in the advanced search as filters, to limit the search to a typical population.

Gender and Age are restricted to normalized values from CINECA synthetic datasets. However, as the diseases are represented in different ways in the CINECA synthetic datasets (see [20]), we preferred to take advantage of our local indexing of these datasets to offer the user a free text field in which he can enter information in the form he wants. Since filters are restrictive by nature (as opposed to an exploratory search field), we estimate that here the user will prefer precision over recall (and limit the research to exact matches).



Variant: BRAF V600E

Example: BRAF V600E

You can also upload a file:

The file must contain a genetic variant per row. Each row must contain two columns: gene and variant. You can download here an example file.

⊟ Advanced query

You can optionally submit additional information about your patient to refine the search.

Disease:

Gender:

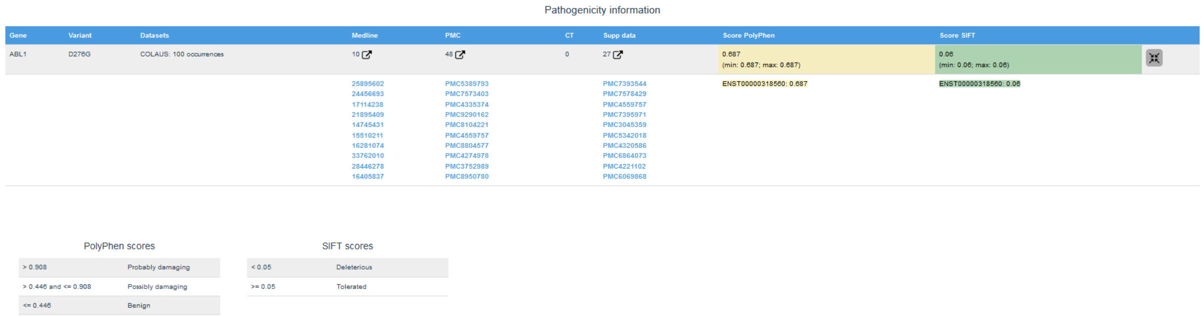
Age:

What is the pathogenicity of ABL1 D276G?

Figure 6: Search interface for the Variant exploration.

After a search, the user is redirected to a results page displaying the calculated SIFT and PolyPhen scores for the given search (Figure 7). Together with the pathogenicity scores, we propose contextual information to support the user's access to evidence, and to allow him/her to validate the predicted level of pathogenicity: counts in the initially selected cohorts, number of related documents per collection, link to access the main sources, and link to the equivalent query via Variomes for more details. A small frame shows the pathogenicity scale and a color code helps to quickly screen through the results. One last column (which is named "Datasets”) connects the mention of this variant with the cohort where it is found. As for anonymity recommendations in Beacons [14-16], we apply k-anonymity to this field [17]. Thus, rare occurrences (below 5) in a given cohort are not displayed.

[4] https://denver.hesge.ch/cineca/demo/step6



Pathogenicity information

Gene	Variant	Datasets	Medline	PMC	CT	Supp data	Score PolyPhen	Score SIFT
ABL1	D276G	COLAUS: 100 occurrences	10	48	0	27	0.687 (min: 0.687; max: 0.687)	0.06 (min: 0.06; max: 0.06)
			25895602 24456693 17114238 21895409 14745431 15510211 16281074 33762010 28446278 16405837	PMC5389793 PMC7573403 PMC4335374 PMC9290162 PMC8104221 PMC4559757 PMC8804577 PMC4274978 PMC3752989 PMC8950780		PMC7393544 PMC7578429 PMC4559757 PMC7395971 PMC3045359 PMC5342018 PMC4320586 PMC6864073 PMC4221102 PMC6069868	ENST00000318560: 0.687	ENST00000318560: 0.06

PolyPhen scores

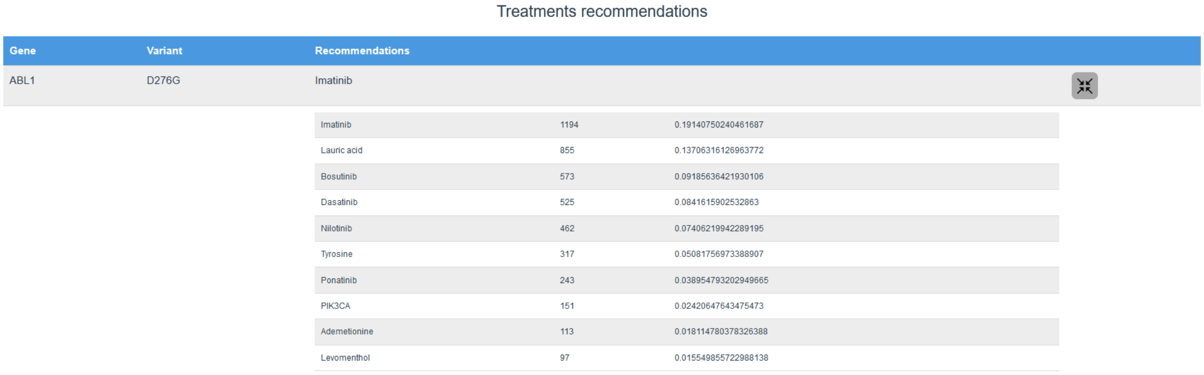
> 0.908	Probably damaging
> 0.446 and <= 0.908	Possibly damaging
<= 0.446	Benign

SIFT scores

< 0.05	Deleterious
>= 0.05	Tolerated

Figure 7: Information on the pathogenicity of the ABL1 variant (D276G). Here, the display shows that this variant has been found 100 times in the COLAUS (synthetic) dataset, 10 times in MEDLINE, 48 times in PMC, 27 times in supplementary materials, and never in clinical trials.

When the user is interested in obtaining a treatment recommendation, the display is simpler as it only provides a ranking of treatments as commonly available in clinical settings (e.g. first line vs. second line treatments). The user can choose to extend the display to obtain the top 10 treatments returned, as well as their actual scores and frequencies of appearance (see Fig. 8).



Treatments recommendations

Gene	Variant	Recommendations
ABL1	D276G	Imatinib

Imatinib	1194	0.19140750240461687
Lauric acid	855	0.13706316126963772
Bosutinib	573	0.09185636421930106
Dasatinib	525	0.0841615902532863
Nilotinib	462	0.07406219942289195
Tyrosine	317	0.05081756973388907
Ponatinib	243	0.038954793202949665
PIK3CA	151	0.02420647643475473
Ademetionine	113	0.018114780378326388
Levomenthol	97	0.015549855722988138

Figure 8: Recommendation of treatments associated in the literature with the ABL1 variant (D276G).

5.3. Application programming interfaces

Pathogenicity of a variant

API endpoint: goldorak.hesge.ch/cineca/api/pathogenicity/variants?

Input: the gene/variant combination using respectively the parameters “gene” and “id”.

Example : to search SPATA7 (V600E)

http://goldorak.hesge.ch/cineca/api/pathogenicity/variants?id=V600E&gene=SPATA7

Output:
JSON object including two children “Frequency” and “Scores”.

- Frequency provides information on the literature available on this variant (MEDLINE, PMC, Clinical Trials, and supplementary materials)
- Scores relates to the computed SIFT andPolyPhen scores

Treatment recommendation for a variant

API endpoint: denver.hesge.ch/CINECA_recommendation_proxy/treatments?

Input: the same combination as for the pathogenicity service, using respectively the parameters “gene” and “id”.

Example : to search AKT1 (E17K)

https://denver.hesge.ch/CINECA_recommendation_proxy/treatment?id=E17K&gene=AKT1

Output:
A JSON file including the jsonArray “Treatment_recommendation” which includes for each entity:

- treatment: the normalized name of the treatment as found in the literature, in the same documents that talk about this gene/variant
- count: the number of occurrence of this treatment in the documents concerned
- score: a confidence score which reflects the relevance of this drug is relevant for the treatment of the searched gene/variant

5.4. Evaluation

In this section, we tentatively explore the performances - efficiency and effectiveness - of the query distribution system. We also sought to assess the validity of our pathogenicity scoring method.

5.4.1. Evaluation of the pathogenicity scoring function

The goal of this evaluation was to compute the correlation between the frequency metrics and each of the SIFT and PolyPhen scores. To this end, we use the data of Varibench[5] [9] provided by GAVIN[6] [10].
The correlation between the frequency measures and SIFT and Polyphen scores is computed using Pearson correlation coefficient from SciPy package[7]. The table below shows the correlations for 10,124 variants. The results show positive weak correlations between the frequency measures and PolyPhen scores and negative weak correlation between the frequency and SIFT scores.

[5] http://structure.bmc.lu.se/VariBench/
[6] https://github.com/molgenis/gavin
[7] https://docs.scipy.org/doc/scipy/

	Medline	PMC	CT	SUPP	Polyphen
SIFT	-0.012	-0.022	-0.006	-0.019	-0.58
PolyPhen	0.033	0.036	0.017	0.021	1

Table 1. Rank correlation between two pathogenicity scores (SIFT and Polyphen) and variant counts in different literature repositories.

From table 1 and as expected, we observe that Polyphen and SIFT scores are inversely correlated (~0.58). More interestingly, when comparing the hit counts from the literature with either SIFT or Polyphen, we see that the highest correlation is obtained with PMC and MEDLINE with respectively 0.036 and 0.033. It suggests that variants occurring in full-text articles and abstracts are more likely pathogenic than variants found in Supplementary Data. This is expected because supplementary data files share more similarity with raw clinical data as stored in cohorts. The results obtained from ClinicalTrials.gov are more difficult to interpret and the correlation may just be an artifact due to the limited number of variants found in clinical trials.

5.4.2. Evaluation of the efficiency

Here, we examine the response time of calling of the services i.e. Variomes and Ensembl VEP. Response time is defined as the time duration between a user sending a request and receiving the corresponding response. This experiment is conducted by executing 1000 web service invocations. By processing the service invocation results, we obtained the results presented in the table below.

Service	Average Time (second)	Standard Deviation (second)	Throughput (kbps)
Ensemble VEP	1.21	0.92	4.35
Variomes	1.28	1.25	2105.5

Table 2. Response time of the query distribution services to four Beacons with external calls to Variomes and Ensemble VEP.

5.4.3. Evaluation of the effectiveness

Here, we investigate the throughput of the services, i.e. Variomes and Ensembl VEP. Throughput is defined as the average rate of successful message size delivery per second. The results for 1000 web service invocations are shown in Table 2. 16 out of 1000 queries were failed on Variomes with error “Read timed out” (the invocations were configured with a timeout of 1200 seconds).

6. Conclusion

With D5.3.2 and D5.4, we built services to aid the characterization of pathogenicity and the recommendation of treatment specific to variant using a combination of services and multimodal data (pathogenicity scales, published evidence, …). Both were implemented in the D5.1 demonstrator workflow which operates on top of WP1's discovery services and

WP3's content normalization. Given the heterogeneity of the selected cohorts, the implemented clinical applications focused on descriptive analysis: given a particular variant or a set of variants, a search performed and returned statistics such as cohort counts, literature hits and pathogenicity scales. When cohorts contain prescription data or for literature hits, it is possible to augment the pathogenicity scale with prescription data. Built on the top of this, the next service in the workflow provides the user with a list of treatments retrieved for the searched variant. The query distribution across the four cohorts is effective and scalability tests have been performed to evaluate the robustness of the system, both internally and by linking external services and data.

6. Possible next steps

The evaluation and improvement of the accuracy of the recommendations proposed by the D5.4 service should be of great interest. Currently, the recommendation service relies on mining for evidence in the literature [21,22] to provide the user with the up-to-date information about treatments that are supported by clinical trials, research papers, and supplementary materials. But we have also seen the significant progress of Bayesian techniques on such tasks. Thus, implementing it in the D5.4 service could be a sensible way to improve the service.

With regards to the evaluation, we are considering either LUCADA (a dataset created and maintained by the National Lung Cancer Audit with the aim of improving the outcome for people diagnosed with lung cancer and mesothelioma), or the various datasets available through recent challenges in precision medicine,

7. References

1. https://zenodo.org/record/6783295
2. Pasche E., Mottaz A., Caucheteur D. et al. (2022). Variomes: a high recall search engine to support the curation of genomic variants. Bioinformatics. 2022 Apr 28;38(9):2595-2601. DOI: 10.1093/bioinformatics/btac146.
3. Gobeill G., Caucheteur D., Michel P.A. et al. (2020). SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts, Nucleic Acids Research, Volume 48, Issue W1, 02 July 2020, Pages W12–W16, DOI: 10.1093/nar/gkaa328.
4. Li M.M., Datto M., Duncavage E.J. et al. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. Journal of Molecular Diagnosis, 19(1), 4-23. DOI: 10.1016/j.jmoldx.2016.10.002.
5. Wagner A.H., Walsh B., Mayfield G. et al. (2018). A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. https://www.biorxiv.org/content/10.1101/366856v2
6. McLaren, W., Gil, L., Hunt, S.E. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122 (2016). DOI: 10.1186/s13059-016-0974-4
7. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001 May;11(5):863-74. doi: 10.1101/gr.176601.

8. Adzhubei, I., Schmidt, S., Peshkin, L. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248–249 (2010). DOI: 10.1038/nmeth0410-248
9. Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013 Jan;34(1):42-9. doi: 10.1002/humu.22204. Epub 2012 Oct 11. PMID: 22903802.
10. van der Velde, K.J., de Boer, E.N., van Diemen, C.C. *et al.* GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biol* 18, 6 (2017). DOI: 10.1186/s13059-016-1141-7
11. Mottaz, A., Pasche, E., Michel, P. A. A., Mottin, L., Teodoro, D. & Ruch, P. (2022). Designing an Optimal Expansion Method to Improve the Recall of a Genomic Variant Curation-Support Service. Studies in health technology and informatics, 294, 839-843.
12. Nahler G. (2009). Anatomical Therapeutic Chemical classification system (ATC). Dictionary of Pharmaceutical Medicine. Springer, Vienna. DOI: 10.1007/978-3-211-89836-9_64.
13. Wishart D.S., Feunang Y.D., Guo A.C. et al. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research, 4, 46(D1), D1074-D1082. DOI: 10.1093/nar/gkx1037.
14. Beacon network link: https://beacon-network.elixir-europe.org/
15. Fiume, M., Cupak, M., Keenan, S. et al. Federated discovery and sharing of genomic data using Beacons. Nature Biotechnology 37, 220–224 (2019). DOI: 10.1038/s41587-019-0046-x.
16. Bernier, A., Liu, H. & Knoppers, B.M. Computational tools for genomic data de-identification: facilitating data protection law compliance. Nature Communications 12, 6949 (2021). DOI: 10.1038/s41467-021-27219-2.
17. Sepas A., Bangash A.H., Alraoui O. et al. (2022). Algorithms to anonymize structured medical and healthcare data: A systematic review. Frontiers in Bioinformatics, 2, 984807. DOI: 10.3389/fbinf.2022.984807.
18. Krysiak K, Danos AM, Saliba J, et al. CIViCdb 2022: evolution of an open-access cancer variant interpretation knowledgebase. Nucleic Acids Res. 2023 Jan 6;51(D1):D1230-D1241. DOI: 10.1093/nar/gkac979.
19. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052.
20. ⊞ AGM_2021_common_fields in synthetic datasets across cohorts https://docs.google.com/spreadsheets/d/1jDZAFq9GvysymJx_YdXYk0WsPQCtsbZmkZgC-esR0Ig/edit#gid=695195232
21. Pasche E, Gobeill J, Teodoro D et al. (2011) Using multimodal mining to drive clinical guidelines development. Studies in Health Technology and Informatics, 169, 477-481. DOI: 10.3233/978-1-60750-806-9-477.
22. Becker M, Böckmann B. (2017) Personalized Guideline-Based Treatment Recommendations Using Natural Language Processing Techniques. Studies in Health Technology and Informatics, 235, 271-275. DOI: 10.3233/978-1-61499-753-5-271.

8. Abbreviations

Abbreviation	Definition
AAI	Authentication and Authorization Infrastructure
AMP	Association for Molecular Pathology
API	Application Programming Interface
ASCO	American Society of Clinical Oncology
ATC	Anatomical Therapeutic Chemical
BB	BioBank
CAP	College of American Pathologists
CIViC	Clinical Interpretation of Variants in Cancer
CT	Clinical Trial
DAC	Data Access Committee
DAQ	Data acquisition
ES	ElasticSearch
GA4GH	The Global Alliance for Genomics and Health (https://www.ga4gh.org), a standards body for health genomics APIs and data models.
GECKO	Genomics Cohorts Knowledge Ontology, an ontology to represent genomics cohort attributes. Browse: https://www.ebi.ac.uk/ols/ontologies/gecko Description of GECKO's development: https://github.com/IHCC-cohorts/GECKO
GUI	Graphical User Interface
JSON	JavaScript Object Notation

LOVD	Leiden Open Variation Database
MeSH	Medical Subject Headings (https://www.nlm.nih.gov/mesh/meshhome.html) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine used for indexing, cataloging, and searching of biomedical and health-related information.
NCIt	National Cancer Institute Thesaurus
PolyPhen-2	Polymorphism Phenotyping v2
QE	Query Expansion
SIB	Swiss Institute of Bioinformatics
SIFT	Sorting Intolerant From Tolerant
UI	User Interface
UML	Unified Modeling Language
UMLS	Unified Medical Language System (https://www.nlm.nih.gov/research/umls/index.html), is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.
WP	Work package

9. Work Packages in CINECA

WP1 - Federated Data Discovery and Querying

WP2 - Interoperable Authentication and Authorisation Infrastructure

WP3 - Cohort Level Meta Data Representation

WP4 - Federated Joint Cohort Analysis

WP5 - Healthcare Interoperability and Clinical Applications

WP6 - Outreach, training and dissemination

WP7 - Ethical and legal governance framework for transnational data-sharing

WP8 - Project Management and coordination

WP9 - Ethics requirements

10. Appendices

None.