Information system for determining expression of genes

(Project for Systems III)

Goran Delić

Definition of the problem

While the human genome has been sequenced, the functional understanding of many genes, especially how and when they are expressed is still limited. Genetic predispositions for various diseases often remain undetected until symptoms arise, and conversely, beneficial genetic traits such as enhanced athletic ability or cognitive potential are rarely identified in advance. This creates a gap in personalized medicine, where individuals and healthcare providers lack access to clear, interpretable information about how a person's unique genetic makeup might a ect their health, abilities, or future risks.

This problem is further complicated by the complexity of gene-environment interactions, epigenetic factors, and the role of ancestry in genetic variability. Current tools often analyze either health history or genetic ancestry separately, failing to synthesize both into meaningful, actionable insights.

To address this issue, I propose to develop an information system called GeneDetective. This system will combine basic medical data with genetic ancestry information to analyze the expression of specific genes, identifying both harmful mutations that could lead to illness and beneficial traits that suggest certain predispositions. By o ering individuals and healthcare professionals a more precise look at gene expression patterns, GeneDetective aims to reduce the uncertainty surrounding genomic data and enable earlier, more personalized interventions.

Functional and nonfunctional requirements of a new system

Functional requirements:

1. User Registration & Authentication
    - Users (patients, doctors, researchers) must be able to register and log in securely.
    - Di erent access levels for roles (patient, geneticist, administrator).
2. Medical Data Input
    - Patients can input personal and medical history data (age, sex, chronic conditions, lifestyle).
3. Genetic Ancestry Input/Integration
    - Users can upload raw genetic data files or manually input known ancestry details.
4. Gene Expression Analysis
    - The system analyzes available data to identify expressed genes associated with diseases or beneficial traits.
    - Includes identification of “at-risk” genes and “advantageous” genes.
5. Reporting Module
    - Generate personalized reports for users, highlighting potential health risks and trait predispositions.
    - Allow exporting to PDF or viewing interactively.
6. Visualization Tools
    - Display gene expression visually (charts, color-coded maps of chromosomes, risk heatmaps, etc.).

7. Recommendation Engine
    - Suggest further testing, preventive care steps, or lifestyle advice based on gene analysis.
8. Data Storage and History
    - Store user data securely for future sessions and allow users to view previous reports and updates.
9. Admin Dashboard
    - Admins can manage users, oversee system performance, and audit security logs.
10. Feedback Mechanism
    - Users can report incorrect results or give feedback for improving the system

Non-functional requirements:

1. The system should respond to gene analysis requests within 10 seconds for uploaded raw data up to 10MB.
2. Throughput: The system should support simultaneous analysis for 100 users without degradation in speed.
3. Genetic and medical data must be accurate, up to date, and stored securely.
4. Integration with external DNA services (via API) should be supported.
5. The data must be formatted in a user-friendly and medically accurate format.
6. All personal and genetic data must be stored in encrypted format (AES-256).
7. Access control must restrict data visibility: only the user and authorized professionals can access results.
8. Must comply with GDPR and other relevant privacy laws.
9. Daily automated backups and o -site data storage are mandatory.
10. Avoid repeated data processing for the same user uploads - implement caching of analysis results.

11. Reduce unnecessary API calls to external services by storing stable ancestry data locally post-verification.
12. The system must be available 24/7 with 99.9% uptime.
13. There will be 3 user types: patients, professionals (e.g., doctors, geneticists), and administrators.
14. The platform should support web and mobile access.
15. Provide a built-in help system, onboarding guide, and short video tutorials.
16. System should be scalable, allowing easy expansion for new gene databases and features.
17. Provide technical documentation and user manuals for all user types.

Feasibility study

The GeneDetective system is technically feasible with current technologies. Tools and frameworks for web-based platforms, secure databases, and genetic data processing are readily available. Open-source bioinformatics libraries like Biopython or scikit-learn for predictive modeling and APIs such as 23andMe can be integrated into the system. Cloud services - AWS, Google Cloud or Azure o er scalable infrastructure for storing sensitive genetic data securely and handling large-scale data analysis.

Since we would be working with highly private human data, there are strict legal and ethical standards which need to be followed, and this part of the system is the most sensitive one. The handling of personal health and genetic information requires adherence to GDPR and biomedical ethics regulations. This includes obtaining informed consent, securing data, and o ering the right to data erasure. In some cultures, discussions of genetic predispositions or talents may be sensitive. This requires thoughtful user communication and optionality in reporting.

When it comes to the economic feasibility, this kind of system has a strong market potential, as it o ers a powerful tool in a growing market of genetic analysis. The system may reduce long-term healthcare costs by identifying health risks early, which can appeal to insurers or public health programs. Development costs can be kept low initially by targeting MVP features and using open-source or existing tools. Future monetization opportunities include tiered subscriptions or medical partnerships.

From a work organization perspective, GeneDetective fits within interdisciplinary teams (bioinformatics, health tech, IT) and can be managed as a cross-functional project. Socially, it addresses rising interest in health optimization and preventive care. However, care must be taken to avoid genetic determinism or discrimination.

Matrix of user roles

Function	Patient/User (1)	Genetic Counselor (2)	System Admin (3)	Researcher (4)
Register and create a profile	YES	YES	YES	YES
Submit medical data	YES	NO	NO	NO
View personal results and gene expression reports	YES	YES	NO	NO
Interpret and annotate genetic results	NO	YES	NO	NO
Modify or delete user data	YES (own data)	Yes (own data)	Yes (with consent)	Yes (own data)
Manage user accounts and roles	NO	NO	YES	NO
Perform system maintenance and backups	NO	NO	YES	NO
Access anonymized genetic data	NO	YES	NO	YES

Export data or reports	YES (own data)	YES (assigned patient)	NO	YES (anonymized)
View system analytics and usage statistics	NO	NO	YES	NO
Contact genetic counselor	YES	YES (response)	NO	YES

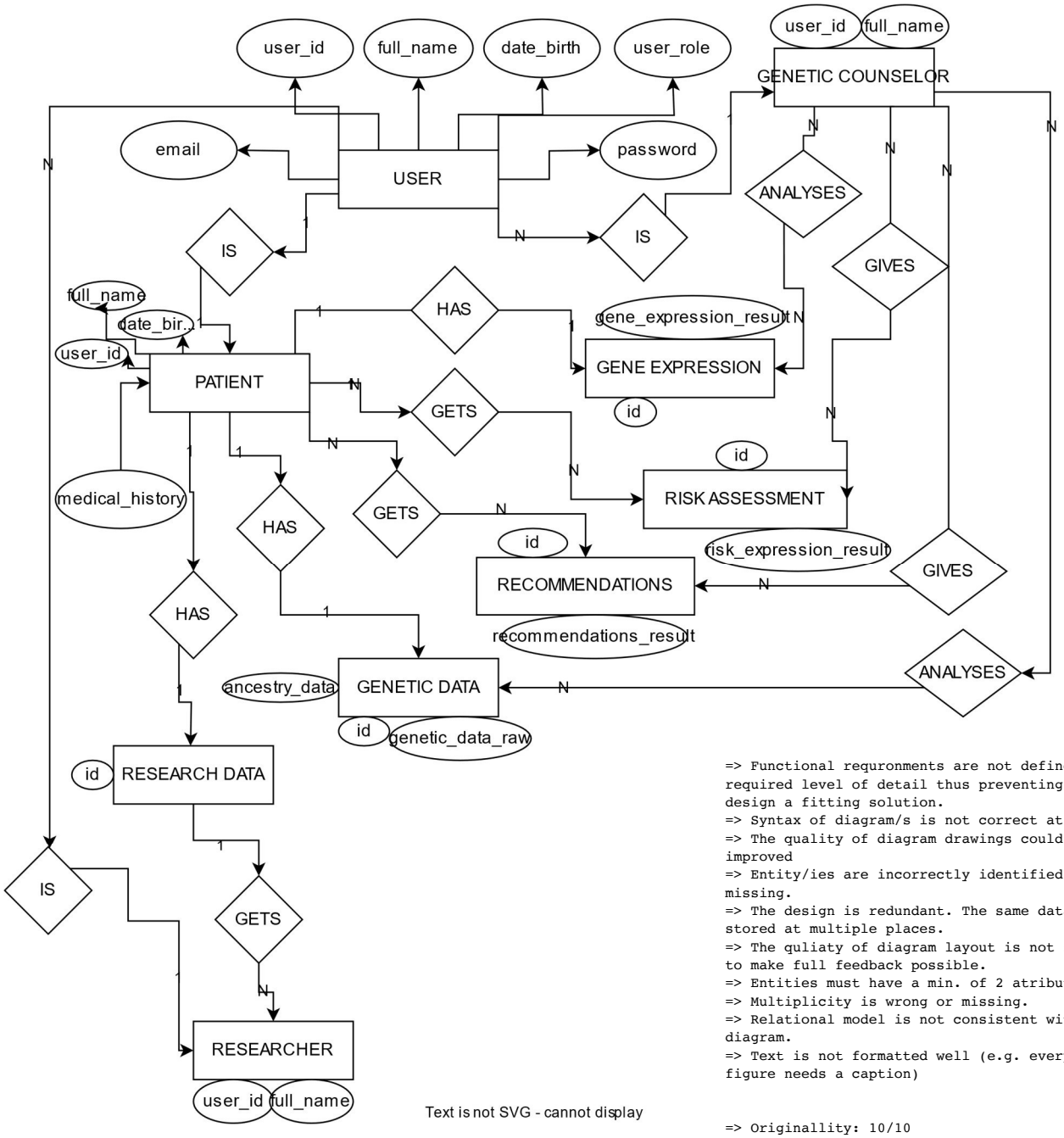
Data dictionary

Entity	Description	Attribute	Type	Description of attribute
User	Patient/User of the system	user_id	Int	Identifier of the user
		full_name	String	User's full name
		date_birth	Date	User's date of birth
		email	String	User's email address
		password	String(hashed)	User's password
		user_role	Int	1,2,3 or 4 – based on the role
Patient	Type of user who is the subject of medical analysis	medical_history	Text	Patient's medical history
		user_id	Int	Identifier for patient
		full_name	String	Full name of the patient
		date_birth	Date	Date of birth of the patient

Genetic counselor	Type of user who is performing the analysis and providing the patient with results	user_id	Int	Identifier for the genetic counselor
		full_name	String	Full name of the genetic counselor
Genetic Data	Information about patient's genetic data	genetic_data_raw	File/JSON	Raw genetic data
		ancestry_data	JSON/Text	Genetic ancestry data
		id	Int	Identifier for the genetic data
Gene expression	Expression data of genes analyzed for the patient	gene_expression_result	JSON/Text	Result of analysis – expressed genes, mutations, markers
		id	Int	Identifier for the gene expression
Risk assessment	Assessment of patient's health based on gene expression	risk_assessment_result	Text	Risk assessment created by the genetic counselor for the patient
		id	Int	Identifier for the risk assessment text
Recommendations	Personalized suggestions based on gene expression	recommendations_result	Text	Personalized health/talent suggestions
		id	Int	Identifier for the recommendations text
Research data	Anonymouse data used for research purposes by researchers	anonymized_id	String	Unique anonymized identifier for research use

Researcher	Receives anonymized patient data for research purposes	user_id	Int	Identifier for the researcher
		full_name	String	Full name of the researcher

ER DIAGRAM

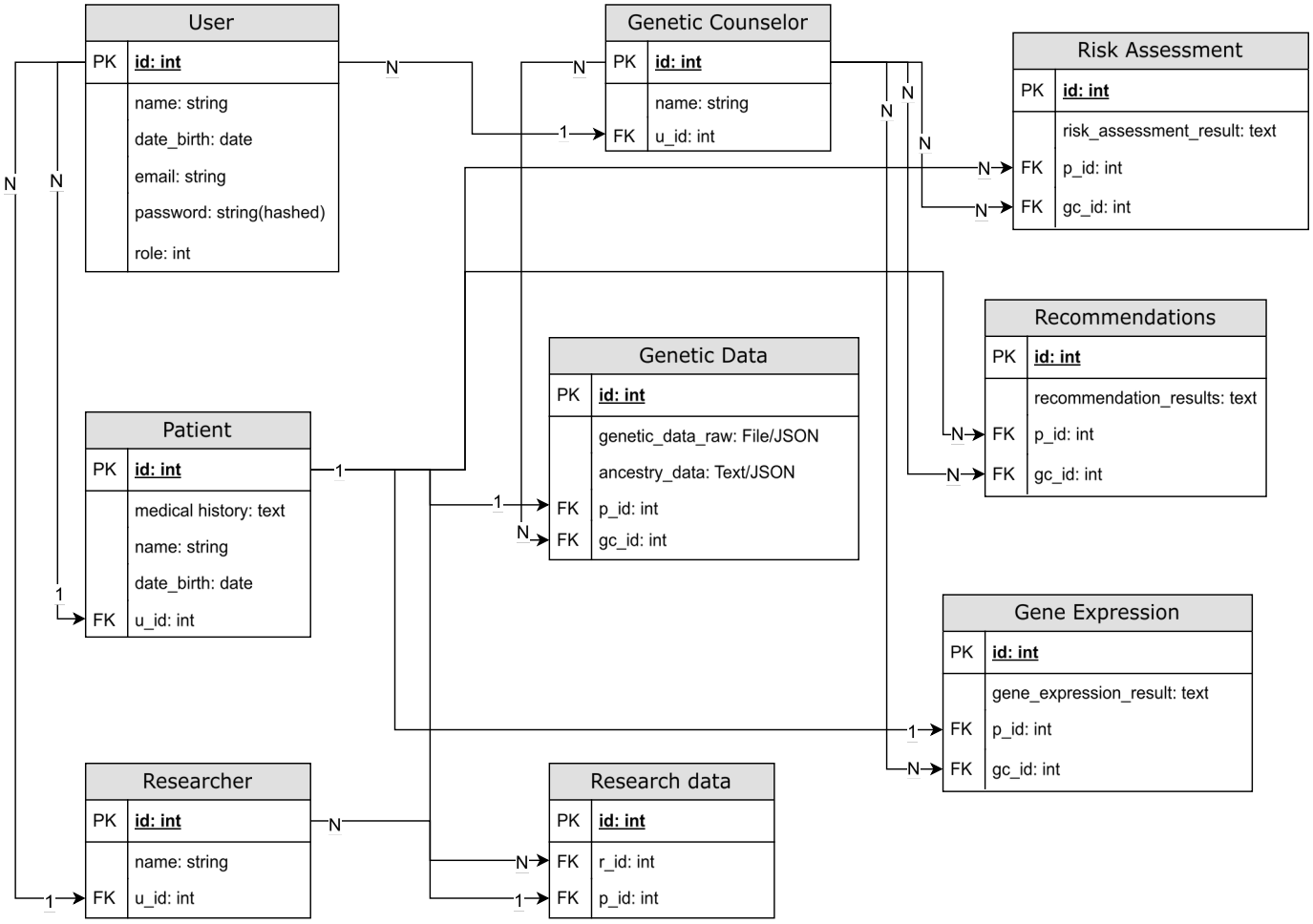


Text is not SVG - cannot display

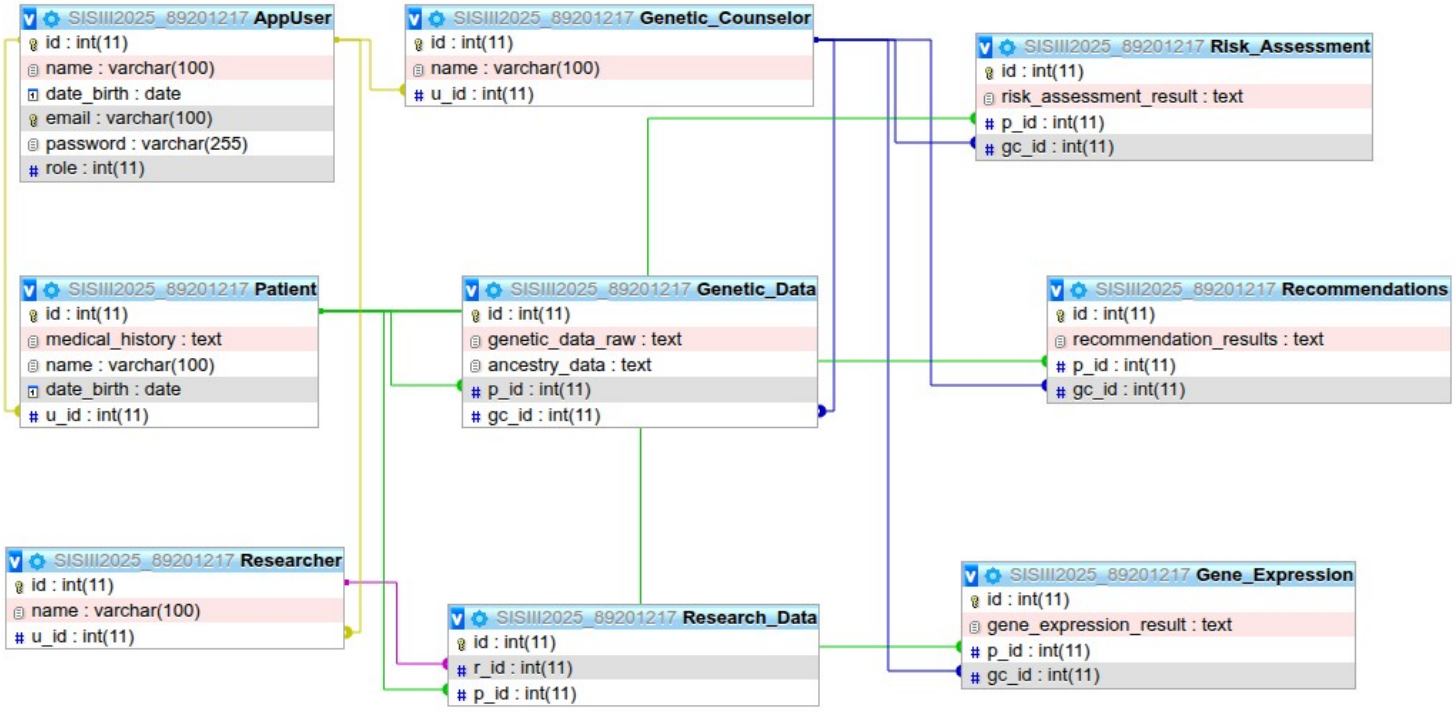
=> Functional requronments are not defined up to required level of detail thus preventing one to design a fitting solution.
=> Syntax of diagram/s is not correct at places
=> The quality of diagram drawings could/should be improved
=> Entity/ies are incorrectly identified or missing.
=> The design is redundant. The same data is stored at multiple places.
=> The quliaty of diagram layout is not sufficeint to make full feedback possible.
=> Entities must have a min. of 2 atributes.
=> Multiplicity is wrong or missing.
=> Relational model is not consistent with the ER diagram.
=> Text is not formatted well (e.g. every table or figure needs a caption)

=> Originallity: 10/10
=> Format and Content Overall: 7.5/15
=> ER diagram and functional requirements: 25/50
=> Relational model and physical database: 0/10
=> UML Class diagram: 5/5
=> UML Sequence diagram: 5/5
=> UML Use-case diagram: 5/5
==========
57.5/100

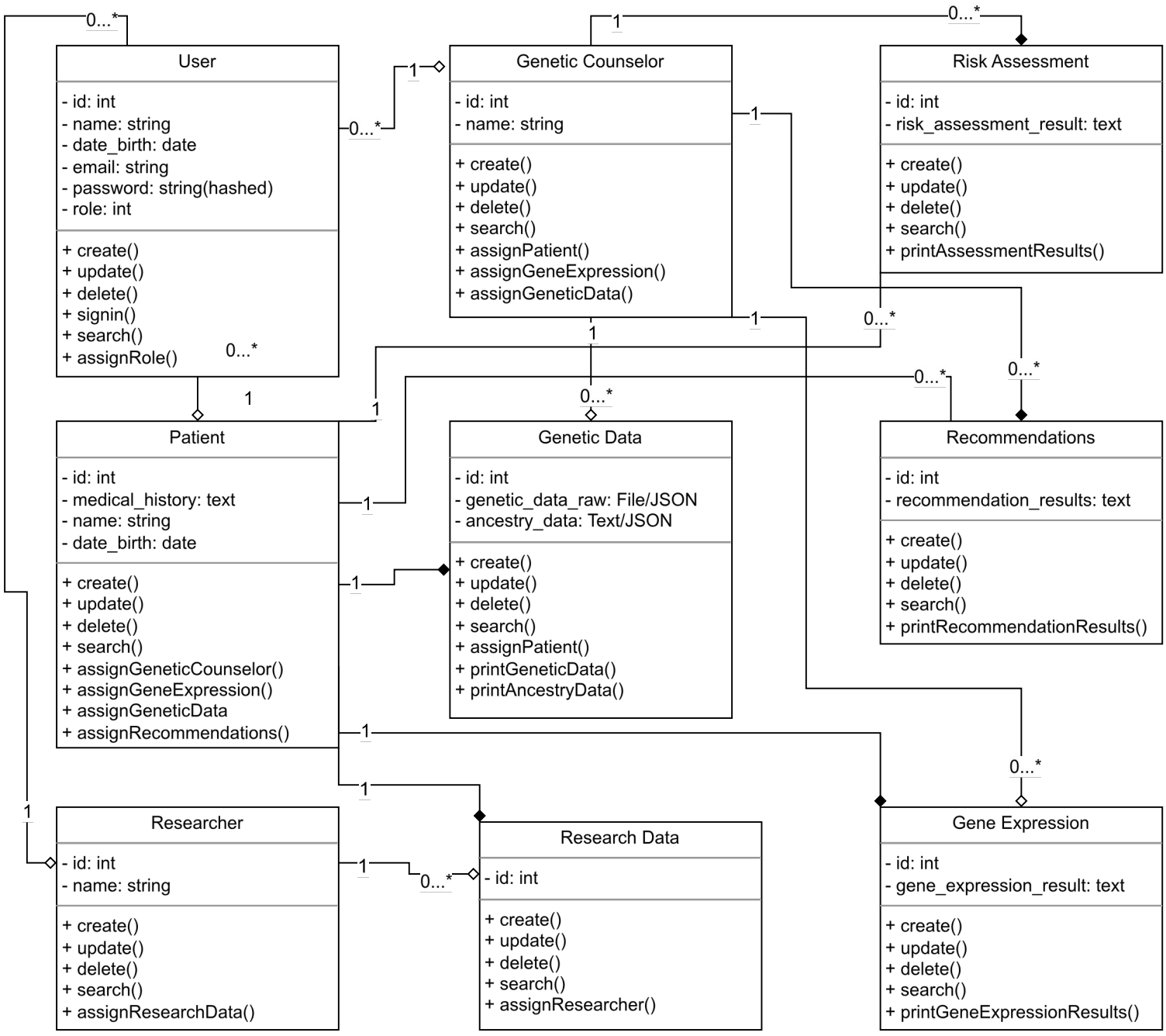
Relational Model



Physical Data Model

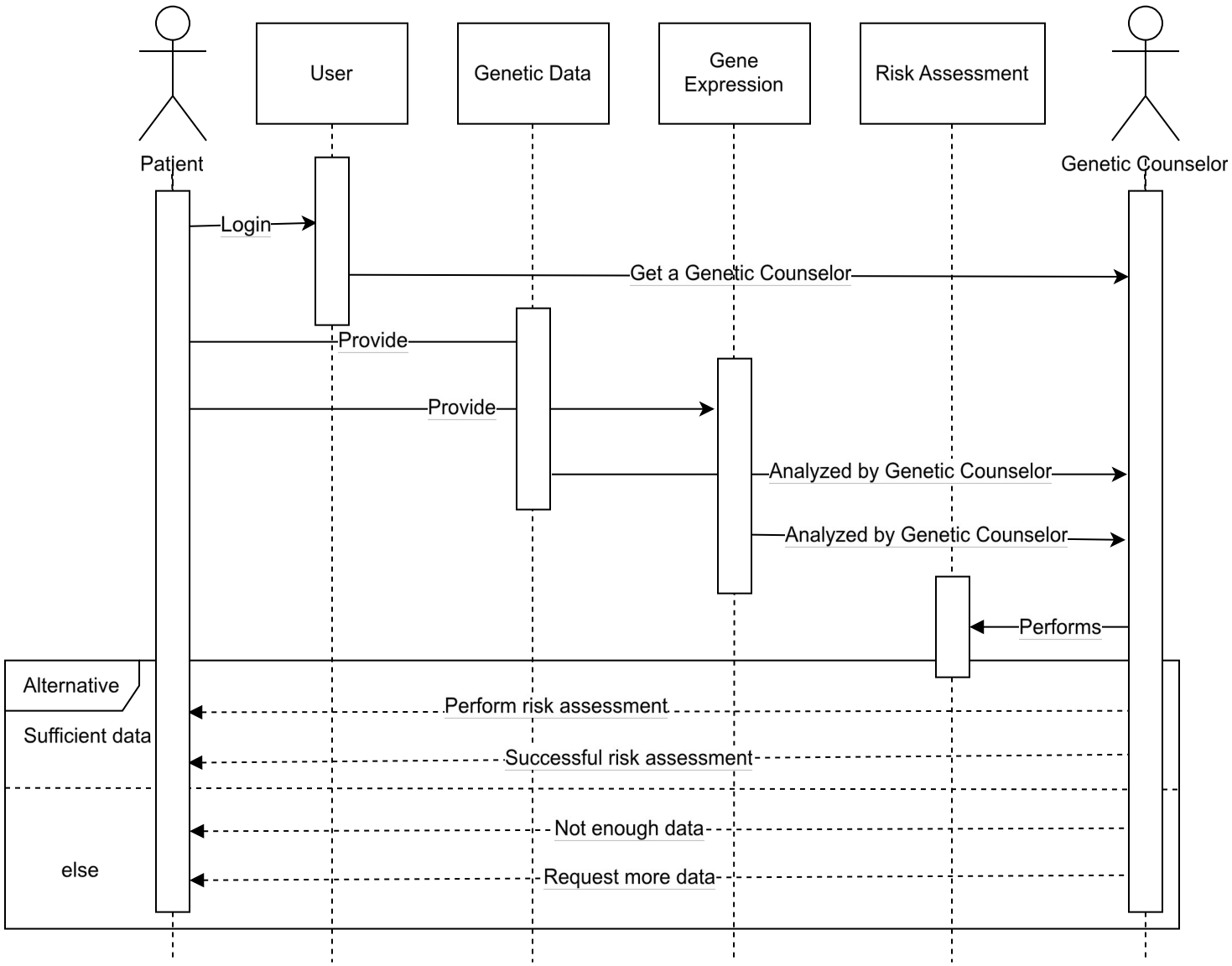


UML class diagram



UML sequence diagram

For a scenario where a patient requests risk assessment



UML use-case diagram