“A genetic disease prioritization platform for gene therapy studies: A wide-ranging review”

Name: Paul Schürmann
Date: 21/06/2023
Student number: 5680263

Supervisor: José Castro Alpizar
Examiner 1: Dr. Sabine Fuchs
Examiner 2: Dr. Peter van Hasselt

Abstract
This review explores the challenge of treating over 10,000 global genetic disorders by targeting their root cause; DNA mutations. Using CRISPR technology as a basis for treating genetic diseases, we review publicly available databases and CRISPR gene editing prediction algorithms for their suitability to screen for diseases with the correct technical, medical and ethical properties that could be considered for gene correction therapy. We discuss the potential for an ethical scoring system to identify disorders most suitable for gene correction therapy. The system, in its conceptual stage and is envisioned to lay the groundwork for a more comprehensive database that could facilitate various applications. These applications could include creating a priority list of diseases for preclinical research, identifying a gene therapy window of opportunity for each disease, proposing a disease list for neonatal screening, and potentially conducting a cost-reduction analysis per disease when treated with gene correction therapy.

Layman’s summary
With over 10,000 genetic diseases worldwide, many remain untreatable or significantly downgrade quality of life. This is largely due to current treatments focusing on disease symptoms, rather than their root cause, DNA errors. Often, these errors involve only a few nucleotide alterations, leading to defective proteins and, in some cases, severe diseases that worsen quality of life.

In 2020, Emmanuelle Charpentier and Jennifer A. Doudna received the Nobel Prize for their discovery of CRISPR, a revolutionary technology capable of locating specific DNA sequences and making precise changes (Zhao et al., 2023). This technology has evolved into CRISPR-prime editing, offering the potential to correct pathogenic mutations in DNA. This advancement could theoretically correct many genetic errors, but it remains unclear which genetic diseases are most suitable to be corrected by this technique.

Several factors determine the potential curability of a disease: technically, the efficiency limitations of DNA correction, delivery of the prime editing machinery to the correct cells in affected organs, and off-target editing risks; medically, the number of cells in an organ that need DNA correction, the reversibility of disease symptoms with treatment, and treatment timing; and ethically, decisions such as when pre-symptomatic treatment should commence and if the disease severity justifies gene editing treatment.

To navigate these complex technical, medical, and ethical considerations, we propose a scoring system to initially screen genetic diseases for potential gene correction therapy treatment. This involves combining various public databases and algorithms. This review presents the concept of our initial version, which could incorporate several databases and demonstrates a hypothetical scoring system. An envisioned version 2 would be more comprehensive and include a scoring system validated by among others fundamental researchers, medical doctors, ethicists, and patients.

Main

Background information

Genetic disorders, which come from small to large changes in DNA, are a big and complex problem for today's healthcare systems. Diseases like Duchenne muscular dystrophy (Duan et al., 2021), cystic fibrosis (Ong & Ramsey, 2023), or sickle cell disease (Kavanagh et al., 2022), despite their origins in minute genetic changes, can lead to severe manifestations, critically impacting individuals' quality of life. The prospect of treating these diseases using gene therapy has stimulated significant interest, leading to a surge of innovative gene therapy tools and the use of public databases that gather vital genetic, phenotypic, and epidemiological data.

One of these gene therapy tools are prime editors. Prime editors are a novel gene correction therapy tool, offering unprecedented precision. Unlike conventional CRISPR genome editing techniques that are dependent on homology-directed repair and often lead to genotoxic double-stranded breaks, prime editing allows for precise, targeted modifications to the genome. These editors can correct a broad range of pathogenic variants, like point mutations, deletions and insertions. The precision and versatility of prime editing offer significant potential for genetic research and therapeutic applications (Zhao et al., 2023).

To understand the potential and limitations of gene correction therapy as a therapeutic application, we must first address a series of technical questions: Can we precisely correct the pathogenic variants for each disease? Can the current tools reach the affected organs and cells, and can these tools correct the genetic defects without causing unacceptable off-target editing? And most important for this research, can we predict the efficiency of these technical processes?

Furthermore, medical considerations are equally critical: Is there a need for new medications given the existing treatment landscape? How confident can we be that the identified gene is the main factor causing the disease? What organs and cells do we need to target, and do these targets align with the available gene correction therapy delivery techniques? Can gene correction therapy reverse the damage already caused by the disease? What is the optimal window of opportunity for treating each disease, and can we run clinical trials in our own medical center?

Lastly, ethical considerations must be considered: Can we develop a standardized scoring system to prioritize diseases for pre-clinical research? Can we compile a list of known genetic diseases suitable for gene correction therapy? Can we determine the window of opportunity for each disease, given that the treatment decisions sometimes need to be made without patient consent due to their young age? Can we expand neonatal screening programs based on the output of our window of opportunity analysis? Can we conduct a cost-benefit analysis of one-time gene correction therapy costs versus the recurrent costs of symptom treatment, such as enzyme replacement therapy?

In the initial version of our research, we explore these questions by reviewing data available in public databases and algorithms (**Figure 1**). We propose a flexible scoring system that can be adapted in collaboration with stakeholders such as researchers, medical doctors, ethicists, and patients. The initial version aims to propose a list of diseases that, based on an adaptable scoring system, may be suitable for gene correction therapy. The approach and sources for the second version are more comprehensive and will be explained in this review, thereby guiding future research and potentially transforming the treatment landscape for genetic disorders.

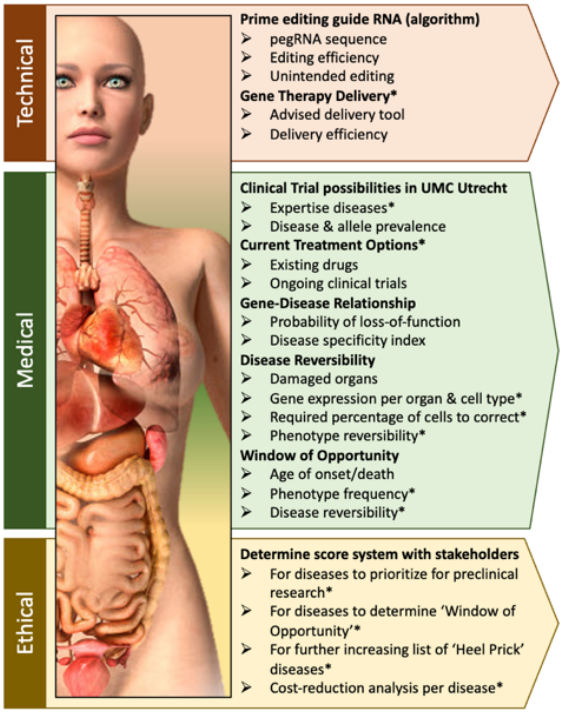

Figure 1: Graphical abstract of proposed technical, medical, and ethical data output. Information without an asterisk * is currently incorporated in the initial version. Information with an asterisk * is not included yet and will potentially be included in version 2.

Publicly available databases and algorithms

Our review incorporates a variety of publicly accessible databases and algorithms, as explained in Table 1. The combined use of these sources supports our primary goal of establishing an understanding of various genetic diseases and their potential for treatment with gene correction therapy.

We collect information from these disparate platforms to develop a scoring system aimed at pinpointing diseases that may be effectively treatable with gene correction therapy. It is important to note that the sources marked with an asterisk (*) are planned for inclusion in the second version of our study and are not part of the initial version. This staged approach allows us to progressively expand and refine our data sources, ensuring progression of our research.

Table 1: Publicly available databases and algorithms used in this research.

Source	Explanation
MONDO	Ontology database that integrates multiple sources to provide a single, up to-date, and coherent disease description. (Vasilevsky et al., n.d.)
OMIM	Online Mendelian Inheritance in Man (OMIM) is an authoritative compendium of human genes and genetic disorders. (Hamosh et al., 2021)
OrphaNet	Information on rare diseases such as disease classifications explaining the type of disease, prevalence and geographic information, average age of onset and death, and inheritance data. (Pavan et al., 2017)
OrphaNet*	Pathogenic Variants Databases, Orphan Drugs, Biobanks Networks, Biobanks, Patients Registries Networks, Patients Registries, Multinational Clinical Trials Networks, National Clinical Trials, Multinational Research Projects Networks, National Research Projects, Patient Organisations Networks, Patient Organisations, Tests Diagnostic & Clinical Laboratories, Expert Centers Networks, Expert Centers, Textual Information datasets.
DisGeNET	A disease-gene platform for discovery of human genes that are associated with diseases and for studying the genetic underpinnings of human diseases. (Piñero et al., 2020)
Ensembl	A resource for reference genomes, genetic variation, gene regulation and functional annotation of genes. (Cunningham et al., 2022)
PRIDICT	A pegRNA algorithm predicting its sequence, on-target, and off-target efficiency based on pathogenic variants as input data. (Kim et al., 2021)
VWS	VWS is the Dutch health ministry that provides the naming of diseases that fall under the heel prick diagnosis. (De Ziekten Die de Hielprik Opspoort, 2022)

HPO*	The Human Phenotype Ontology (HPO) is a standardized vocabulary of phenotypic abnormalities happening in human diseases, providing a platform for computational analysis of genetic and clinical data. (Köhler et al., 2021)
GTEx*	The Genotype-Tissue Expression (GTEx) project is an extensive resource that provides insights into the mechanisms of gene regulation by studying human gene expression and regulation across multiple tissues and individuals. (Lonsdale et al., 2013)

* Will be integrated in version 2

Data extraction and inclusion

Application Programming Interfaces (APIs) are sets of rules and protocols that enable different software applications to communicate and share data. APIs define the methods and data formats that a program can use to communicate with other software, serving as a bridge between different software systems. GitHub, a popular platform for version control and code sharing, provides a robust API allowing users to interact programmatically with its services, including accessing repositories or user profiles. By using these APIs, it is possible to extract data from platforms like OMIM, Orpha, Disgenet, Ensembl, and PRIDICT (a pegRNA prediction algorithm from GitHub), and information from websites like MONDO, DisGeNET, and VWS, enabling you to retrieve specific data sets, automate tasks, and integrate different data sources into your research (**Figure 2**).

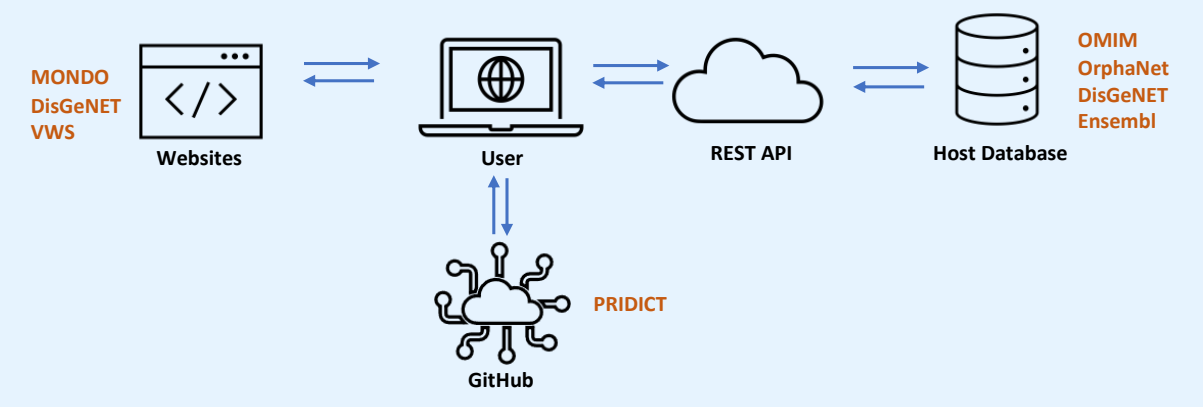

Figure 2: Visualization of data extraction. Databases and algorithms are explained in Table 1.

We collected 9,573 rare diseases from MONDO and OMIM. 7,266 disease numbers correspond to OrphaNet diseases, of which 3,907 are unique. OrphaNet provides epidemiology data generalized per disease rather than disease-subtypes, explaining why duplicate values exist. DisGeNET API links OMIM numbers to 7,326 disease IDs containing the related pathogenic variants. For each variant ID, DisGeNET API provides the chromosome number and coordinate, gene symbol, and allele frequencies. Furthermore, DisGeNETs gene symbol contains more data, such as gene-disease relationships. By combining DisGeNETs pathogenic variant details with Ensembl Rest API sequence, an input sequence for the PRIDICT algorithm was generated. PRIDICT algorithm was modified to process all pathogenic variants to determine pegRNA efficiencies and off-target effects. From the 9,573 diseases, only 2,249 rows (read: diseases) in the database contain information of all the relevant databases (Mondo, OMIM, OrphaNet, DisGeNET, Ensembl, and PRIDICT), resulting in 654 ophthalmic, 116 immunologic, 145 hematologic, and 108 hepatic diseases (**Figure 3**). Nonetheless, all 9,573 diseases were used in the prioritization scheme, since diseases containing less information are ranked lower anyway.

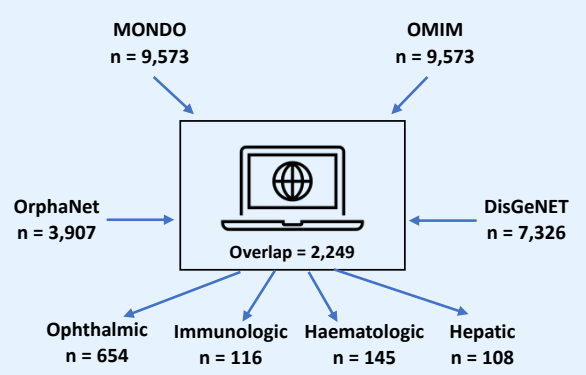

Figure 3: Visualization of data extraction. n is number of input or output diseases. Input databases point towards the box. Output diseases point away from the box (Ophthalmic, Immunologic, Hematologic, Hepatic).

Technical considerations

Technical considerations embody if the prime editor machineries can correct the pathogenic variants in the affected cells and if the machineries can be delivered to the related organs. The initial model should provide the gene editing efficiency and the affected organs. Version 2 should provide in the future, as output, one or more delivery methods that reach the affected organs and cell types, including the predicted versus the required editing and delivery efficiency.

To repair the DNA error, the initial version will employ prime editing, while in future versions other gene editing mechanisms can be included as well. The PRIDICT algorithm is designed to generate prime editor guide RNAs (pegRNAs), predict editing efficiency, and anticipate unintended editing for second-generation prime editors (PE2-NGG). While PE2-NGG is an older prime editor and suboptimal due to the constraints of the protospacer adjacent motif (PAM), newer prime editors like SpRY and SpG, which are nearly PAM-less, have improved flexibility and performance (Liang et al., 2022). However, prediction algorithms are not yet available for these newer versions.

Gene correction therapy delivery to the affected organs is another critical technical aspect but will not be included into the first version. The PRIDICT algorithm predicts the editing efficiency based on a lentiviral delivery vector, but these predictions are based on *in vitro* delivery to the human cell lines, HEK293T and K562, and may not be representative of an *in vivo* setting (Kim et al., 2021). Many different delivery systems exist, including multiple viral vectors, virus-like particles, extracellular vesicles, lipid nanoparticles, and more systems, each with varying delivery efficiencies. Since delivery techniques differ in their efficiency to reach specific areas, new literature studies should be done to incorporate them into the next version of our system (Butt et al., 2022). The number of cells needed to be corrected for phenotype restoration is a topic that will be discussed under medical considerations; however, this information will be excluded in the initial version. In an ideal second version, we would measure the total effectiveness of *in vivo* gene correction therapy by considering factors such as the efficiency of drug delivery and gene editing, the capabilities of drug delivery systems, and the number of cells in each affected organs that need successful gene correction.

While the initial version of our model does not include *in vivo* drug delivery efficiency, we will focus on diseases that affect cells in the bone marrow, liver, and/or eye. This is because these organs are amenable to *ex vivo* gene editing, which is less controversial and potentially safer (Ferrari et al., 2021; Newby et al., 2021; Suh et al., 2022; Zabaleta et al., 2022). For example, hematopoietic stem cells and liver cells can be treated outside the body and then retransplanted, offering a potential advantage. The eye is also a target that is expected to cause fewer systemic effects (Suh et al., 2022). This strategic focus will allow us to make meaningful progress with our research, while laying the groundwork for more complex and comprehensive models in the future.

Technical output

Version 1

- pegRNA sequence (DisGeNET, Ensembl, PRIDICT)
- pegRNA editing efficiency (DisGeNET, Ensembl, PRIDICT)
- Unintended editing (DisGeNET, Ensembl, PRIDICT)
- Affected organs (OrphaNet)

Version 2

- Advised delivery methods (Unknown)
- Delivery efficiency to specific organs and cell types (GTEx, Unknown)
- Required editing efficiency to reverse phenotype (Unknown

Medical considerations

Medical considerations for disease correction primarily focus on the patients benefit for being treated with gene correction therapy and the feasibility of initiating clinical trials in a specific medical center. This necessitates an understanding of several key aspects. These include the current treatment options available, the relationship between the specific gene and the disease, the potential for reversing the disease phenotypes, and the window of opportunity for effective intervention. Also, the likeliness that a gene correction therapy clinical trial can be started in a specific medical center will be investigated. By considering these factors, we can identify the diseases that our medical center, UMC Utrecht, should prioritize for the development of gene therapies.

For identifying those diseases, we aim to determine the likeliness of having potential clinical trial participants. The initial step is to assess the probability of finding patients at the UMC Utrecht. This process in our system involves an examination of medical centers worldwide for their expertise in specific diseases. Since we are still waiting for this data, this will be included in version 2. We will develop a list of diseases where the UMC Utrecht has established expertise. Then, we include disease prevalence data (OrphaNet) with related pathogenic variant frequencies (DisGeNET) to evaluate the likelihood that the UMC Utrecht will have patients of a specific disease with a specific mutation (preferably with a founder mutation or 'homogeneity of mutations') who could participate in clinical trials, assuming that one gene editing machinery will be used per clinical trial (included in version 1). Furthermore, early detection is a beneficial factor in disease prioritization, leading us to prioritize diseases included in the neonatal heel prick screening, which are obtained from the Dutch Health Ministry (VWS) and implemented in version 1. Also, to understand if it is beneficial for patients to join a gene correction therapy clinical trial, we must consider the existing treatments for this disease, however these are excluded in version 1.

Next, we proceeded to evaluate the likeliness that a disease will initiate for a specific pathogenic variant. This is an important factor to determine whether it should be considered to treat a patient pre-symptomatically with gene correction therapy. We investigated the relationships between genes and diseases to ascertain if a specific gene is a valid target for a given disease by using DisGeNET data. This process involved assessing the likelihood that a pathogenic variant in the gene would result in a loss of function, known as the Probability of Loss-of-function Index (PLI). A high PLI score suggests that the gene cannot tolerate loss-of-function mutations without causing disease. However, it is important to note that while PLI gives us insight into the potential for a mutation to cause disease, it does not directly measure the concept of penetrance; the probability that a person carrying a particular genetic variant will develop the disease. Penetrance is influenced by a range of factors, including other genetic variants, and environmental influences. Therefore, we aim to prioritize genes with a high disease specificity index (DSI), since this explains that the pathogenic variant is the primary factor causing the disease and other genetic variants become less important.

Therefore, we assume a high DSI score would indicate that the gene is closely linked to one disease. We included in the initial version that high PLI and high DSI scores may imply a higher likelihood of a disease. Nonetheless, it does not guarantee disease initiation, strengthening the motivation to search for better penetrance models for future versions.

The disease-organ-cell relationship is another critical consideration; therefore, we analyze which organs and cell types are affected by the disease. In the first version, affected organs are derived from diseases listed in Orphanet, which are classified based on a hierarchical disease classification system. In the second version, affected organs and cells will be identified based on gene expression levels sourced from GTEx (or more recent sources), as local production may also contribute to the damage. Also, it will be important information to estimate the minimal number of cells that should be corrected with gene correction therapy to restore the phenotype. Unfortunately, this type of information is not available yet. If this information will be available in the future, we will compare this with the editing capacity as determined in the technical section, to ensure that the required number of cell corrections can be achieved.

Simultaneously, we must ascertain whether the damage is permanent or reversible. Since genetic diseases can cause permanent damage, we strive to identify the frequency of such specific phenotypes and estimate the likelihood of phenotype damage reversibility using a machine learning-based scoring system. Reversibility is not directly available information; the Human Phenotype Ontology database tracks the frequency of certain phenotypes for all diseases using phenotype IDs. In a potential future project, a machine learning model, potentially assisted by an artificial intelligence tool, could be trained by medical doctors or researchers to score phenotypes based on the permanence of the damage. Low scores would indicate irreversible damage, such as bone deformities, and high scores would indicate reversible damage. However, this approach also has limitations, such as reliance on machine learning-determined information that, without verification by doctors or scientists, could be hard to rely on. This methodology could be incorporated in version 2.

Lastly, information on disease severity (age of death) and age of onset can be extracted from OrphaNet. By combining this data with the phenotype reversibility scoring and phenotype frequency, we can estimate the window of opportunity for treating patients with gene correction therapy before irreversible damage begins to accumulate. In summary, these medical considerations form an important part of our research, providing a roadmap for further enhancing our understanding of the complex dynamics between medical considerations and prioritization of diseases for gene therapies.

Medical output

Version 1:

- Medical expertise centers per disease (OrphaNet)
- Disease prevalence (OrphaNet)
- Allele prevalence or homogeneity (DisGeNET and GNOMAD)
- PLI: Probability of Loss-of-function (DisGeNET)
- DSI: Disease Specificity Index (DisGeNET)
- Affected organs – Classifications (OrphaNET)
- Age of onset/death (OrphaNet)

Version 2:

- Affected cell types (GTEx)
- Percentage of cells to target (Unknown)
- Phenotype reversibility (HPO)
- Phenotype frequency (HPO)

Ethical considerations & Score system

The integration of ethical considerations with database information can serve several objectives, one of which is the development of a scoring system to assess which diseases should be prioritized for gene correction therapy research (Table 1). As shown in the table, assigning scores to technical and medical aspects includes an ethical dimension. The creation of ethical guidelines for gene correction therapy should be a collective decision undertaken by a diverse group of stakeholders, including ethicists, patients, researchers, and medical doctors. This collaborative decision-making process ensures that the guidelines embody a range of perspectives and address all potential ethical concerns. Given that the collective decision has not yet taken place, provisional values are currently assigned to the scoring model, allowing the associated graph to be more illustrative (**Figure 4**).

Table 2: Example scoring system for genetic diseases with affected in bone marrow, eye and/or liver.

Technical considerations				
Parameter	**Data (score)**	**Total**	**Status**	**Source**
Gene editing: Efficiency	Between 0 – 100 *	**	Included	PRIDICT
Gene editing: Off-target score	Between 0 – 100 *	**	Included	PRIDICT
Drug Delivery: Efficiency		n/a	Excluded	Various sources
Drug Delivery: Stem cell transplantation related disease (Classifications)	Included in 'stem cell related disorder' classification.	10	Included	OrphaNet
Medical considerations				
Parameter	**Data (score)**	**Max score**	**Status**	**Source**
Clinical Trial: Medical expertise center for the disease	n/a	n/a	Excluded	OrphaNet
Clinical Trial: Prevalence	<1 / 1 000 000 (0.5), 1-9 / 1 000 000 (5), 1-9 / 100 000 (7.5), 1-5 / 10 000 (10), 6-9 / 10 000 (10), >1 / 1 000 (10)	10	Included	OrphaNet
Clinical Trial: Allele Frequency (and 'homogeneity of mutations')	Between 0 – 1 (multiplied by X)	10	Included	DisGeNET
Clinical Trial: Heel Prick disease	Included in heel prick screening	5	Included	VWS
Treatment options: Existing drugs	n/a	n/a	Excluded	OrphaNet
Treatment options: Ongoing clinical trials	n/a	n/a	Excluded	OrphaNet
Gene-Disease: PLI	Between 0 – 1 *	**	Included	DisGeNET
Gene-Disease: DSI	Between 0 – 1 *	**	Included	DisGeNET
Gene-Disease: Type of Inheritance	X-Linked Recessive (?), X-Linked Dominant (?), Autosomal Recessive (5), Autosomal Dominant (0.5)	n/a	Included	OrphaNet
Affected organs: Classifications	If bone marrow, liver, and/or eyes (10). For every other organ -2.5 (per organ).	10	Included	OrphaNet
Affected organs: Gene expression per organ	n/a	n/a	Excluded	GTEx
Affected organs: Number of cells to target	n/a	n/a	Excluded	Unknown
Disease severity: Age of death	Antenatal (2.5), Neonatal (12.5), Infancy (25), Child (22.5), Adolescent (12.5), Adult (2.5), Elderly (0)	25	Included	OrphaNet
Disease severity: Phenotype severity score	n/a	n/a	Excluded	HPO
Disease severity: Phenotype prevalence	n/a	n/a	Excluded	HPO
Disease reversibility: Embryonal damage (Classifications)	n/a	10	Included	OrphaNet
Disease reversibility: Phenotype reversibility	n/a	n/a	Excluded	HPO
Disease reversibility: Age of onset	Antenatal (1), Neonatal (5), Infancy (10), Child (7.5), Adolescent (5), Adult (1), Elderly (0)	10	Included	OrphaNet

* We used a fictive number and combined all the parameters with an asterisk.
** The combined score with all the parameters with a double asterisk is maximum of 20.

By applying this scoring system to our database of 10,000 diseases, we have created a sidebar plot that prioritizes these diseases for gene-editing research (**Figure 4**). One limitation is that only 2,249 out of the 10,000 diseases have complete data, and of these, 1,023 diseases affect the eye, liver, and/or bone marrow. This figure indicates that the top diseases have the highest priority. However, these outcomes are debatable as they are based on provisional scores. Currently, factors such as early age of death, indicative of disease severity, carry significant weight in our model. However, this may not be the most relevant considering gene correction therapy, as disease severity for example does not necessarily correlate with how treatable a disease is with gene correction therapy. Another example; Very long-chain acyl-coa dehydrogenase deficiency (VLCADD) is ranked at the top but has been scored lower because it damages organs other than the liver, eye, and bone marrow, such as the heart and neurons (data not shown). Therefore, it raises the question of whether this disease can even be treated with gene correction therapy if we are unable reach the neurons and heart. Another example is ranked fifth position, namely peroxisome biogenesis disorder, which presents embryonic anomalies that are potentially irreversible. Therefore, it raises the question if this disease deserves the fifth place in the priority list. Importantly, now that the scores are more illustrative, it is crucial to develop a suitable scoring setup with an ethicist and perhaps a mathematician. This setup could be filled out by all stakeholders (ethicists, patients, researchers, and medical doctors) via a survey, ensuring that the most critical diseases appear at the top.

Figure 4: Top 50 prioritized diseases based on the current scoring system. 10,000 diseases are prioritized according to the scorings in Table 1. The top 50 diseases illustrated in this side bar plot contain the highest scores suggesting that they should be most suitable for gene correction therapy treatment (according to the provisional score system).

The timing of disease detection plays a crucial role in making informed treatment decisions. Numerous genetic diseases initiate symptoms early in life and can cause irreversible damage if not treated before this damage starts accumulating. Consequently, early detection methods like the heel prick test, despite their current limitation of screening only 26 conditions, are crucial for preventing this kind of damage. The timing of treatment also presents ethical dilemmas. In certain circumstances, there may be situations where the optimal treatment window coincides with an age when the patient cannot provide informed consent. In these cases, the decision falls on others, such as parents. Besides, sometimes presymptomatic treatment might be ethically justified. Such an approach, however, demands a robust understanding of the person's likelihood to develop the disease. Also, the reversibility of the disease is an important factor. These predictions are challenging; however, we aim to include these in version 2. Thus, a key goal for the second version of our model is to incorporate an understanding of the optimal window for gene correction therapy treatment per disease, as depicted in our proposed figure 5.

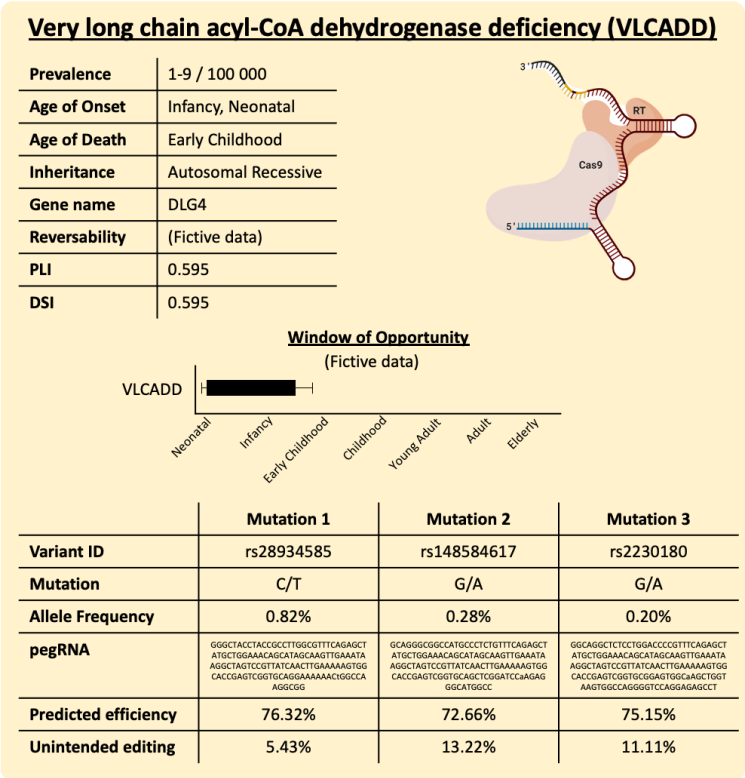

Figure 5: Example figure for data output VLCADD. Optional format for output data per disease containing epidemiological data, window of opportunity data (this is fictive data), and pegRNA data.

Considering the additional capabilities of this database opens a wealth of intriguing possibilities. For instance, neonatal screenings could be enhanced through the inclusion of genetic testing, enabling us to provide a genetic disease list that are treatable at young ages via gene correction therapy. Additionally, the database could serve as a tool in cost-reduction studies, where we could analyze and compare the financial implications of managing genetic diseases with and without gene correction therapy. By providing insights into these areas, the database might hold significant potential to guide both medical practices and healthcare economics.

Discussion

Our comprehensive review of the gene correction therapy landscape has shown several key challenges and considerations that span technical, medical, and ethical dimensions. One of the central challenges in gene correction therapy is the lack of public databases, limiting the range and depth of data available for research. This issue is further complicated when we need to compare thousands of diseases with minimal overlapping characteristics, often leading to reliance on less specific properties such as age of onset, age of death, and prevalence. To overcome these hurdles, we utilized various extra resources like OrphaNet, OMIM, DisGeNET, and MONDO. However, this strategy is not without its complications. For instance, the use of genotype-phenotype interactions poses a risk for errors. Efforts have been made to cross-reference and correlate data, such as Orphanet data, but it is not always the optimal approach, especially for single orpha diseases involving multiple OMIM or DisGeNET diseases.

In terms of technical considerations, the application of PRIDICT, a prime editing algorithm, is crucial for genetic correction. PRIDICT offers insights into the potential of CRISPR prime editing by designing pegRNA sequences and predicting both intended and unintended editing efficiencies. But PRIDICT has its limitations, particularly as it was designed for second-generation prime editors, which carry restrictions related to the protospacer adjacent motif (PAM). This situation underscores the need for its updating ability to incorporate newer pegRNA prediction algorithms. Additionally, drug delivery, a significant aspect of gene correction therapy, is not included in the current version of PRIDICT. Future updates should incorporate this aspect for a more holistic approach to genetic diseases.

In a medical perspective, several questions require attention. Although the included data can provide insights, there is much room for improvement in areas like current treatment options, disease reversibility, the identification of affected organs, estimating the number of cells to target for effectively reversing disease phenotypes, and identifying medical expertise centers to conduct clinical trials for gene correction therapy studies. Details included in our initial version like age of onset, death, anomalies during embryogenesis, and transplant related disorders are helpful, but if there is need for new medications remains unclear to determine the disease severity using our initial version. Furthermore, the current databases, such as DisGeNET, provide insights in gene-disease relationships but are insufficient to address current complex issues such as penetrance, whether gene correction therapy can reverse existing disease damage, or identify the optimal treatment window of opportunity. To accomplish these tasks, proposed methods should be developed as explained earlier.

Ethically, the use of a scoring system assessed by stakeholders should be validated by comparing the top diseases in the results with other genetic diseases that have (nearly) approved gene-editing therapies. These diseases would ideally rank highly, thus lending credibility to the scoring systems.

In conclusion, while our initial version serves as a starting point for prioritizing gene correction therapy, there remains a clear necessity for further refinement. We recognize that multiple disease parameters are currently partly data deficient due to a lack of information from publicly available databases. Despite these challenges, the applications of current and future versions carry the potential that may revolutionize the gene correction therapy landscape. As gene editing techniques evolve, our model will also advance, offering a fundamental tool to simplify and address the wide array of upcoming technical, medical, and ethical questions related to treating gene therapies.

References

Butt, M. H., Zaman, M., Ahmad, A., Khan, R., Mallhi, T. H., Hasan, M. M., Khan, Y. H., Hafeez, S., Massoud, E. E. S., Rahman, M. H., & Cavalu, S. (2022). Appraisal for the Potential of Viral and Nonviral Vectors in Gene Therapy: A Review. In *Genes* (Vol. 13, Issue 8). MDPI. https://doi.org/10.3390/genes13081370

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

De ziekten die de hielprik opspoort. (2022, June 1). N/a.

Duan, D., Goemans, N., Takeda, S., Mercuri, E., & Aartsma-Rus, A. (2021). Duchenne muscular dystrophy. In *Nature Reviews Disease Primers* (Vol. 7, Issue 1). Nature Research. https://doi.org/10.1038/s41572-021-00248-3

Ferrari, S., Vavassori, V., Canarutto, D., Jacob, A., Castiello, M. C., Javed, A. O., & Genovese, P. (2021). Gene Editing of Hematopoietic Stem Cells: Hopes and Hurdles Toward Clinical Translation. *Frontiers in Genome Editing*, *3*. https://doi.org/10.3389/fgeed.2021.618378

Hamosh, A., Amberger, J. S., Bocchini, C., Scott, A. F., & Rasmussen, S. A. (2021). Online Mendelian Inheritance in Man (OMIM®): Victor McKusick's magnum opus. *American Journal of Medical Genetics, Part A*, *185*(11), 3259–3265. https://doi.org/10.1002/ajmg.a.62407

Kavanagh, P. L., Fasipe, T. A., & Wun, T. (2022). Sickle Cell Disease: A Review. In *JAMA* (Vol. 328, Issue 1, pp. 57–68). American Medical Association. https://doi.org/10.1001/jama.2022.10233

Kim, H. K., Yu, G., Park, J., Min, S., Lee, S., Yoon, S., & Kim, H. H. (2021). Predicting the efficiency of prime editing guide RNAs in human cells. *Nature Biotechnology*, *39*(2), 198–206. https://doi.org/10.1038/s41587-020-0677-y

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The human phenotype ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217. https://doi.org/10.1093/nar/gkaa1043

Liang, F., Zhang, Y., Li, L., Yang, Y., Fei, J. F., Liu, Y., & Qin, W. (2022). SpG and SpRY variants expand the CRISPR toolbox for genome editing in zebrafish. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-31034-8

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics* (Vol. 45, Issue 6, pp. 580–585). https://doi.org/10.1038/ng.2653

Newby, G. A., Yen, J. S., Woodard, K. J., Mayuranathan, T., Lazzarotto, C. R., Li, Y., Sheppard-Tillman, H., Porter, S. N., Yao, Y., Mayberry, K., Everette, K. A., Jang, Y., Podracky, C. J., Thaman, E., Lechauve, C., Sharma, A., Henderson, J. M., Richter, M. F., Zhao, K. T., … Liu, D. R. (2021). Base editing of haematopoietic stem cells rescues sickle cell disease in mice. *Nature*, *595*(7866), 295–302. https://doi.org/10.1038/s41586-021-03609-w

Ong, T., & Ramsey, B. W. (2023). Cystic Fibrosis: A Review. In *JAMA* (Vol. 329, Issue 21, pp. 1859–1871). American Medical Association. https://doi.org/10.1001/jama.2023.8120

Pavan, S., Rommel, K., Marquina, M. E. M., Höhn, S., Lanneau, V., & Rath, A. (2017). Clinical practice guidelines for rare diseases: The orphanet database. *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0170365

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, *48*(D1), D845–D855. https://doi.org/10.1093/nar/gkz1021

Suh, S., Choi, E. H., Raguram, A., Liu, D. R., & Palczewski, K. (2022). Precision genome editing in the eye. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(39). https://doi.org/10.1073/pnas.2210104119

Vasilevsky, N. A., Roncaglia, P., & Ross, J. E. (n.d.). *Mondo: Unifying diseases for the world, by the world*. https://doi.org/10.1101/2022.04.13.22273750

Zabaleta, N., Torella, L., Weber, N. D., & Gonzalez-Aseguinolaza, G. (2022). mRNA and gene editing: Late breaking therapies in liver diseases. In *Hepatology* (Vol. 76, Issue 3, pp. 869–887). John Wiley and Sons Inc. https://doi.org/10.1002/hep.32441

Zhao, Z., Shang, P., Mohanraju, P., & Geijsen, N. (2023). Prime editing: advances and therapeutic applications. In *Trends in Biotechnology*. Elsevier Ltd. https://doi.org/10.1016/j.tibtech.2023.03.004