TABLE OF CONTENTS

Bibliometrics: A lifelong ascent or the Caudine Forks for the modern researcher? **1**
A. Pession

Enzyme replacement therapy for lysosomal storage diseases: learning from the past to shape the future **5**
S. Fecarotta, A. Tarallo, G. Parenti

Communication challenges in newborn screening: what the public knows (and doesn't) **22**
M. Costanzo, S. Bianco, M. Caterino, A. Cevenini, M. Cozzolino, C. Cazzorla, M. Bani, M. Ruoppolo

How reliable are large language models in metabolic dietetics? a cross-sectional benchmarking evaluation **31**
M. Tosi, M. Muscarella, G. Gugelmo, A. Pinto, M. Suárez González, S. Poloni, A. Höller, V. Perico, A. Favaro, A. Tavian, A. Belvedere, A. Dianin

Mind the bias: mapping and mitigating cognitive errors across the expanded newborn screening pathway **46**
S. Russo, S. Gasperini, V. Crescitelli, M.C. Faraguna, R. Pretese, C. Bonfanti, S. Gigante, K. Pozzi, A. Cerizza, G. La Marca, M. Bani

JIM Journal of Innate Metabolism
www.jim.simmesn.org

Copyright © 2026 Journal of Innate Metabolism
Registered at No. 8618 on 02/16/2024 in the periodical press register of Bologna
Publisher: Verduci Editore s.r.l. via Gregorio VII, 186 - 00165 Rome (Italy) - P.I. 03761621006
Printed: August 2026 - Print on Web Srl, Isola del Liri (FR)

QUARTERLY SCIENTIFIC JOURNAL FOR HEALTHCARE PROFESSIONALS
Volume 3 - Issue 3 - August 2026

ISSN: 3034-9346 — e-ISSN: 3034-9362.

JIM
JIM 2026; 3 (3): e1195
DOI: 10.61012/JIM_202608_1195
Category: *Editorial*

BIBLIOMETRICS: A LIFELONG ASCENT OR THE CAUDINE FORKS FOR THE MODERN RESEARCHER?

A. PESSION

• • •

Full Professor of Pediatric Alma Mater Studiorum, University of Bologna, Bologna, Italy

Corresponding Author
Andrea Pession, Ph.D; e-mail: andrea.pession@unibo.it

Keywords: Bibliometrics, Research assessment, h-index, Journal impact factor, CiteScore, Citation indicators, Responsible metrics, Research integrity.

"We are scaling the mountain of metrics, only to realize we have lost sight of the horizon of discovery."

Contemporary academic aphorism

The landscape of contemporary scientific evaluation has undergone a profound transformation, evolving from a qualitative realm of peer appreciation into a highly quantified, algorithmic matrix. Today, the career trajectories of researchers, the allocation of institutional funding, and the perceived prestige of scientific output are inextricably bound to quantitative indicators. As an international community, we must critically evaluate whether this pervasive reliance on bibliometric indices represents a triumphant meritocratic journey – the metaphorical climb of a lifetime, where each peer-reviewed publication marks a grueling step toward the academic summit – or whether it has devolved into the Caudine Forks, forcing researchers to pass through a humiliating, bureaucratic yoke that deforms the very nature of scientific inquiry. From one perspective, the rise of bibliometrics has been heralded as a major democratic advancement, mapping the steep terrain of a lifelong academic ascent by replacing subjective favoritism with objective criteria. To navigate this terrain, the scientific community developed specific indicators, each designed to capture a different dimension of impact.

The **Journal Impact Factor** (JIF), originally conceived by Eugene Garfield in the mid-twentieth century, emerged as the foundational metric for measuring journal prestige[1]. Calculated annually by Clarivate Analytics within the Journal Citation Reports, the JIF for a specific year *X* is mathematically defined as the ratio between the citations received in that year by items published in the two preceding years, and the total number of citable items published during those same 2 years:

$$\mathrm{JIF}_X = \frac{\text{Citations received in year } X \text{ by items published in years } (X-1) \text{ and } (X-2)}{\text{Total number of citable items published in years } (X-1) \text{ and } (X-2)}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The primary strength of this formula lies in its ability to evaluate the average citation momentum of a journal over a tight, 2-year window, offering a historical benchmark that institutions have trusted for decades. However, its major weakness is that it measures the vehicle, not the passenger; a high JIF does not guarantee the quality of an individual paper. Furthermore, while the numerator encompasses citations to any document type, the denominator is strictly restricted to substantive research contributions, specifically original research articles and review papers, leaving the index vulnerable to artificial inflation through uncounted editorial front matter.

To address these structural imbalances, Elsevier's Scopus introduced **CiteScore**, a metric that expands the evaluation window and applies a transparent, fully inclusive accounting policy[2]. The mathematical formula for calculating the CiteScore for a specific year *X* utilizes a 4-year rolling window and is expressed as follows:

$$\text{CiteScore}_X = \frac{\text{Citations received over 4 years } (X-3 \text{ to } X) \text{ from documents published over 4 years } (X-3 \text{ to } X)}{\text{Total number of documents published over 4 years } (X-3 \text{ to } X)}$$

This equation introduces critical methodological deviations from traditional JIF. By extending the temporal window to 4 years and including the current year of calculation, CiteScore reduces the volatility of short-term citation spikes and offers a stabilized reflection of a journal's citation longevity. Crucially, its denominator enforces a strict "no-exemptions" policy, incorporating every single published document type, including editorials, letters, notes, and conference papers, thereby preventing journals from manipulating their scores through selective front-matter publication.

While CiteScore offers a more transparent and comprehensive overview, its all-inclusive nature is occasionally criticized by traditionalists for lacking the strict selectivity associated with highly curated databases. A direct comparative evaluation of these two indicators reveals a fundamental philosophical and operational divide in journal matrix tracking. While the JIF prioritizes immediate citation velocity, acting as an acute indicator of short-term scientific interest, CiteScore captures long-term citation persistence, which is particularly beneficial for disciplines characterized by slower citation half-lives. Furthermore, their contrasting mathematical treatments of the denominator radically alter editorial behavior. Under the JIF framework, editors are structurally incentivized to publish high volumes of uncounted "front matter", such as extensive editorials or letters to the editor, that actively cite the journal's own research papers, thereby artificially inflating the numerator without penalizing the denominator. CiteScore completely neutralizes this loophole through its "no-exemptions" policy; because every written item counts toward the denominator, any expansion of front matter automatically dilutes the overall score unless those items receive a proportionally high volume of external citations. Consequently, the JIF rewards selective, high-velocity publishing models, whereas CiteScore provides a more egalitarian, transparent, and structurally stable representation of a journal's total editorial output across a broader multidisciplinary indexing database[2].

On an individual level, the **h-index**, introduced by Jorge Hirsch, revolutionized author evaluation by mathematically fusing productivity and citation impact into a single, elegant number[3]. Formally, a researcher has an index *h* if *h* of their total N_p papers received at least *h* citations each, while the remaining N_p - *h* papers have received fewer than *h* citations each. Mathematically, if an author's publications are ranked in descending order based on their citation counts (C_i), the h-index is defined as:

$$h = \max\{i \in \mathrm{N} : C_i \geq i\}$$

The undeniable strength of this formulation is its absolute robustness against both the "one-hit wonder" phenomenon and the inflation of unreferenced volume; it ensures that an investigator cannot achieve elite status based on a single highly cited paper or through sheer, low-quality prolificacy. Nevertheless, its structural weaknesses are profound: it fails to account for the career stage of the researcher, as it can only increase over time, inherently penalizing early-career investigators. Furthermore, it completely ignores author placement, rewarding a middle co-author on a massive multicenter trial equally to a first or corresponding author who drove the entire project[3].

To resolve these cross-disciplinary and career-stage discrepancies, modern bibliometrists rely on advanced normalized indices, such as **Elsevier's Field-Weighted Citation Impact (FWCI)** and **Clarivate's Category Normalized Citation Impact (CNCI)**[4,5]. For a specific publication *k*, the normalized citation impact is mathematically formulated as the ratio of the actual number of citations received by that paper (c_k) to the expected baseline average number of citations (e_k) received by all global publications of the same document type, published in the same year, and indexed within the same subject category or discipline:

$$\text{Normalized Citation Impact}_K = \frac{c_K}{e_K}$$

When evaluating an entire publication portfolio of an author consisting of *N* papers, the FWCI or CNCI represents the mean of these individual ratios. For a structural portfolio where the global baseline expectations vary across multiple co-existing disciplines, the institutional definition is formulated as:

$$\text{Portfolio FWCI / CNCI} = \frac{1}{N}\sum_{K=1}^{N}\frac{c_K}{e_K}$$

A resulting value of exactly 1.0 indicates that the researcher's output matches the global average. Values greater than 1.0 (e.g., 1.45) signify that the portfolio's impact is 45% higher than the global baseline for that exact field and year.

The undeniable strength of these normalized metrics is their mathematical ability to eliminate discipline-specific structural biases, allowing an equitable comparison between researchers across completely distinct clinical fields; their weakness, however, rests in their statistical complexity, making them less intuitive and harder to implement for traditional university tenure committees. To appreciate the necessity of these normalized parameters, one must contrast them directly with raw, unnormalized measurements like a foundational Citation Index – originally engineered by Eugene Garfield to capture absolute citation frequencies. Garfield's core Citation Index tabulates absolute counts without mathematical filtering, meaning every single reference carries equal weight regardless of structural context. Consequently, a traditional Citation Index functions as a raw, field-dependent accumulation ledger, making it highly sensitive to the massive volume variations inherent to different scientific cultures. In contrast, modern normalized metrics, such as the FWCI and CNCI, decouple citation performance from disciplinary boundaries. The mathematical architecture of both FWCI and CNCI dynamically establishes a shifting expected baseline average based on publication year, document type, and subject classification. While a paper with ten citations in molecular biology might look identically positioned to a paper with 10 citations in the humanities within a standard Citation Index, the normalized framework immediately exposes the hidden truth: the former sits well below its field's expected average, while the latter dramatically outperforms its disciplinary baseline. By computing these contextual ratios, FWCI and CNCI transform raw data into equitable cross-disciplinary impact vectors, although they sacrifice the immediate, transparent arithmetic simplicity of Garfield's classical model.

When these measures are used uncritically, a darker pathophysiological process initiates within the academic body corporate, exposing the severe structural damage caused by a "**publish or perish**" culture. The modern researcher, midway through this lifelong ascent, is increasingly forced to pass under the Caudine Forks of algorithmic dictatorship, where these indices are manipulated through self-citation loops, reciprocal citation cartels, and the artificial slicing of robust studies into multiple minimum publishable units.

In this landscape of perpetual quantitative pursuit, the doors to a dangerous academic trap have swung wide open: the phenomenon of **predatory publishing**. Where traditional channels impose biblical review times and staggering rejection rates, these predatory outlets offer an illusory escape route. Behind the promise of rapid, effortless publication lies a purely lucrative business, devoid of any genuine peer review. For the desperate researcher, crushed by the mandate to inflate their H-index, predatory publishing becomes a modern siren song, corrupting the foundational principles of the open access movement[6]. To navigate these murky waters, the scientific community long sought a guiding light. The most famous attempt was Beall's List, the pioneering catalog curated by librarian Jeffrey Beall to unmask deceptive publishers[7]. However, the fate of that list underscores the complexity of the crisis: shuttered in 2017 under the weight of intense legal pressure and controversy, its disappearance left a void that contemporary databases (such as Cabells) and institutional whitelists struggle to fill, forcing researchers

to rely on archived open-source versions of the directory[8]. Consequently, for the modern researcher stripped of a universal compass, vigilance becomes yet another bureaucratic burden. The risk is dramatic: paying to step into an ambush that not only generates zero scientific value but leaves an indelible stain on one's professional reputation, turning the long-sought lifelong ascent into academic suicide.

This systemic tension necessitates a definitive paradigm shift. Scientific publishing must transcend the reductionist view of treating papers as mere currency to inflate an individual's metrics. We must actively support international frameworks and active projects that champion responsible research assessment. Chief among these is the **San Francisco Declaration on Research Assessment (DORA)**, which explicitly denounces the misuse of journal-based metrics as a surrogate for individual paper quality[9,10].

Furthermore, this vision is systematically operationalized by the **Coalition for Advancing Research Assessment (CoARA)**. Rather than offering mere abstract declarations, CoARA binds its signatories to ten core commitments designed to establish a systemic overhaul of evaluation practices[11]. Central to this framework is the mandate to recognize the diverse outputs, practices, and activities that maximize research quality, moving academic validation far beyond traditional journal articles. Signatories explicitly pledge to base research assessment primarily on qualitative judgment, reinstating peer review as the bedrock of evaluation while ensuring that quantitative indicators play a strictly supportive, responsible role. Crucially, the agreement demands that institutions abandon the inappropriate uses of journal- and publication-based metrics – specifically targeting the inappropriate application of the JIF and h-index as shorthand markers for individual quality. It also requires organizations to avoid using rankings of research organizations as a primary metric in the evaluation of individual researchers. To bridge the gap between policy and practice, CoARA enforces tangible operational accountability. Every member organization is required to commit resources to advance assessment reform, share lessons learned through an international network of knowledge exchange, and submit a comprehensive Action Plan within one year of signing the agreement. By participating in specialized Working Groups and National Chapters, institutions actively test and pilot new, inclusive assessment tools. Editorial boards, funding bodies, and university committees must lead this transition, shifting the evaluative framework away from a reactive compliance with algorithmic metrics toward a proactive defense of scientific integrity, ensuring that research evaluation serves as a strategic cornerstone for progress, transforming the modern researcher's path back into a noble, rewarding, and lifelong ascent.

Conflict of Interest:
The author declares no conflict of interest.

Ethics Approval and Informed Consent:
Not required due to the nature of the article.

ORCID ID:
Andrea Pession: 0000-0002-0379-9562

REFERENCES

1. Garfield E. Citation indexes for science: A new dimension in documentation through association of ideas. Science 1955; 122: 108-111.
2. James C, Colledge L, Meester W, Azoulay N, Plume A. CiteScore metrics: Creating journal metrics from the Scopus citation index. Learned Publishing 2019; 32: 367-374.
3. Hirsch JE. An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 2005; 102: 16569-16572.
4. Elsevier. Research Metrics Guidebook: Field-Weighted Citation Impact (FWCI) [Internet]. Amsterdam: Elsevier; 2022. Available at: https://supportcontent.elsevier.com/RightNow%20Next%20Gen/SciVal/ACAD_RL_ElsevierResearchMetricsBook_WEB.pdf (accessed 2 June 2026).
5. Clarivate. InCites Benchmarking & Analytics Help: Category Normalized Citation Impact (CNCI) [Internet]. Philadelphia: Clarivate; 2025. Available at: https://incites.zendesk.com/hc/en-gb/articles/25087312115601-Category-Normalized-Citation-Impact-CNCI (accessed 2 June 2026).
6. Beall J. Predatory publishers are corrupting open access. Nature 2012; 489: 179.
7. Beall J. What I learned from predatory publishers. Biochem Med (Zagreb) 2017; 27: 273-279.
8. Beall's List of Potential Predatory Journals and Publishers [Internet]. [Post-2017/Archived version]. Available at: https://beallslist.net/ (accessed 2 June 2026).
9. San Francisco Declaration on Research Assessment (DORA). San Francisco, CA: DORA; 2013. Available at: https://sfdora.org (accessed 2 June 2026).
10. Cagan R. San Francisco Declaration on Research Assessment. Dis Models Mech 2013; 6: 869-870.
11. Coalition for Advancing Research Assessment (CoARA). Agreement on Reforming Research Assessment. Brussels: CoARA; 2022. Available at: https://coara.eu (accessed 1 June 2026).

JIM
JIM 2026; 3 (3): e1207
DOI: 10.61012/JIM_202608_1207
Category: *Review*

ENZYME REPLACEMENT THERAPY FOR LYSOSOMAL STORAGE DISEASES: LEARNING FROM THE PAST TO SHAPE THE FUTURE

S. FECAROTTA[1], A. TARALLO[1,2], G. PARENTI[1,2]

• • •

[1]Department of Translational Medical Sciences, Federico II University, Naples, Italy
[2]Telethon Institute of Genetics and Medicine, Pozzuoli, Italy

Corresponding Author
Giancarlo Parenti, MD; e-mail: parenti@unina.it

ABSTRACT – Enzyme replacement therapy (ERT) has represented a major therapeutic breakthrough for lysosomal storage diseases (LSDs) over the past three decades, translating fundamental insights into lysosomal enzyme trafficking and cross-correction into effective clinical interventions. Since its first application in Gaucher disease, ERT has been extended to multiple LSDs, leading to substantial improvements in survival, organ function, and quality of life. However, accumulated long-term clinical experience has revealed several important limitations that preclude ERT from being universally effective.

This review discusses key factors influencing the efficacy of ERT, including enzyme biodistribution, receptor-mediated uptake, pharmacokinetics, and the impact of physiological barriers such as the blood–brain barrier. Tissue-specific variability in response, immune reactions, and the influence of disease stage and secondary pathogenic mechanisms are examined as major determinants of therapeutic outcome. In addition, challenges related to manufacturing, cost, and treatment burden are considered.

Advances in biotechnology have led to the development of second- and third-generation enzymes with improved targeting properties, reduced immunogenicity, and enhanced stability. Novel approaches, including glycoengineering, chimeric enzymes capable of crossing biological barriers, nanoparticle-based delivery systems, and alternative production platforms such as plant-based expression systems, are expanding the therapeutic landscape. At the same time, emerging strategies aimed at modulating immune responses and optimizing treatment protocols are contributing to improved clinical management.

Finally, the evolving role of ERT is discussed in the context of emerging therapies, particularly gene therapy. While gene-based approaches hold great promise, ERT is likely to remain an essential component of treatment, as a complementary strategy or a bridge to more definitive therapies. Lessons learned from decades of clinical use provide a foundation for future innovations aimed at overcoming current limitations and improving outcomes.

Keywords: Blood-brain-barrier, Central nervous system, Drug delivery, Enzyme replacement therapy, Gene therapy, Immunogenicity, Inborn errors of metabolism, Lysosomal storage diseases.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

THE HISTORY OF ENZYME REPLACEMENT THERAPY

Enzyme replacement therapy (ERT) was devised for the treatment of lysosomal storage diseases (LSDs) more than 35 years ago. Pivotal studies[1] identified the mechanism underlying the sorting and trafficking of lysosomal enzymes and provided the rationale for a strategy based on replacing a missing lysosomal enzyme through the infusion of the normal enzyme. Based on these studies, ERT was implemented and translated into human therapy for the first time in patients affected by Gaucher disease[2] (Figure 1).

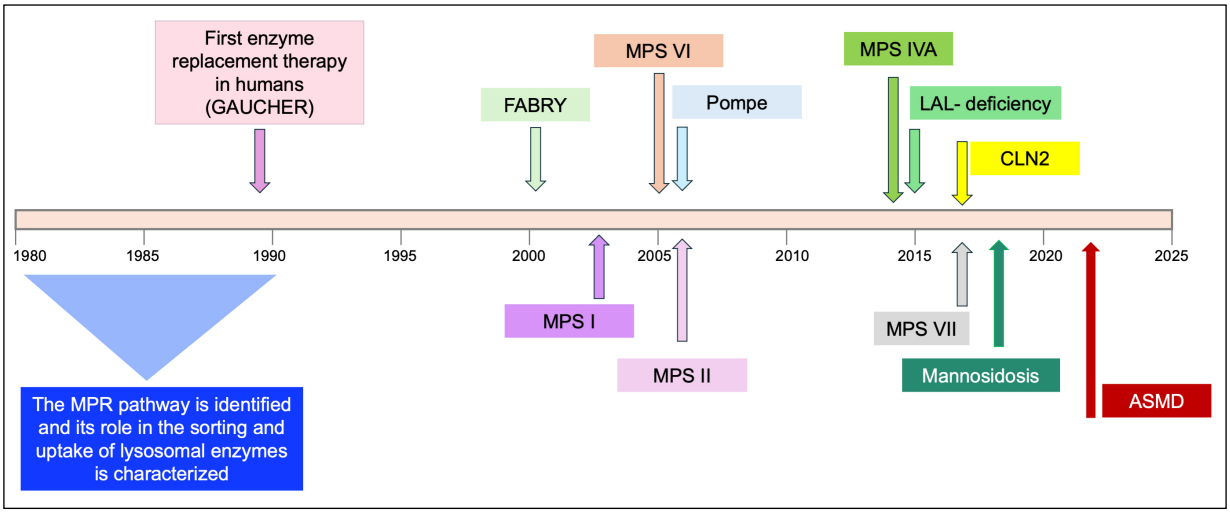


Figure 1. History and first therapies for each disease. The first studies that identified the mechanism underlying the sorting and trafficking of lysosomal enzymes and provided the rationale for ERT date to the 1980s. In the early 1990s ERT was translated into human therapy in patients affected by Gaucher disease. Since then, ERT was progressively extended to several other LSDs. The figure lists the first drugs developed for each different LSD.

Thanks to the success of this approach in the non-neuronopathic forms of Gaucher disease[3], ERT was progressively extended to several other LSDs, on the hypothesis that it could be equally effective for other disorders (Figure 1). The currently approved enzyme replacement therapies are summarized in Table 1.

Overall, thousands of patients with different LSDs have been treated, and there is now sufficient insight – at least for some of these disorders – to evaluate the results of substantial commitment and efforts of patients, physicians and institutions, to assess the yield of huge investments in research and in clinical development, and to indicate new directions and shape future actions to treat lysosomal disorders.

In assessing the results of ERT, we must consider that, although the approach is based on a common strategy and on common molecular mechanisms, ERT is not a “one-for-all strategy”. There are peculiar aspects, specific for individual LSDs (and possibly for single groups of patients), that impact therapeutic efficacy.

We must also consider that ERT has substantially evolved over time, for example with novel manufacturing processes and modifications introduced in the recombinant enzymes, and with new-generation therapeutic enzymes that have been produced to overcome some of the hurdles and drawbacks that emerged from real-world experience. Also, we must compare ERT with newer approaches that have been introduced as alternative or complementary therapies for the treatment of LSDs.

This article is a concise review of some of the issues that have emerged after decades of ERT and is not intended to be comprehensive. State-of-the-art and extensive reviews are available and can be easily found in the literature, for example van der Beek et al[4] (2025), Poswar et al[5] (2019), or Desnick and Schuchman[6] (2012), and are not within the scope of this article. The purpose of our review is rather to address selected aspects that are critical for evaluating the outcomes of enzyme replacement therapy and for supporting the further development of this therapeutic approach.

Table 1. Approved enzyme replacement therapies, including second-generation enzymes.

Disease	Enzyme	Route of administration
Type 1 Gaucher disease	Imiglucerase Velaglucerase alfa Taliglucerase alfa	IV
Fabry disease	Agalsidase alfa Agalsidase beta Pegunigalsidase alfa	IV
Mucopolysaccharidosis type I	Laronidase	IV
Mucopolysaccharidosis type II	Idursulfase alfa Idursulfase beta (in selected countries) Pabinafusp alfa (in Japan)	IV
Mucopolysaccharidosis type VIA	Elosulfase alfa	IV
Mucopolysaccharidosis type VI	Galsulfase	IV
Mucopolysaccharidosis type VII	Vestronidase alfa	IV
GSD type II (Pompe disease)	Alglucosidase alfa Avalglucosidasi alfa Cipaglucosidasi alfa (+ miglustat)	IV
Lysosomal acid lipase deficiency	Sebelipase alfa	IV
Alpha-Mannosidosis	Velmanase alfa	IV
Neuronal ceroid lipofuscinosis type 2 (CLN2)	Cerliponase alfa	ICV
Acid sphingomyelinase deficiency	Olipudase alfa	IV

Intracerebroventricular (ICV), intravenous (IV), glycogen storage disease (GSD).

MATERIALS AND METHODS

A comprehensive literature search was conducted using the PubMed and Scopus databases to identify relevant articles on enzyme replacement therapy for lysosomal storage diseases. The search strategy included combinations of the following keywords: “enzyme replacement therapy”, “lysosomal storage disorders”, “Pompe disease”, “Gaucher disease”, “Fabry disease”, “mucopolysaccharidosis”, “acid lipase deficiency”, “CLN2” “immune response”, “pharmacokinetics”, “pathophysiology” and “gene therapy”. One hundred and six articles published in English up to 2025 were considered.

Original research articles, clinical trials, systematic reviews, and relevant preclinical studies were included based on their relevance to the topic. Particular attention was given to studies addressing efficacy, limitations, immune responses, and novel therapeutic strategies. Case reports, abstracts without full text, and studies not directly related to enzyme replacement therapy or lysosomal storage diseases were excluded. Additional relevant publications were identified through manual screening of reference lists of selected articles.

THE RATIONALE FOR ERT

Translating ERT from the level of proof-of-concept molecular studies into actual therapies for patients has been a complex process. To be effective in disorders like LSDs and to be translatable into patients’ care, a treatment must meet several important requisites, related to the biology and pathophysiology of the diseases, to manufacturing procedures, and to the viability of this approach in clinical practice.

First, the development of a treatment should be supported by robust biological evidence, specifically with regard to the cellular pathway that allows for uptake and delivery of enzymes to lysosomes, where their action is required to replace the missing activity.

Enzyme manufacturing procedures are also critical. Therapeutic enzymes must be produced in sufficient amounts according to "good manufacturing procedures". These processes must be feasible, scalable, and sustainable.

The therapeutic enzyme should be administered to patients *via* a suitable and efficient route that is compatible with patients' conditions and with lifelong treatment.

Once administered intravenously, recombinant enzymes begin their journey through the patient's circulation, which must lead to optimal biodistribution to tissues and correction of the missing enzyme activity in the right cells and enzyme compartment (Figure 2). This means that enzyme proteins should be stable in the circulation and penetrate physiological barriers.

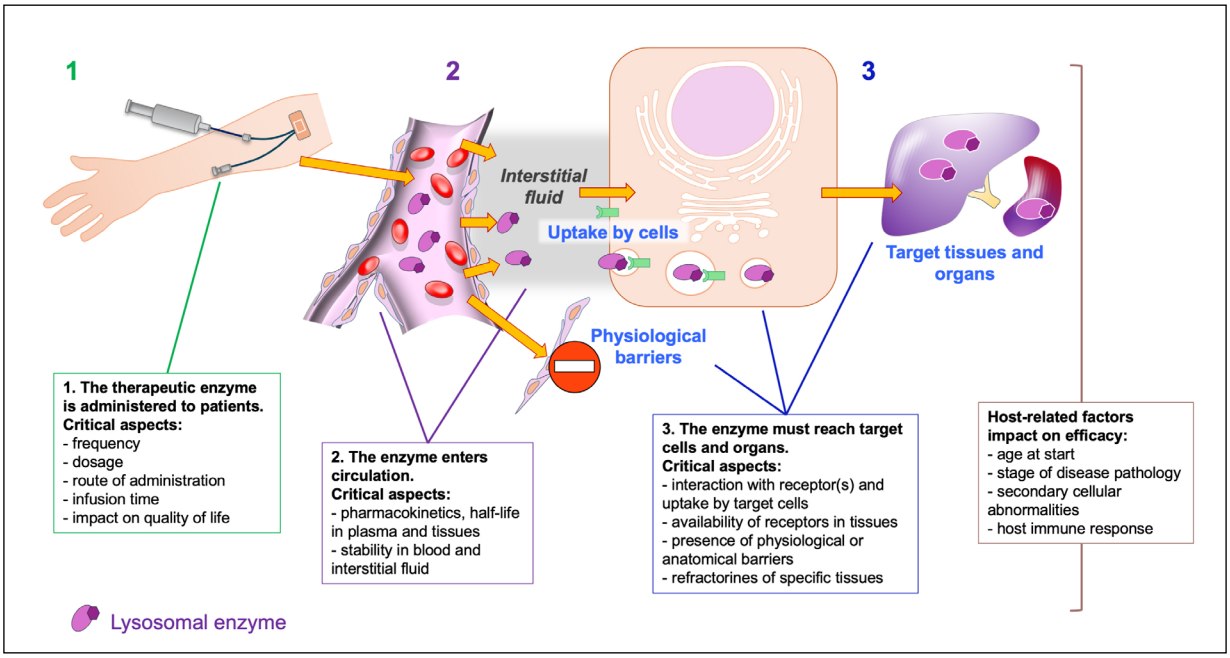


Figure 2. Important factors that influence recombinant enzymes trafficking, delivery to target tissues, and efficacy. (**1**) Therapeutic enzymes are administered to patients intravenously. (**2**) Once administered, recombinant enzymes start their journey in the patient circulation and the interstitial fluid. (**3**) Enzymes must reach target tissues where they must be taken up through receptor-mediated processes and reach therapeutic concentrations in tissues and organs. Boxes indicate factors that are relevant for each of these steps and impact on efficacy. In addition (right), host-related factors also impact therapeutic efficacy.

After trafficking through the circulation and interstitial fluid and reaching target tissues, enzymes must be taken up *via* receptor-mediated processes to reach therapeutic concentrations.

In addition to enzyme properties, another issue is the emergence of host-related factors as critical determinants of therapy success.

Aspects related to the conduct of studies to assess efficacy also warrant careful analysis. ERT should be tolerated by patients and translate into clinical benefits. It is often difficult to identify reliable outcome measures to evaluate treatment efficacy and patient outcomes.

The main critical factors influencing the efficacy of ERT are summarized in Table 2 and discussed in the following sections.

Table 2. Main limitations and challenges associated with enzyme replacement therapy.

Critical issue	Example (disease)	Underlying mechanism/notes
Limited delivery to target tissues	Pompe disease (skeletal muscle), Gaucher disease (bone)	Low receptor density (M6P receptor), impaired uptake, tissue-specific barriers
Restricted penetration across physiological barriers	Mucopolysaccharidoses, CLN2	Blood–brain barrier limits CNS delivery
Suboptimal biodistribution	Pompe disease, Fabry disease	Preferential uptake by liver and spleen rather than affected tissues
Need for lifelong, repeated infusions	All LSDs	Short half-life of recombinant enzymes
Infusion-associated reactions	Pompe disease, Fabry disease	Hypersensitivity, complement activation
Development of anti-drug antibodies (ADA)	Pompe disease (CRIM-negative), Fabry disease	Neutralizing antibodies reduce efficacy
Variable clinical response across tissues	Pompe disease (heart *vs*. skeletal muscle), Gaucher disease (bone)	Differences in vascularization, receptor expression, intracellular trafficking
High treatment costs	All LSDs	Complex manufacturing and chronic administration
Limited effect on advanced disease	Most LSDs	Irreversible tissue damage at late stages
Incomplete correction of secondary pathways	Pompe disease, Gaucher disease	Persistent autophagy, inflammation, cellular dysfunction
Burden on quality of life	All LSDs	Hospital-based infusions, time-consuming treatment

Neuronal ceroid lipofuscinosis type 2 (CLN2), lysosomal storage diseases (LSDs), cross-reactive immunologic material-negative (CRIM-negative), central nervous system (CNS).

THE BIOLOGICAL EVIDENCE SUPPORTING ERT DEVELOPMENT

The intuition that an exogenous enzyme could correct an enzyme deficiency in an LSD and the consequent pathological accumulation of undegraded substrate originated from seminal studies[7] that were performed between the late 1970s and the 1980s.

Two lines of research were particularly relevant in this respect.

First, studies[8] performed between the late 1970s and the 1980s provided information on the sorting of lysosomal enzymes from the endoplasmic reticulum, where they are synthesized, to the lysosomes. Additional information was derived from studies[9] that clarified the concept of cross correction and recognized the therapeutic potential of the exchange of lysosomal enzymes between different cells.

All these studies showed that the mannose-6-phosphate receptor (MPR) pathway plays a central role in this process. Lysosomal enzymes synthesized in the endoplasmic reticulum (Figure 3A) acquire mannose-6-phosphate (M6P) residues in the Golgi, so that they can be delivered to lysosomes *via* the MPR biosynthetic pathway. As the MPR is also expressed at the cellular plasma membrane, exogenous lysosomal enzymes can be internalized by defective cells through the endocytic pathway and reach the lysosomal compartment, thus replacing a missing enzyme.

The MPR is a multifunctional, ubiquitously expressed membrane glycoprotein of a molecular weight of approximately 300 kDa that binds various ligands through its modular extracellular region (Figure 3B). Only four of the domains present on MPR are responsible for the recognition and binding of M6P [or mannose-6-(bis)phosphate] and/or phosphodiester M6P-GlcNAc residues exposed on the N-glycans of lysosomal enzymes, whereas other domains mediate interaction with different ligands, such as insulin-like growth factor 2 (IGF2)[10,11].

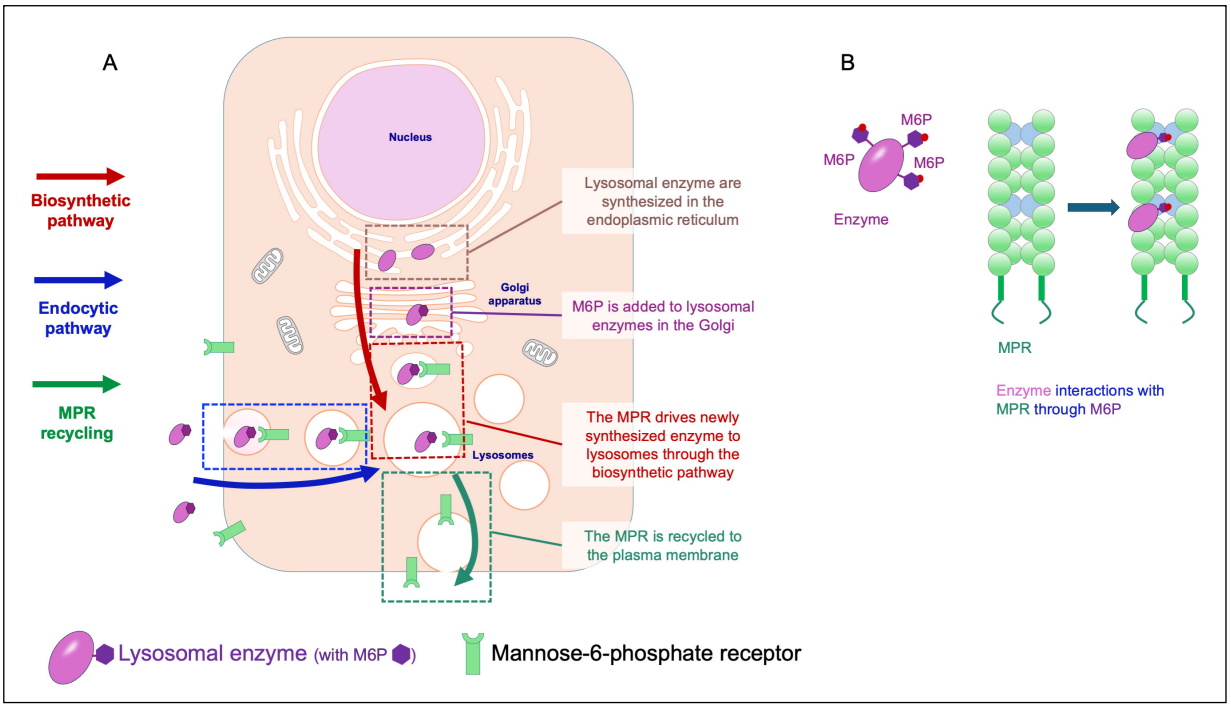


Figure 3. The MPR pathway and the trafficking/uptake of lysosomal enzymes. **A**, Enzymes are synthesized in the endoplasmic reticulum (brown dotted box). Oligosaccharide chains of lysosomal enzymes are processed in the Golgi apparatus, and M6(bis)P is added (purple dotted box), and lysosomal enzymes are delivered to lysosomes (red dotted box). Lysosomal enzymes can also be internalized through the endocytic pathway, exploiting the MPR present at the plasma membrane, and delivered to the lysosomes (blue dotted box). The MPR dissociates from the enzymes and is then recycled to the plasma membrane (green dotted box). **B**, Lysosomal enzymes are endowed with one or more M6P residues that are exposed on the oligosaccharide chains. M6P residues interact with the MPR, a protein that is composed of two subunits with repetitive motifs. Six out of its 15 homologous domains have ligand-binding properties. Among them, domains 3, 5, 9 and 15 are those involved in binding of mannose-6-phosphate (M6P) and/or phosphodiester M6P-GlcNAc residues that are exposed on the N-glycan chains of lysosomal enzymes. Two M6P binding sites (pink) are available in each MPR molecule.

Other enzymes, for example beta-glucocerebrosidase (GBA), the enzyme that is defective in Gaucher disease and was the first enzyme developed for ERT, exploit different receptors for their trafficking. The Cluster of Differentiation 206 (CD206) receptor interacts with Mannose (MAN) and drives GBA into macrophages and cells of the reticuloendothelial system[12]. Additional and alternative processes of lysosomal targeting also exist, such as those that are mediated by lysosomal integral membrane protein (LIMP-2) and sortilin[13]. In the following paragraphs, we will focus on the MPR pathway as a prototype of the cellular processes involved in lysosomal enzyme trafficking.

The notions on MPR structure and function have important clinical implications, particularly to understand what factors impact or limit the uptake and trafficking of lysosomal enzymes. For example, it is important and relevant for ERT that one receptor can bind multiple molecules of lysosomal enzymes simultaneously (at least two, and potentially three, depending on the structure and the specific lysosomal enzyme). It is also essential that recombinant enzymes must be endowed with M6P or M6(bis)P residues to be internalized by cells and correctly trafficked to lysosomes.

The need for concomitant action of both MPR and M6P in determining the success or the failure of ERT is illustrated by the experience in Pompe disease. Albeit effective in improving patients' cardiac and motor functions, and in extending their survival, ERT with alglucosidase alpha (the first-generation recombinant human alpha-glucosidase) has shown limitations, particularly in correcting skeletal muscle pathology and function. Both MPR and M6P were thought to be responsible for these limitations. Specifically, the relative refractoriness of muscles was ascribed to both sub-optimal amounts of M6P in the enzyme preparation[14], and to the relatively low abundance of MPR available at the plasma membrane in skeletal muscles[15].

To address the scarcity of receptors in skeletal muscles in Pompe disease, two alternative strategies have been explored. One is based on drugs that impact MPR expression at the plasma membrane of muscles in Pompe disease patients and favor response to ERT in these organs. Specifically, beta-2-agonists have been shown in preclinical studies[16,17] to be effective in this respect and have been brought to clinical translation[18-20].

Alternatively, enzyme preparations enriched in M6P (or MPR ligands) have been developed to facilitate interactions with the MPR and delivery of recombinant enzymes to target tissues. Also in this case, the example is Pompe disease, with two second-generation enzymes that have been approved for clinical use[14,21,22].

Even though much information has been gathered on the molecular basis of ERT and many therapeutic recombinant enzymes for different diseases have been translated into clinical use, there are issues that need to be fully clarified. For example, how many M6P residues on an enzyme molecule are required for optimal interactions with MPR? Is the optimal number of M6P residues the same for all enzymes? What is the most limiting factor in the uptake of the enzymes, the amounts of MPR at the plasma membrane (that may be saturated by an excess of ligand), or the number of M6P or M6(bis)P on the N-glycan chains of the enzyme? Is it possible that the same lysosomal enzyme exploits different routes for uptake in cells?

A deeper understanding of these mechanisms is of biological interest and has important clinical implications, ultimately enabling the rational design of improved therapeutic enzymes and the optimization of treatment strategies.

ENZYME MANUFACTURING

The procedures required for the manufacturing of recombinant enzymes have varied in the past and may continue to evolve significantly with the introduction of newer technologies. The first preparation (alglucerase) of therapeutic beta-glucocerebrosidase was extractive, derived and purified from human placentae[2] (Figure 1). This procedure required the collection and processing of large amounts of placentae to obtain large quantities of pure enzyme and was associated with the risk of transmitting infectious agents.

With the advent of recombinant DNA techniques (methods in which expression vectors, such as plasmids, are used to introduce a gene of interest into host cells, enabling the production of the corresponding recombinant protein), therapeutic enzymes were produced in various lines of mammalian cells. In this way, it was possible to address the issues related to the use of extractive enzymes. However, other challenges emerged, such as the need for high yield, efficient purification, and sterility. Indeed, the risk of contamination materialized in the years 2019-2010 with an incident in one of the manufacturing plants. This event led to a shortage of some enzymes used for ERT and to the need for some patients to temporarily stop treatment or switch to alternative therapies[23]. This episode showed how important the production of high-quality therapeutic enzymes is and how this can be a limiting factor for a therapy like ERT that is based on repeated, lifelong infusions.

Even if production technologies have evolved over time, the high costs of recombinant enzymes remain a major challenge associated with ERT. The costs of the enzyme preparations may represent a heavy burden for national health services or limit access to ERT in underdeveloped countries. This issue has been analyzed in detail[22-26]. The treatment of a single patient may cost as much as several hundred thousand euros per year.

The final marketing price of the enzymes is the result of several factors, including manufacturing procedures and the need for specialized factory plants equipped to produce enzymes with recombinant DNA techniques according to good manufacturing practice (GMP). Investments required for preclinical and clinical research also heavily impact the price of ERT. In general, the development of a new enzyme must be financially supported for long periods, as much as 10-15 years, which is the period usually required for clinical development of a drug[27].

A possible strategy to address issues in enzyme manufacturing procedures may involve using alternative, innovative expression platforms to generate recombinant therapeutic proteins, such as plants or plant cells. The use of plants as bioreactors for therapeutic production (plant molecular “pharming”) has advantages in terms of potentially lower costs (largely due to high scalability)[28] and absence of human and animal pathogens that can contaminate animal-based production systems. Seed production systems are also attractive, as therapeutic proteins produced in this way exhibit high, prolonged stability when stored as dry seeds.

Potential drawbacks may derive from the different glycosylation of enzymes produced in plant systems. As mentioned previously, proper glycosylation and the presence of ligands (such as M6P or MAN) on the N-glycan chains of the enzyme are critical for the uptake of lysosomal enzymes. The protein glycosylation machinery in plants is different from that of mammalian cells, and the N-glycosylation profile in plant-derived enzymatic proteins is different from that obtained in mammalian cells[29]. Nevertheless, plant-derived enzymes so far developed are endowed with glycans that allow tissue delivery and are active. Currently, two enzyme preparations produced in plants have already been granted approval from regulatory agencies[30,31] for the treatment of Gaucher disease and Fabry disease[32-36], while others are still in a phase of pre-clinical development[35]. In this context, multiple plant-derived enzymes are currently being explored, particularly for Pompe disease, where recombinant acid α-glucosidase (GAA) has been expressed in various plant platforms, including tobacco chloroplasts, transgenic seeds, rice cell cultures, and glycoengineered systems, with encouraging results in terms of enzymatic activity, immune modulation, and glycogen clearance in disease models. Similar approaches have also been applied to other lysosomal storage disorders, including the production of α-galactosidase for Fabry disease (complementing the already approved pegunigalsidase), α-L-iduronidase for mucopolysaccharidosis type I, and lysosomal acid lipase for Wolman disease, further supporting the potential of plant-based platforms for enzyme replacement therapy[37].

The evolution of technologies has provided significant innovation and has addressed some of the shortcomings of ERT. As the targeting of recombinant enzymes was sub-optimal in some tissues, for example in skeletal muscles in Pompe disease patients, second-generation enzymes enriched in M6P and with improved targeting properties have been developed[21], one of them in combination with a stabilizing molecule[38].

Engineered enzymes that are able to cross physiological barriers have also been developed and will be discussed in another paragraph in this review.

New technologies will likely offer additional opportunities in the coming years to enhance evolution and innovation in this field. It is possible to envisage that the main goals of research will be to develop convenient and affordable platforms for the production of recombinant enzymes to further reduce costs and explore innovative approaches for optimal tissue delivery (such as nanoparticles)[39]. In addition to enhancing biodistribution and tissue targeting, nanoparticle-based systems may also contribute to reducing immunogenicity by shielding the enzyme from immune recognition[40]. In parallel, significant efforts are expected to be directed toward the design of second- and third-generation enzymes with enhanced targeting properties, such as chimeric constructs. Reduced immunogenicity may be achieved through protein and glycoengineering strategies, as well as formulation approaches such as PEGylation[41,42]. In addition, alternative strategies aimed at modulating the immune response, including oral tolerance approaches, have also been explored in preclinical models[43].

Further advances are likely to derive from approaches aimed at increasing enzyme stability and catalytic efficiency, leading to the development of so-called “super-enzymes”. In this context, evolution-guided strategies such as ancestral sequence reconstruction have shown that engineered variants can display increased catalytic activity and improved substrate clearance in preclinical models, potentially enabling lower dosing and improved therapeutic outcomes[44].

ADMINISTRATION TO PATIENTS

ERT is based on repeated, periodic intravenous infusions (in most cases every week or fortnightly, depending on the enzyme half-life and the desired rate of substrate clearance). Infusion times vary and in general last several hours. In young patients, repeatedly finding venous access is often difficult and may cause pain and discomfort, thus requiring the implantation of intravenous devices. The use of implantable intravenous devices, such as ports or central intravenous catheters, may be associated with serious risks (for example, infections). In rare instances (for example for cerliponase), these problems are exacerbated by the need for intrathecal administration.

For these reasons, the impact of ERT on patients’ quality of life is heavy. To alleviate the burden of periodic hospital infusions on patients, home therapy programs have been proposed and implemented for some LSDs in several countries[26,45]. However, home therapy is not available for all ERTs and not for all patients. In principle, a patient who is stable for months and has not developed major immune IgE-mediated reactions may be eligible for home administration. However, there are disparities in the availability of home therapy in different countries, sometimes in different regions of the same country.

The impact of repeated, lifelong infusions may also be addressed by exploiting new technical developments. For example, new-generation enzymes with prolonged half-life and stability may be injected less frequently and reduce the burden on patients[35]. Combination protocols of ERT and therapies such as substrate-reducing agents may also be envisaged to enhance efficacy and potentially reduce dosing requirements. For example, co-dosing of glycogen synthase type 1 (GYS1) inhibitors and ERT with recombinant human alpha-glucosidase has been effective in preclinical and clinical studies[46,47] in Pompe disease.

Similar combinatorial approaches are being investigated in other lysosomal storage disorders. In type 3 Gaucher disease, substrate reduction therapy has been used as an alternative or, in selected cases, in combination or sequence with enzyme replacement therapy to optimize disease control and reduce substrate accumulation[48].

CIRCULATION, PHARMACOKINETICS AND STABILITY OF RECOMBINANT ENZYMES

Once the enzyme has been infused into the patient, it must travel through the bloodstream, penetrate any necessary physiological barriers, and be available to target tissues and cells. Within target cells, it must reach its final destination in the specific cellular compartment, the lysosomes, where its action is required. This trip to the lysosomes is not without risks for the enzyme.

Pharmacokinetics and blood half-life are the results of intrinsic properties of the enzyme, such as stability and conformation, and by other factors, for example uptake by target cells and non-productive clearance by tissues where its activity is not desired[49]. Stability of the enzyme may substantially vary among different enzymes or different enzyme preparations of the same enzyme[50]. Stability depends on several factors, including subcellular localization, molecular crowding in a certain environment or organelle, interactions with other proteins, and pH[51]. pH is particularly important. Lysosomal enzymes have evolved to exert their function and to have optimal stability in an acidic environment such as the lysosomal compartment. For example, the stability of alglucosidase (rhGAA), the enzyme used since 2006 for the treatment of Pompe disease, was shown to be affected by pH[52] and by persistence in blood[53]. Exposure to non-optimal conditions may also occur in the interstitial fluid[54] and, possibly, in the endocytic compartment of cells, before reaching the lysosomes. A way to address this issue has been translated into human therapy by co-dosing a second-generation recombinant alpha-glucosidase preparation and a stabilizer molecule for the treatment of Pompe disease[22].

UPTAKE, TISSUE DISTRIBUTION, DELIVERY TO TARGET TISSUES

A major determinant of efficacy of a recombinant enzyme is its ability to reach target tissues, where its action is required to correct disease pathology.

In general, visceral organs appear to respond better, possibly due to diffuse vascularization that allows enzymes to reach therapeutic concentrations in these tissues. Thus, LSDs primarily involving viscera, for example Gaucher disease type 1[55] or lysosomal acid lipase deficiency[56], can be treated with better chances of success.

This concept is exemplified by Wolman disease, a predominantly visceral disorder in which ERT has shown marked efficacy and improved survival[57]. However, despite these encouraging results, residual disease manifestations may persist in macrophage-rich compartments, including the monocyte–macrophage system (MAS)[58]. In these sites, substrate clearance may remain incomplete, suggesting that even in well-perfused organs, cell-specific uptake and intracellular delivery of the enzyme represent limiting factors[59].

Some tissues appear particularly refractory to ERT. For example, in Gaucher disease, bone is difficult to correct[60]. Bone refractoriness in Gaucher disease reflects an imbalance between impaired osteoblast function and increased osteoclast activity, associated with a pro-inflammatory microenvironment driven by lipid accumulation[60,61]. Dysregulation of molecular signals that direct osteoblast commitment (sirtuin 1, runt-related transcription factor 2 protein) and increased oxidative stress may also contribute to poor response of bone manifestations to ERT[62].

A similar tissue-specific response is observed in Pompe disease, where cardiac involvement, especially in infantile-onset forms, improves rapidly with treatment, whereas skeletal muscles show a slower and often incomplete response. This difference is mainly attributed to reduced uptake of the recombinant enzyme in skeletal muscle, lower expression of mannose-6-phosphate receptors, and impaired trafficking and processing within myofibers[4,63].

The presence of physiological or anatomical barriers is one of the factors that limit the biodistribution of therapeutic enzymes. The blood-brain barrier (BBB) is perhaps the most studied and better characterized barrier, although physiological barriers may also exist in other tissues, for example the blood-retina barrier (BRB) in the eye[64].

In most LSDs, brain involvement may occur, causing progressive neurological deterioration and handicap[65]. In these disorders, correcting brain pathology and function is a primary goal, and failure to reach therapeutic levels in the central nervous system (CNS) is a major limitation. Recombinant lysosomal enzymes are large molecules that are unable to cross a selective barrier like the BBB. Thus, intravenous administration of the recombinant enzymes used for ERT (as they have been designed and manufactured over the past decades) is not expected to have therapeutic effects on the central nervous system.

In principle, different approaches may be exploited to overcome this problem. A straightforward strategy would be to change the route of administration and inject the therapeutic enzyme directly into the CNS.

A second approach would be to modify the enzyme's properties by designing signal domains that enable trafficking across the brain microvascular endothelium and other cells that comprise the BBB.

Another emerging strategy involves encapsulating therapeutic enzymes within nanocarriers, including liposomes, polymeric nanoparticles, solid lipid nanoparticles, and extracellular vesicles, to improve enzyme stability, biodistribution, and tissue-specific delivery.

Direct intrathecal administration has been proposed for mucopolysaccharidoses I, II and III[66,67]. In patients with neuronopathic mucopolysaccharidosis type II, therapy was based on combining weekly intravenous and monthly intrathecal administration of idursulfase, the enzyme used for ERT in this disorder. While intrathecal administration lowered cerebrospinal fluid concentrations of heparan sulfate, its effect on cognitive outcomes appeared variable[68]. A potential drawback of this approach in LSDs associated with complex phenotypes like mucopolysaccharidoses would be the lack of efficacy on visceral manifestations, thus requiring a combination of intravenous and intrathecal injections. Therefore, it would be more appropriate to use this approach in disorders characterized by predominant neurological impairment.

For children with symptomatic neuronal ceroid lipofuscinosis type 2 (CLN2) due to tripeptidyl-peptidase 1 deficiency, recombinant human cerliponase alfa has been approved for clinical use and can be administered intracerebroventricularly. The first study in 23 children showed that patients treated with cerliponase alfa had a slower rate of decline in a motor–language score on the CLN2 Clinical Rating Scale compared with historical control participants[69]. Re-assessment of cerliponase alfa efficacy in 24 CLN2 patients confirmed slowed decline in motor and language function[70]. However, this therapy must be used in a suitable setting by experienced teams to prevent complications like intrathecal infections (about 40%).

In a cohort of patients affected by mucopolysaccharidosis type IIIA, due to heparan-N-sulfatase (HNS) deficiency, intrathecal administration of a recombinant human HNS showed that although cognitive decline continued to be apparent in most patients, improvements were observed in a small subgroup[71].

The use of chimeric enzymes has also reached clinical development. To this purpose, engineered recombinant enzymes are fused to peptides that allow penetration of the BBB and brain delivery through interaction with specific receptors, such as the transferrin, insulin and low-density lipoprotein (LDL) receptors. Pabinafusp alpha, a human iduronate-2-sulphatase fused to an anti-human transferrin receptor antibody, has been approved in Japan for the treatment of neuronopathic mucopolysaccharidosis type II. Clinical trials[72,73] showed consistent reductions in heparan sulfate concentrations. Some, but not all, patients showed positive effects on the Bayley Scales of Infant and Toddler Development. Trials using a similar strategy are underway for mucopolysaccharidoses type I and II. Initial results are promising in terms of biomarkers, but it is too early to draw conclusions about benefits on cognitive outcome[74,75].

The use of enzyme-loaded carriers, such as nanoparticles, also holds promise but remains in a preliminary phase. These systems, based on innovative, biocompatible nanomaterials, have the potential to enhance the efficacy of ERT and minimize side effects. They may improve tissue targeting and passage across the BBB by decorating the nanocarrier surface with targeting molecules (peptides, aptamers, and antibodies)[76]. In addition, they may address some of the limitations of ERT, such as immunologic reactions and enzyme stability. This approach has been proposed for different LSDs, for example Krabbe disease [due to galactosylceramidase (GALC) deficiency][77,78].

HOST-RELATED FACTORS

In addition to the intrinsic properties of the enzyme, pharmacokinetics and delivery of enzymes to target tissues, host-related factors play a critical role in determining the success of ERT. This is a broad and complex topic that includes many aspects.

Some factors may appear obvious, for example, the severity of the disease. Most LSDs encompass broad spectra of presentations that range from severe (and often early-onset) manifestations to attenuated and late-onset phenotypes. In principle, it would be reasonable to expect that ERT shows variable efficacy in the different forms and that milder forms may respond better. However, this is not consistently true. For example, in Pompe disease, a severe and life-threatening manifestation such as hypertrophic cardiomyopathy is efficiently corrected by ERT, thus improving patients' survival, while skeletal muscle involvement, typical of the late-onset forms, is more difficult to correct and appears refractory to treatment[4].

Other factors involved in the differential response of cells or tissues to the treatment are related to the pathophysiology of the disease. For example, the stage of disease pathology, and secondary tissue and cell abnormalities triggered by storage[79] are emerging as important challenges.

One of the pathways that are affected in several LSDs is autophagy[80]. In Pompe disease, some studies[81,82] have demonstrated that the recombinant enzyme used for ERT is aberrantly delivered to sites of autophagic accumulation and is ineffective.

A consequence of the impairment of the autophagic flux is oxidative stress[83]. It has been shown that oxidative stress impacts the level of correction of alpha-glucosidase activity by ERT *in vitro* and *in vivo*. In fibroblasts and in the animal model of Pompe disease, there is an inverse correlation between the level of correction of alpha-glucosidase activity by ERT and the level of tissue stress. The use of known antioxidants facilitated the correction of activity, suggesting a complementary approach to improving ERT efficacy in this disorder.

The age at start of therapy is related to these aspects. It is reasonable to think that an early start of therapy, before the development of massive tissue pathology, may prevent irreversible abnormalities. In this regard, the availability of reliable measures of disease progression and the development of guidelines may serve as valuable tools for defining the optimal time to start therapy.

Newborn screening programs may also be a valuable tool for timely identification of patients, particularly for disorders that need to be treated as soon as possible, for example metachromatic leukodystrophy and Pompe disease[84,85]. In recent years, screening for lysosomal storage disorders has been progressively implemented in several countries, including pilot and regional programs. In Italy, a pilot program for metachromatic leukodystrophy in Tuscany demonstrated the feasibility of a two-tier diagnostic approach[86], while a long-term newborn screening program in the North-East region confirmed the clinical utility of early diagnosis in Pompe disease, enabling prompt treatment and improved outcomes[87].

Complications due to immune responses to ERT are emerging as a significant challenge in patients exposed to ERTs in several LSDs.

Some patients, particularly those carrying null mutations leading to the complete absence of the enzyme protein [cross-reactive immunologic material-negative (CRIM-negative)], may develop high antibody titers that can cause adverse reactions, including mild to life-threatening manifestations, or affect the efficacy of ERT, leading to treatment failure[88].

Strategies to mitigate patients' immune responses have been developed, mainly based on the use of antihistamines and antipyretics, or slowing the rate of infusion without interruption or discontinuation of treatment[89]. Prophylactic tolerance induction protocols have been developed. Currently, standard protocols to mitigate antibody response are based on multi-drug approaches with the combination of rituximab, methotrexate, and intravenous immunoglobulin (IVIG)[90,91]. Moreover, additional strategies have been explored, including immunomodulatory approaches targeting plasma cells or promoting antigen-specific tolerance[88,91,92]. These advances underscore the need for early identification of patients at higher immunological risk and support the adoption of tailored treatment strategies to improve long-term efficacy and safety[91].

However, it could be expected that the evolution of technologies will provide additional therapeutic strategies. The development of engineered and less immunogenic enzymes may represent an option to mitigate the activation of immune responses. Engineering stable and non-immunogenic enzymes has already been proposed in other medical fields, for example, for the treatment of cancer[93].

Another potential technology is based on blood-circulating immune cells as producers of the needed therapeutic protein[94].

Another emerging therapeutic approach relies on gene-based strategies that enable sustained endogenous production of the therapeutic enzyme. Among these, liver-directed gene therapy has shown particular promise by inducing antigen-specific immune tolerance through the generation of regulatory T cells, thereby reducing anti-drug antibody formation and enhancing the long-term efficacy of enzyme replacement therapy[95].

EFFICACY AND ASSESSMENT OF EFFICACY

The steps in the clinical development and assessment of ERTs for LSDs are, for several aspects, comparable to those required in other fields of medical sciences. However, there are some peculiarities of LSDs that may be associated with specific challenges.

First, LSDs are rare or ultra-rare diseases. Enrolling a sufficient number of patients in a clinical study to obtain reliable and significant results and to assess therapeutic efficacy is often difficult. This is even more complicated by the fact that within the same disorder, broad clinical variability is common, and it may be difficult to select homogeneous patient cohorts in a study.

The rarity and the clinical variability of LSDs also impact the feasibility of natural history studies, which are essential as comparators to assess efficacy. For several LSDs, these studies are quite demanding and require the collaboration of multi-center networks.

Reaching conclusive results is also difficult because of the slow progression of manifestations in many LSDs. As most approvative trials last one or two years, it is difficult to expect appreciable, clinically meaningful, and statistically significant improvements. The insight deriving from decades of real-life experience in the field of LSDs tells us that only after a very long period it is possible to ascertain the true effects of ERT.

Additional challenges are posed by the fact that LSDs are often multisystem diseases[96]. As mentioned in the previous paragraphs, some tissues or organs may respond in different ways to therapies, either due to local conditions or to different bioavailability of therapeutic enzymes.

For these reasons, it is difficult to design a study and select *a priori* the endpoints that are expected to show significant changes and to reflect the efficacy of a therapy. Ideally, the endpoints of a trial should be clinical measures. However, clinical measures often require invasive procedures or are difficult to obtain as they need specific skills and expertise of assessing doctors and participation from patients that are highly debilitated or affected by variable intellectual impairment. For some studies, for example, intracerebroventricular cerliponase[69,70,97], efficacy endpoints were based on longitudinal assessments of neurological function, including the rate of decline in motor and language performance. It would be preferable that the assessment of therapies is based on multiple evaluations, including patient-reported outcome measures that are emerging as valuable tools for this purpose[98] and surrogate measures, such as biomarkers[99].

Assessing the real value and role of biomarkers as complementary measures of efficacy will be a challenge for the future[100]. A few new therapies have been approved by regulatory agencies exclusively based on changes in biochemical markers. In LSDs, most biochemical markers are substrate-related (GAGs, GLC4, lysoGB3, etc). Only in a few cases are markers more functional and more reflective of secondary changes involved in the pathophysiology of different disorders (cytokines, neurofilament light chain, microRNAs)[101-103].

Future efforts in this regard are expected to be challenging and should follow different trajectories. It is indispensable to invest in natural history studies and registries, with concerted action by networks based on multicenter collaborations. Centers involved in these actions should be qualified and expert in the management and follow-up of LSDs, a requisite that is not common for these rare disorders.

It will be important to invest in research to better understand the reasons for the different responses of different tissues that are currently known only in part. Expanding our knowledge on LSD pathophysiology is expected to provide clues to clarify this aspect.

CONCLUSIONS

The information on the sorting of lysosomal enzymes and on the MPR pathway, showing that a lysosomal enzyme provided by an external source can be taken up by cells and replace a missing activity, indicated that LSDs are an ideal target for ERT. Indeed, after the first attempts in Gaucher disease, several ERTs were granted regulatory approval and were introduced in the treatment of many other LSDs.

Real-life experience, however, has shown that the “ideal” situation described in preclinical studies and in the first trials does not always match the results obtained in real life. Not all LSDs are the same.

A therapeutic approach is not necessarily good for all. As clinical experience expands and knowledge of the biology of LSDs increases, it is increasingly recognized that issues, shortcomings, and hurdles make each disease (or perhaps each category of patients) a moving target.

In general, the most impressive successes have been obtained in "visceral" LSDs, like Gaucher disease. In other disorders, where sanctuary tissues are involved, such as bone, eye or central nervous system, the results are erratic. The experience of many years, however, was not fruitless and provided directions for future improvements. New technologies, already in an advanced stage of development, will likely address the challenges that have emerged over the past decades, with second (or third)-generation enzymes, co-dosing with other therapeutic agents, optimization of therapeutic protocols, new routes of administration, and more efficient manufacturing procedures. In addition, the experience of many years and thousands of LSD patients treated with ERT may indicate new candidate diseases for this approach.

New therapeutic options for the treatment of LSDs are emerging. Gene therapy is probably the most compelling and promising. They may offer several advantages, including the potential for a single administration, reduced impact on patients' quality of life, improved targeting of critical tissues and organs, and prolonged therapeutic efficacy. Clinical experience in disorders such as metachromatic leukodystrophy has provided proof of concept for these benefits. In particular, hematopoietic stem cell gene therapy for metachromatic leukodystrophy has shown durable clinical efficacy and stabilization of neurological function[104,105]. Similarly, early-phase and preclinical studies in Pompe disease suggest that gene therapy may enable sustained and systemic enzyme expression, with improved tissue targeting through continuous secretion and receptor-mediated uptake in affected tissues[106].

Is there still room for ERT at the time of gene therapy? The answer is probably yes. Procedures to start gene therapy (particularly for *ex vivo* approaches) require time. For *in vivo* gene therapies, time is also required for full transgene expression in a patient. Thus, a pre-treatment with ERT may be advisable, particularly in early-onset and rapidly progressive disorders to stabilize patients and preserve organ function while awaiting the therapeutic effect of gene transfer[107,108].

Also, assessing the long-term effects of gene therapy will need a long time (as it happened for ERT). In this context, maintaining ERT as a complementary or fallback option may represent a safer strategy, particularly in cases where gene therapy proves insufficient or is not applicable to specific patient subgroups.

Artificial Intelligence-Assisted Technologies:

The authors used artificial intelligence (AI)-based language tools to improve the clarity and quality of the English language in some parts of the manuscript. No AI tools were used for the generation of scientific content. All authors take full responsibility for the content of the manuscript.

Authors' Contributions:

Simona Fecarotta contributed to the conception and design of the review, performed the literature review, and drafted the manuscript. Antonietta Tarallo contributed to the analysis and interpretation of the literature, critically revised the manuscript for important intellectual content, and contributed to drafting specific sections. Giancarlo Parenti conceived and supervised the review, contributed to the interpretation of the literature, critically revised the manuscript, and provided final approval of the version to be published.

All authors have contributed to the work, have read and approved the final version of the manuscript, and agree with its submission to this journal.

Conflict of Interest:

The authors have no conflict of interest to declare in relation to this article.

Ethics Approval and Informed Consent:

Not applicable due to the design of the study.

Funding:

No funding is declared for this article.

ORCID ID:

Simona Fecarotta: 0000-0002-8449-361X
Antonietta Tarallo: 0009-0004-6251-6333
Giancarlo Parenti: 0000-0002-6287-5748

REFERENCES

1. Kornfeld S. Lysosomal enzyme targeting. Biochem Soc Trans 1990; 18: 367-374.
2. Barton NW, Brady RO, Dambrosia JM, Di Bisceglie AM, Doppelt SH, Hill SC, Mankin HJ, Murray GJ, Parker RI, Argoff CE. Replacement therapy for inherited enzyme deficiency—macrophage-targeted glucocerebrosidase for Gaucher's disease. N Engl J Med 1991; 324: 1464-1470.
3. Brady RO. Enzyme replacement for lysosomal diseases. Annu Rev Med 2006; 57: 283-296.
4. van der Beek NAME, Theunissen MTM, van den Hout JMP, Pijnappel WWM, Schoser B, Laforêt P, Parenti G, van Doorn PA, van der Ploeg AT. Clinical insights in enzyme replacement therapy for metabolic storage disorders: lessons from Pompe disease. Lancet Neurol 2025; 24: 230-245.
5. Poswar FO, Vairo F, Burin M, Michelin-Tirelli K, Brusius-Facchin AC, Kubaski F, Souza CFM, Baldo G, Giugliani R. Lysosomal diseases: overview on current diagnosis and treatment. Genet Mol Biol 2019; 42: 165-177.
6. Desnick RJ, Schuchman EH. Enzyme replacement therapy for lysosomal diseases: lessons from 20 years of experience and remaining challenges. Annu Rev Genomics Hum Genet 2012; 13: 307-335.
7. Neufeld E. Replacement of genotype specific proteins in mucopolysaccharidosis enzyme therapy in genetic disease. In: Desnick RJ, Bernlohr RW, Krivit W, editors. Birth defects. Baltimore, MD: Williams & Wilkins; 1973. pp. 27-30.
8. Sly WS, Fischer HD. The phosphomannosyl recognition system for intracellular and intercellular transport of lysosomal enzymes. J Cell Biochem 1982; 18: 67-85.
9. Hickman S, Neufeld EF. A hypothesis for I-cell disease: defective hydrolases that do not enter lysosomes. Biochem Biophys Res Commun 1972; 49: 992-999.
10. Bohnsack RN, Misra SK, Liu J, Ishihara-Aoki M, Pereckas M, Aoki K, Ren G, Sharp JS, Dahms NM. Lysosomal enzyme binding to the cation-independent mannose 6-phosphate receptor is regulated allosterically by insulin-like growth factor 2. Sci Rep 2024; 14: 26875.
11. Olson LJ, Peterson FC, Castonguay A, Bohnsack RN, Kudo M, Gotschall RR, Canfield WM, Volkman BF, Dahms NM. Structural basis for recognition of phosphodiester-containing lysosomal enzymes by the cation-independent mannose 6-phosphate receptor. Proc Natl Acad Sci U S A 2010; 107: 12493-12498.
12. Grabowski GA, Leslie N, Wenstrup R. Enzyme therapy for Gaucher disease: the first 5 years. Blood Rev 1998; 12: 115-133.
13. Coutinho MF, Prata MJ, Alves S. A shortcut to the lysosome: the mannose-6-phosphate-independent pathway. Mol Genet Metab 2012; 107: 257-266.
14. Zhu Y, Li X, Kyazike J, Zhou Q, Thurberg BL, Raben N, Mattaliano RJ, Cheng SH. Conjugation of mannose 6-phosphate-containing oligosaccharides to acid alpha-glucosidase improves the clearance of glycogen in Pompe mice. J Biol Chem 2004; 279: 50336-50341.
15. Wenk J, Hille A, von Figura K. Quantitation of Mr 46000 and Mr 300000 mannose 6-phosphate receptors in human cells and tissues. Biochem Int 1991; 23: 723-731.
16. Koeberl DD, Luo X, Sun B, McVie-Wylie A, Dai J, Li S, Banugaria SG, Chen YT, Bali DS. Enhanced efficacy of enzyme replacement therapy in Pompe disease through mannose-6-phosphate receptor expression in skeletal muscle. Mol Genet Metab 2011; 103: 107-112.
17. Koeberl DD, Li S, Dai J, Thurberg BL, Bali D, Kishnani PS. β2 agonists enhance the efficacy of simultaneous enzyme replacement therapy in murine Pompe disease. Mol Genet Metab 2012; 105: 221-227.
18. Koeberl DD, Case LE, Desai A, Smith EC, Walters C, Han SO, Thurberg BL, Young SP, Bali D, Kishnani PS. Improved muscle function in a phase I/II clinical trial of albuterol in Pompe disease. Mol Genet Metab 2020; 129: 67-72.
19. Safety and Efficacy of Albuterol in Individuals With Late-onset Pompe Disease. https://clinicaltrials.gov. ClinicalTrials.gov Identifier: NCT01885936.
20. Safety and Efficacy of Clenbuterol in Individuals With Late-onset Pompe Disease and Receiving Enzyme Replacement Therapy. https://clinicaltrials.gov. ClinicalTrials.gov Identifier: NCT01942590.
21. Zhu Y, Jiang JL, Gumlaw NK, Zhang J, Bercury SD, Ziegler RJ, Lee K, Kudo M, Canfield WM, Edmunds T, Jiang C, Mattaliano RJ, Cheng SH. Glycoengineered acid alpha-glucosidase with improved efficacy in a mouse model of Pompe disease. Mol Ther 2009; 17: 954-963.
22. Schoser B, Roberts M, Byrne BJ, Sitaraman S, Jiang H, Laforêt P, Toscano A, Castelli J, Díaz-Manera J, Goldman M, van der Ploeg AT, Bratkovic D, Kuchipudi S, Mozaffar T, Kishnani PS. Safety and efficacy of cipaglucosidase alfa plus miglustat versus alglucosidase alfa in late-onset Pompe disease. Lancet Neurol 2021; 20: 1027-1037.
23. Hollak CE, vom Dahl S, Aerts JM, Belmatoug N, Bembi B, Cohen Y, Collin-Histed T, Deegan P, van Dussen L, Giraldo P, Mengel E, Michelakakis H, Manuel J, Hrebicek M, Parini R, Reinke J, di Rocco M, Pocovi M, Sa Miranda MC, Tylki-Szymanska A, Zimran A, Cox TM. Therapeutic measures in response to restricted supply of imiglucerase. Blood Cells Mol Dis 2010; 44: 41-47.
24. Wyatt K, Henley W, Anderson L, Anderson R, Nikolaou V, Stein K, Klinger L, Hughes D, Waldek S, Lachmann R, Mehta A, Vellodi A, Logan S. Effectiveness and cost-effectiveness of enzyme and substrate replacement therapies. Health Technol Assess 2012; 16: 1-543.
25. Katsigianni EI, Petrou P. A systematic review of economic evaluations of enzyme replacement therapy in lysosomal storage diseases. Cost Eff Resour Alloc 2022; 20: 51.
26. Heinrich R, Claus F, Schoenfelder T. Health care costs of home care enzyme replacement therapy for patients with lysosomal storage diseases in Germany. Orphanet J Rare Dis 2024; 19: 462.
27. Kaitin KI, DiMasi JA. Pharmaceutical innovation in the 21st century: new drug approvals in the first decade. Clin Pharmacol Ther 2011; 89: 183-188.
28. Pierce OM, McNair GR, He X, Kajiura H, Fujiyama K, Kermode AR. N-glycan structures and mannose-phosphorylation of plant recombinant human alpha-L-iduronidase. Plant Mol Biol 2017; 95: 593-606.
29. Uthailak N, Kajiura H, Misaki R, Fujiyama K. Production of recombinant β-glucocerebrosidase in glycoengineered Nicotiana benthamiana. J Biosci Bioeng 2022; 133: 481-488.
30. ELELYSO. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2012/022458s000ltr.pdf
31. ELFABRIO. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/appletter/2023/761161Orig1s000ltr.pdf

32. Zimran A, Wajnrajch M, Hernandez B, Pastores GM. Taliglucerase alfa: safety and efficacy across clinical studies in Gaucher disease. Orphanet J Rare Dis 2018; 13: 36.
33. Linhart A, Dostálová G, Nicholls K, West ML, Tøndel C, Jovanovic A, Giraldo P, Vujkovac B, Geberhiwot T, Brill-Almon E, Alon S, Chertkoff R, Rocco R, Hughes D. Pegunigalsidase alfa in Fabry disease: results from BRIDGE study. Orphanet J Rare Dis 2023; 18: 332.
34. Wallace EL, Goker-Alpan O, Wilcox WR, Holida M, Bernat J, Longo N, Linhart A, Hughes DA, Hopkin RJ, Tøndel C, Langeveld M, Giraldo P, Pisani A, Germain DP, Mehta A, Deegan PB, Molnar MJ, Ortiz D, Jovanovic A, Muriello M, Barshop BA, Kimonis V, Vujkovac B, Nowak A, Geberhiwot T, Kantola I, Knoll J, Waldek S, Nedd K, Karaa A, Brill-Almon E, Alon S, Chertkoff R, Rocco R, Sakov A, Warnock DG. Pegunigalsidase alfa versus agalsidase beta in Fabry disease: BALANCE study. J Med Genet 2024; 61: 520-530.
35. Holida M, Linhart A, Pisani A, Longo N, Eyskens F, Goker-Alpan O, Wallace E, Deegan P, Tøndel C, Feldt-Rasmussen U, Hughes D, Sakov A, Rocco R, Almon EB, Alon S, Chertkoff R, Warnock DG, Waldek S, Wilcox WR, Bernat JA. Pegunigalsidase alfa in Fabry disease: phase III study. J Inherit Metab Dis 2025; 48: e12795.
36. Hughes D, Gonzalez D, Maegawa G, Bernat JA, Holida M, Giraldo P, Atta MG, Chertkoff R, Alon S, Almon EB, Rocco R, Goker-Alpan O. Long-term efficacy of pegunigalsidase alfa in Fabry disease. Genet Med 2023; 25: 100968.
37. Shanmugaraj B, Ravi K, Baladevan K. Plant molecular farming for orphan drug production. Biotechnol Lett 2025; 47: 56.
38. Hopkin RJ, Byrne BJ, Dimachkie MM, Kishnani PS, Mozaffar T, Roberts M, Schoser B, van der Beek NAME, van der Ploeg AT, Wenninger S, Brudvig J, Fox B, Holdbrook F, Jain V, Johnson F, Zhang J, Parenti G. Miglustat as enzyme stabilizer for Pompe disease. Ther Adv Rare Dis 2026; 7: 26330040261425686.
39. Abasolo I, Seras-Franzoso J, Moltó-Abad M, Díaz-Riascos V, Corchero JL, Pintos-Morell G, Schwartz S Jr. Nanotechnology-based approaches for lysosomal storage disorders. Wiley Interdiscip Rev Nanomed Nanobiotechnol 2021; 13: e1684.
40. Del Grosso A, Parlanti G, Mezzena R, Cecchini M. Nanotechnology-driven enzyme replacement strategies. Adv Drug Deliv Rev 2022; 188: 114464.
41. Lenders M, Pollmann S, Terlinden M, Brand E. Anti-drug antibodies in Fabry disease and pegunigalsidase alfa. Mol Ther Methods Clin Dev 2022; 26: 323-330.
42. Schiffmann R, Goker-Alpan O, Holida M, Giraldo P, Barisoni L, Colvin RB, Jennette CJ, Maegawa G, Boyadjiev SA, Gonzalez D, Nicholls K, Tuffaha A, Atta MG, Rup B, Charney MR, Paz A, Szlaifer M, Alon S, Brill-Almon E, Chertkoff R, Hughes D. Pegunigalsidase alfa phase 1/2 trial. J Inherit Metab Dis 2019; 42: 534-544.
43. Su J, Sherman A, Doerfler PA, Byrne BJ, Herzog RW, Daniell H. Oral delivery of acid alpha-glucosidase epitopes in Pompe mice. Plant Biotechnol J 2015; 13: 1023-1032.
44. Hendrikse NM, Sandegren A, Andersson T, Blomqvist J, Makower Å, Possner D, Su C, Thalén N, Tjernberg A, Westermark U, Rockberg J, Svensson Gelius S, Syrén PO, Nordling E. Ancestral lysosomal enzymes for therapy. iScience 2021; 24: 102154.
45. Toscano A, Musumeci O, Sacchini M, Ravaglia S, Siciliano G, Fiumara A, Verrecchia E, Maione M, Gentile J, Fischetto R, Crescimanno G, Taurisano R, Sechi A, Gasperini S, Cianci V, Maggi L, Parini R, Lupica A, Scarpa M. Home-based ERT in Pompe and MPS I. Orphanet J Rare Dis 2023; 18: 338.
46. Ullman JC, Mellem KT, Xi Y, Ramanan V, Merritt H, Choy R, Gujral T, Young LEA, Blake K, Tep S, Homburger JR, O'Regan A, Ganesh S, Wong P, Satterfield TF, Lin B, Situ E, Yu C, Espanol B, Sarwaikar R, Fastman N, Tzitzilonis C, Lee P, Reiton D, Morton V, Santiago P, Won W, Powers H, Cummings BB, Hoek M, Graham RR, Chandriani SJ, Bainer R, DePaoli-Roach AA, Roach PJ, Hurley TD, Sun RC, Gentry MS, Sinz C, Dick RA, Noonberg SB, Beattie DT, Morgans DJ Jr, Green EM. Glycogen synthase inhibition in Pompe disease. Sci Transl Med 2024; 16: eadf1691.
47. Ullman JC, Dick RA, Linzner D, Minga T, Tep S, Satterfield TF, Xi Y, Beattie DT, Marmon T, Neutel JM, Chung B, Leeds JM, Noonberg SB, Green EM, Bernstein HS. First-in-human glycogen synthase inhibitor study in Pompe disease. Clin Pharmacol Ther 2024; 116: 1580-1592.
48. Lee NC, Chien YH, Wang CH, Wong SL, Peng SS, Tsai FJ, Hwu WL. Eliglustat combined with ERT in Gaucher disease type 3. Mol Genet Metab Rep 2022; 31: 100867.
49. Xu S, Lun Y, Frascella M, Garcia A, Soska R, Nair A, Ponery AS, Schilling A, Feng J, Tuske S, Valle MCD, Martina JA, Ralston E, Gotschall R, Valenzano KJ, Puertollano R, Do HV, Raben N, Khanna R. Next-generation ERT in Pompe disease. JCI Insight 2019; 4: e125358.
50. Friedman B, Vaddi K, Preston C, Mahon E, Cataldo JR, McPherson JM. Pharmacological properties of recombinant glucocerebrosidase. Blood 1999; 93: 2807-2816.
51. Parenti G, Andria G, Valenzano KJ. Pharmacological chaperone therapy for lysosomal storage disorders. Mol Ther 2015; 23: 1138-1148.
52. Porto C, Ferrara MC, Meli M, Acampora E, Avolio V, Rosa M, Cobucci-Ponzano B, Colombo G, Moracci M, Andria G, Parenti G. Allosteric chaperone enhancement of α-glucosidase. Mol Ther 2012; 20: 2201-2211.
53. Khanna R, Flanagan JJ, Feng J, Soska R, Frascella M, Pellegrino LJ, Lun Y, Guillen D, Lockhart DJ, Valenzano KJ. Pharmacological chaperone AT2220 in Pompe disease. PLoS One 2012; 7: e40776.
54. Do HV, Khanna R, Gotschall R. Challenges in treating Pompe disease. Ann Transl Med 2019; 7: 291.
55. Grabowski GA. Phenotype and treatment of Gaucher disease. Lancet 2008; 372: 1263-1271.
56. Vijay S, Brassier A, Ghosh A, Fecarotta S, Abel F, Marulkar S, Jones SA. Sebelipase alfa in lysosomal acid lipase deficiency. Orphanet J Rare Dis 2021; 16: 13.
57. Bashir A, Tiwari P, Duseja A. Enzyme replacement therapy in LAL-D: systematic review. Ther Adv Rare Dis 2021; 2: 26330040211026928.
58. Hannah WB, Ryan K, Pendyal S, Burrow TA, Harley SE, Cordell M, McCall CM, Mavis AM, Tan QK, Kishnani PS. Clinical insights from Wolman disease. Am J Med Genet A 2022; 188: 3364-3368.
59. Demaret T, Lacaille F, Wicker C, Arnoux JB, Bouchereau J, Belloche C, Gitiaux C, Grevent D, Broissand C, Adjaoud D, Abi Warde MT, Plantaz D, Bekri S, de Lonlay P, Brassier A. Sebelipase alfa in Wolman disease. Orphanet J Rare Dis 2021; 16: 507.
60. Camou F, Berger MG. Gaucher disease: state of the art and perspectives. J Intern Med 2025; 298: 155-172.
61. Ormazabal ME, Pavan E, Vaena E, Ferino D, Biasizzo J, Mucci JM, Serra F, Cifù A, Scarpa M, Rozenfeld PA, Dardis AE. Osteoclastogenic activation in Gaucher disease. Int J Mol Sci 2023; 24: 11204.

62. Grabowski GA, Kishnani PS, Alcalay RN, Prakalapakorn SG, Rosenbloom BE, Tuason DA, Weinreb NJ. Challenges in Gaucher disease: expert perspectives. Mol Genet Metab 2025; 145: 109074.
63. Kishnani PS, Corzo D, Nicolino M, Byrne B, Mandel H, Hwu WL, Leslie N, Levine J, Spencer C, McDonald M, Li J, Dumontier J, Halberthal M, Chien YH, Hopkin R, Vijayaraghavan S, Gruskin D, Bartholomew D, van der Ploeg A, Clancy JP, Parini R, Morin G, Beck M, De la Gastine GS, Jokic M, Thurberg B, Richards S, Bali D, Davison M, Worden MA, Chen YT, Wraith JE. Recombinant human acid alpha-glucosidase in infantile Pompe disease. Neurology 2007; 68: 99-109.
64. Chen M, Luo C, Zhao J, Devarajan G, Xu H. Immune regulation in the aging retina. Prog Retin Eye Res 2019; 69: 159-172.
65. Platt FM, d'Azzo A, Davidson BL, Neufeld EF, Tifft CJ. Lysosomal storage diseases. Nat Rev Dis Primers 2018; 4: 27.
66. Eisengart JB, Pierpont EI, Kaizer AM, Rudser KD, King KE, Pasquali M, Polgreen LE, Dickson PI, Le SQ, Miller WP, Tolar J, Orchard PJ, Lund TC. Intrathecal enzyme replacement for Hurler syndrome. Genet Med 2019; 21: 2552-2560.
67. Dickson PI, Chen AH. Intrathecal enzyme replacement therapy for mucopolysaccharidosis I. Curr Pharm Biotechnol 2011; 12: 946-955.
68. Muenzer J, Burton BK, Harmatz P, Gutiérrez-Solana LG, Ruiz-Garcia M, Jones SA, Guffon N, Inbar-Feigenberg M, Bratkovic D, Rust S, Hale M, Wu Y, Yee KS, Whiteman DAH, Alexanderian D. Intravenous and intrathecal idursulfase in MPS II. J Inherit Metab Dis 2025; 48: e12790.
69. Schulz A, Ajayi T, Specchio N, de Los Reyes E, Gissen P, Ballon D, Dyke JP, Cahan H, Slasor P, Jacoby D, Kohlschütter A. Intraventricular cerliponase alfa for CLN2 disease. N Engl J Med 2018; 378: 1898-1907.
70. Schulz A, Schwering C, Wibbeler E, Westermann LM, Hagenah L, Lezius S, Jha A, Hunt A, Slasor P, Reisewitz P, Nickel M. Real-world outcomes in CLN2 disease treated with cerliponase alfa. Front Neurol 2025; 16: 1516026.
71. Wijburg FA, Whitley CB, Muenzer J, Gasperini S, Del Toro M, Muschol N, Cleary M, Sevin C, Shapiro E, Alexanderian D. Intrathecal heparan-N-sulfatase in MPS IIIA. Mol Genet Metab 2021; 134: 175-181.
72. Okuyama T, Eto Y, Sakai N, Nakamura K, Yamamoto T, Yamaoka M, Ikeda T, So S, Tanizawa K, Sonoda H, Sato Y. Pabinafusp alfa for neuronopathic MPS II. Mol Ther 2021; 29: 671-679.
73. Yamamoto R, Kawashima S. Pharmacological properties of pabinafusp alfa in MPS II. Nihon Yakurigaku Zasshi 2022; 157: 62-75.
74. Giugliani R, Giugliani L, de Oliveira Poswar F, Donis KC, Corte AD, Schmidt M, Boado RJ, Nestrasil I, Nguyen C, Chen S, Pardridge WM. Valanafusp alpha in MPS I. Orphanet J Rare Dis 2018; 13: 110.
75. Muenzer J, Burton BK, Harmatz P, Rajan DS, Jones SA, van den Hout JMP, Mitchell JJ, Bhalla A, Engmann NJ, Zubizarreta I, Dong W, Model F, Watts RJ, Troyer MD, Chin PS, Ho C. Brain-penetrant enzyme therapy for MPS II. N Engl J Med 2026; 394: 39-50.
76. Leal AF, Inci OK, Seyrantepe V, Rintz E, Celik B, Ago Y, León D, Suarez DA, Alméciga-Díaz CJ, Tomatsu S. Molecular Trojan horses for lysosomal storage diseases. Mol Genet Metab 2023; 140: 107648.
77. Del Grosso A, Galliani M, Angella L, Santi M, Tonazzini I, Parlanti G, Signore G, Cecchini M. Brain-targeted enzyme-loaded nanoparticles in Krabbe disease. Sci Adv 2019; 5: eaax7462.
78. Del Grosso A, Carpi S, Colagiorgio L, De Sarlo M, Gagliardi M, Cecchini M. Cellular effects of GALC dosing in Krabbe disease. Adv Biol (Weinh) 2025; 9: e00147.
79. Parenti G, Medina DL, Ballabio A. The evolving view of lysosomal storage diseases. EMBO Mol Med 2021; 13: e12836.
80. Lieberman AP, Puertollano R, Raben N, Slaugenhaupt S, Walkley SU, Ballabio A. Autophagy in lysosomal storage disorders. Autophagy 2012; 8: 719-730.
81. Fukuda T, Ahearn M, Roberts A, Mattaliano RJ, Zaal K, Ralston E, Plotz PH, Raben N. Autophagy and enzyme mistargeting in Pompe disease. Mol Ther 2006; 14: 831-839.
82. Spampanato C, Feeney E, Li L, Cardone M, Lim JA, Annunziata F, Zare H, Polishchuk R, Puertollano R, Parenti G, Ballabio A, Raben N. TFEB as therapeutic target in Pompe disease. EMBO Mol Med 2013; 5: 691-706.
83. Tarallo A, Damiano C, Strollo S, Minopoli N, Indrieri A, Polishchuk E, Zappa F, Nusco E, Fecarotta S, Porto C, Coletta M, Iacono R, Moracci M, Polishchuk R, Medina DL, Imbimbo P, Monti DM, De Matteis MA, Parenti G. Oxidative stress correction enhances ERT in Pompe disease. EMBO Mol Med 2021; 13: e14434.
84. Laugwitz L, Shenker A, Sluys EF, Pintat S, Whiteman D, Chanson C. Newborn screening for metachromatic leukodystrophy. Int J Neonatal Screen 2025; 11: 103.
85. Desai AK, Rodriguez-Rassi E, Parikh S, Russo RS, Kronn D, Hamm JA, Chang IJ, Ortiz D, McPheron M, Lydigsen H, DeArmey S, Young SP, Kishnani PS. Early ERT in infantile Pompe disease via newborn screening. Genet Med Open 2025; 4: 103478.
86. Malvagia S, Bettiol A, Porcaro M, Mura M, Funghini S, Ombrone D, Forni G, Scolamiero E, Coppi F, Damiano R, Cereda C, Simonetti S, Lonetti A, Daniotti M, Caciotti A, Morrone A, Calbi V, Fumagalli F, Aiuti A, Procopio E, Guerrini R, la Marca G. Newborn screening for metachromatic leukodystrophy in Tuscany. Int J Neonatal Screen 2025; 11: 30.
87. Gragnaniello V, Pijnappel PWWM, Burlina AP, In 't Groen SLM, Gueraldi D, Cazzorla C, Maines E, Polo G, Salviati L, Di Salvo G, Burlina AB. Newborn screening for Pompe disease in Italy. Mol Genet Metab Rep 2022; 33: 100929.
88. Kishnani PS, Dickson PI, Muldowney L, Lee JJ, Rosenberg A, Abichandani R, Bluestone JA, Burton BK, Dewey M, Freitas A, Gavin D, Griebel D, Hogan M, Holland S, Tanpaiboon P, Turka LA, Utz JJ, Wang YM, Whitley CB, Kazi ZB, Pariser AR. Immune responses to enzyme replacement therapies. Mol Genet Metab 2016; 117: 66-83.
89. Gragnaniello V, Deodato F, Gasperini S, Donati MA, Canessa C, Fecarotta S, Pascarella A, Spadaro G, Concolino D, Burlina A, Parenti G, Strisciuglio P, Fiumara A, Casa RD. Immune responses to alglucosidase in Pompe disease. Ital J Pediatr 2022; 48: 41.
90. Desai AK, Rosenberg AS, Kishnani PS. IVIG timing and immune tolerance in Pompe disease. Clin Immunol 2020; 219: 108541.
91. İnci A, Ezgü FS, Tümer L. Immune tolerance induction in enzyme replacement therapy. Paediatr Drugs 2024; 26: 287-308.
92. Doerfler PA, Nayak S, Corti M, Morel L, Herzog RW, Byrne BJ. Immune tolerance approaches in Pompe disease. Mol Ther Methods Clin Dev 2016; 3: 15053.
93. Tseng YH, Lin HP, Lin SY, Chen BM, Vo TNN, Yang SH, Lin YC, Prijovic Z, Czosseck A, Leu YL, Roffler SR. Engineering stable immunoenzymes. J Control Release 2024; 369: 179-198.
94. Schaible P. Modifying enzyme replacement therapy. J Cell Mol Med 2023; 27: 165-173.
95. Doerfler PA, Todd AG, Clément N, Falk DJ, Nayak S, Herzog RW, Byrne BJ. AAV9 vectors for immune tolerance in Pompe disease. Hum Gene Ther 2016; 27: 43-59.

96. Rajkumar V, Dumpa V. Lysosomal storage disease. StatPearls 2023.
97. Schulz A, Specchio N, de Los Reyes E, Gissen P, Nickel M, Trivisano M, Aylward SC, Chakrapani A, Schwering C, Wibbeler E, Westermann LM, Ballon DJ, Dyke JP, Cherukuri A, Bondade S, Slasor P, Cohen Pfeffer J. Cerliponase alfa in CLN2 disease. Lancet Neurol 2024; 23: 60-70.
98. Kishnani PS, Shohet S, Raza S, Hummel N, Castelli JP, Sitaraman Das S, Jiang H, Kopiec A, Keyzor I, Hahn A. PROMIS outcomes in Pompe disease. J Patient Rep Outcomes 2024; 8: 13.
99. Schiffmann R. Biomarkers in lysosomal storage diseases. J Inherit Metab Dis 2025; 48: e70034.
100. Rossi A, Malvagia S, la Marca G, Parenti G, Brunetti-Pierri N. Biomarkers for gene therapy in lysosomal diseases. Mol Ther 2024; 32: 2930-2938.
101. Pandey MK. Neuroinflammation in lysosomal storage diseases. Biomedicines 2023; 11: 1067.
102. Weinhofer I, Rommer P, Berger J. Blood biomarkers in lysosomal diseases. J Inherit Metab Dis 2025; 48: e70032.
103. Tarallo A, Carissimo A, Gatto F, Nusco E, Toscano A, Musumeci O, Coletta M, Karali M, Acampora E, Damiano C, Minopoli N, Fecarotta S, Della Casa R, Mongini T, Vercelli L, Santoro L, Ruggiero L, Deodato F, Taurisano R, Bembi B, Dardis A, Banfi S, Pijnappel WWP, van der Ploeg AT, Parenti G. microRNAs as biomarkers in Pompe disease. Genet Med 2019; 21: 591-600.
104. Biffi A, Montini E, Lorioli L, Cesani M, Fumagalli F, Plati T, Baldoli C, Martino S, Calabria A, Canale S, Benedicenti F, Vallanti G, Biasco L, Leo S, Kabbara N, Zanetti G, Rizzo WB, Mehta NA, Cicalese MP, Casiraghi M, Boelens JJ, Del Carro U, Dow DJ, Schmidt M, Assanelli A, Neduva V, Di Serio C, Stupka E, Gardner J, von Kalle C, Bordignon C, Ciceri F, Rovelli A, Roncarolo MG, Aiuti A, Sessa M, Naldini L. Lentiviral gene therapy in metachromatic leukodystrophy. Science 2013; 341: 1233158.
105. Tucci F, Scaramuzza S, Aiuti A, Mortellaro A. Ex vivo hematopoietic stem cell gene therapy. Mol Ther 2021; 29: 489-504.
106. Kishnani PS, Koeberl DD. Liver depot gene therapy for Pompe disease. Ann Transl Med 2019; 7: 288.
107. Ma X, Zhuang L, Ma W, Li J, Mao Y, Wang J, Wang X, Xu J, Yu S, Gu R, Wang Y, Li Z, Jiang X, Zhang S, He F, Yang X, Zhu L, Zhang S, Zhang Y, Li T, Li Q, Wu Z, Zhang C, Zhang Y, Dong X, Xiong H, Wu X, Feng Z. AAV9 gene therapy for infantile-onset Pompe disease. N Engl J Med 2025; 392: 2438-2446.
108. Parenti G, Brunetti-Pierri N. Gene therapy for infantile-onset Pompe disease. N Engl J Med 2025; 392: 2477-2478.

JIM
JIM 2026; 3 (3): e1209
DOI: 10.61012/JIM_202608_1209
Category: *Original Article*

COMMUNICATION CHALLENGES IN NEWBORN SCREENING: WHAT THE PUBLIC KNOWS (AND DOESN'T)

M. COSTANZO[1,2], S. BIANCO[1,2], M. CATERINO[1,2], A. CEVENINI[1,2], M. COZZOLINO[2,3], C. CAZZORLA[4], M. BANI[5], M. RUOPPOLO[1,2]

• • •

[1]Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy
[2]CEINGE–Biotecnologie Avanzate Franco Salvatore, Naples, Italy
[3]European School of Molecular Medicine, University of Milan, Milan, Italy
[4]Division of Inherited Metabolic Diseases, Department of Diagnostic Services, University Hospital of Padua, Padua, Italy
[5]Department of Medicine and Surgery, University of Milan-Bicocca, Milan, Italy

Corresponding Author
Margherita Ruoppolo, MD, Ph.D; e-mail: margherita.ruoppolo@unina.it

***ABSTRACT* – Objective:** Extended newborn screening (ENBS) is a worthwhile public health program for the early detection of inherited metabolic disorders, facilitating timely diagnosis and treatments that have significantly decreased the morbidity and mortality associated with these conditions. Despite its importance, public awareness of ENBS remains limited. All families deserve access to clear, accurate, and relevant information to help them navigate the newborn screening system. To assess whether the level of knowledge and perceptions of NBS among the general population are appropriate, a public engagement initiative was conducted.

Subjects and Methods: A public engagement initiative was realized in the city of Naples (Italy) to involve passers-by in the completion of a survey titled "Extended neonatal screening – how much do you know?". The questionnaire was distributed as a paper-based or online form, including 15 items focused on demographics and knowledge of ENBS, with the option to share personal experiences and feedback on the dissemination activity.

Results: We interacted with approximately 600 people, and half completed the questionnaire. The study population was heterogeneous in terms of sex, age, and educational background, with around half belonging to the biomedical/healthcare field. Globally, only 33% of respondents were aware of ENBS. Despite this, appreciation of the outreach activity and the materials provided was high regardless of the knowledge attained.

Conclusions: The data collected in this study reveal a concerning lack of awareness, particularly among those who have already had children. These findings indicate that sporadic recognition or outreach efforts, although beneficial, are insufficient on their own. More sustained and structured schemes are required within clinical settings, where healthcare personnel should be deliberately trained and supported to enhance education and communication with parents.

Keywords: Expanded newborn screening, Metabolic disorders, NBS, Survey, Knowledge gap, Communication challenge.

INTRODUCTION

Newborn screening (NBS) is one of the most effective public health strategies for the early identification of inherited genetic and metabolic disorders[1-4]. Empowering timely diagnosis and early treatments can help prevent the progression of such severe diseases, as leaving them untreated often leads to brain damage, multi-organ failure, or even death[5]. In fact, NBS has significantly ameliorated the prognosis and quality of life through an organized approach that coordinates the monitoring and care of patients applying significant

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

surveillance activities[6]. The application of NBS was pioneered by Robert Guthrie, who in 1959 developed a bacterial inhibition assay to detect elevated phenylalanine levels using dried blood spots (DBS). This approach marked the beginning of the first NBS program in the United States, providing a successful proof of concept for early diagnosis of phenylketonuria (PKU). This allowed prompt treatment of pre-symptomatic neonates with the possibility to prevent permanent damages[7]. The panel of pathologies examined through NBS has gradually increased over the years. In the late 1990s, the introduction of tandem mass spectrometry (MS/MS) into NBS laboratories revolutionized the clinical paths, allowing comprehensive, fast, sensitive and accurate quantification of metabolite markers in body fluids[8]. MS/MS enabled the screening for aminoacidopathies (including PKU), organic acidemias, fatty acid oxidation defects, and urea cycle defects, leading to the modern concepts of expanded newborn screening (ENBS)[9-11]. The history of neonatal screening in Italy began with Legislative Decree No. 104 of 1992, which introduced the screening for three diseases: congenital hypothyroidism (CH), cystic fibrosis (CF), and PKU. Subsequently, thanks to scientific progress and over 20 years of practice, it was possible to extend neonatal screening activities to many more inherited metabolic diseases. The panel of diseases subjected to ENBS was established in 2016 (Law 167 of 19 August 2016 and subsequent Ministerial Decree of 13 October 2016) and currently includes over 40 metabolic disorders. Then, following the Prime Ministerial Decree of 12 January 2017, ENBS became part of the Essential Assistance Levels (LEA) and was made mandatory across the entire Italian territory. The ENBS program at the regional or interregional level is a system divided into the following main functional structures, namely: neonatal screening laboratory, laboratory for biochemical diagnostic confirmation, laboratory for genetic diagnostic confirmation, regional reference clinical centers, and regional coordination center of the screening system. The structure and functions of the current Italian organization have been described by Ruoppolo et al[12]. Effective implementation of ENBS not only requires a robust technical infrastructure but also relies on informed parental engagement, especially during critical phases such as post-screening communication and follow-up. Despite its proven clinical and societal benefits, awareness and understanding of NBS among the general population remain limited, particularly outside biomedical contexts. For this reason, increasing public knowledge is crucial.

The psychosocial impact of receiving unexpected newborn screening results or unsolicited genetic information represents a major concern in ENBS, especially in the case of positive results. There is an additional consideration that screening tests can generate false-positive results, inducing strong psychosocial effects on parents and families. However, research suggests that parental stress and anxiety can be significantly mitigated through better parental education about ENBS, improvement of the communication schemes adopted, personalized genetic counseling, and proactive developmental guidance[13,14]. Positive tests, confirmed diagnoses, false positives, or variants of uncertain significance (VUS) can cause significant psychosocial distress, anxiety, confusion, and feelings of helplessness[13,15]. In cases of false positives or VUS, this distress is often intensified by uncertainty, as parents struggle to interpret ambiguous health information and worry about potential future risks[15]. Studies[14,16] have shown that such responses can lead to increased parental stress, altered perceptions of child vulnerability and, in some cases, long-term effects on the parent-child relationship. Notably, even when follow-up testing confirms the absence of disease, lingering concerns and altered parenting behaviors may persist[17,18]. To reduce the psychosocial burden, tailored interventions have been proposed, including early and empathetic communication from trained healthcare professionals, ideally following structured counseling protocols that anticipate emotional reactions[19]. Providing clear, balanced, and jargon-free explanations of screening results, supported by visual or written materials, can enhance understanding and reduce fear[20]. Incorporating psychosocial support by genetic counselors or psychologists and ensuring continuity of care through follow-up consultations may alleviate parental anxiety[21]. Moreover, it is essential to determine the appropriate timing for delivering information, as educating parents on this subject can enhance information retention and contribute to better care for both children and parents[22]. Improved parental education about newborn screening, effective communication strategies, individualized genetic counseling, and anticipatory guidance can mitigate unexpected communications associated with diseases with a non-acute neonatal onset, rendering diagnosis acceptance for families less difficult[13,23].

This work aimed to assess public awareness and attitudes toward ENBS, exploring informational needs and opinions of people from Naples, Italy. By collecting survey responses, we sought to identify knowledge gaps and opportunities to improve communication strategies related to neonatal screening practices. At present, in fact, the implementation of effective educational initiatives is hindered by fragmented communication practices, restricted counseling time during perinatal care, limited resources, and a siloed newborn screening system, all of which create barriers to effective educational experience. Engaging families, especially those from historically underserved communities, is essential to increase knowledge, build confidence, and ultimately improve outcomes for children and their families[24].

SUBJECTS AND METHODS

To evaluate the degree of public knowledge about ENBS and awareness of its relevance within the national healthcare system, a survey was delivered among passers-by in the most frequented areas of the city of Naples (Campania, Italy) during the Christmas days (22-24 December 2025). The interviews were conducted through questionnaire distribution, after which informational material was provided to participants to offer detailed screening-related information. Individuals were invited to participate in the survey regardless of their level of expertise, ensuring the inclusion of both informed and non-specialist respondents. Participants who consented to take part in this initiative were provided with a questionnaire that could be completed either on paper or online, according to their preference. The survey instrument was developed by the authors and comprised 15 questions aimed at collecting demographic information as well as assessing respondents' knowledge and awareness of ENBS. The full list of questions proposed during the survey is reported below, while their relative multiple-choice answers are reported in the Supplementary Material.

Questionnaire on the "Extended neonatal screening – how much do you know?"

1. What is your sex?
2. How old are you?
3. What is your level of education?
4. Do you work in the biomedical/healthcare field?
5. Did you know about neonatal screening?
6. Have you ever heard of any of the diseases included in neonatal screening?
7. Do you have children?
8. At the time of childbirth, have you received information about neonatal screening?
9. Do you have a family member/acquaintance with a hereditary metabolic disease?
10. If yes, was it identified by neonatal screening or subsequently?
11. How do you rate this outreach activity?
12. Indicate your appreciation of the dissemination material.
13. What did you learn during this outreach activity?
14. Do you have any questions about extended neonatal screening?
15. If you like, tell us about your experience with extended newborn screening.

Questionnaires were collected directly by the authors, and all data were handled in accordance with ethical standards, ensuring anonymity and confidentiality. All subjects provided written informed consent for inclusion in the study and anonymous treatment of the data. Indeed, no personally identifiable information was recorded, stored, or processed at any stage of the study. The contribution to this survey was voluntary, and participants were adequately informed about the initiative prior to questionnaire administration. After data collection, all responses were aggregated into a single matrix dataset. Questionnaires were screened for completeness and internal consistency. Incomplete questionnaires, defined as those with more than 20% missing responses, were excluded from further analysis. For questionnaires with a limited number of missing items, available data were retained without imputation, given the descriptive nature of the study, and labeled as NS (Not Specified).

Statistical Analysis

The finalized dataset was subsequently analyzed using descriptive statistical methods[25], including frequency distributions, percentages, and summary measures, to characterize respondents' demographic features and levels of awareness related to ENBS. Statistical analyses and graphical representations were performed using GraphPad Prism software version 10 (San Diego, CA, USA)[26].

RESULTS

A public engagement initiative was carried out in the city of Naples with the aim of assessing the level of knowledge of ENBS within the general population and identifying potential communication challenges affecting its understanding and perception. Through the administration of a structured questionnaire to passers-by, the initiative allowed the collection of quantitative and qualitative data on participants' demographic characteristics, awareness and parental knowledge of neonatal screening practices, and perceptions of the outreach activity itself.

Demographic Characteristics of the Survey Participants

The questionnaires were compiled by 298 people, representing the 51% of the passers-by (total of 583) (Figure 1A), and all related data on their demographic characteristics are detailed in Table 1. People were invited to complete the survey through a paper modality or *via* an online form, with the majority of them choosing the paper modality (62%) (Figure 1B). Globally, the mean response rate to the full questionnaire was 99%, with only sporadic missing responses (1%), mainly involving demographic information. The population that responded to the survey showed a prevalence of female subjects (66%) (Figure 1C). However, it was well distributed in terms of age ranges, which were divided into three groups: young people (under 25 years old), middle-aged adults (between 25 and 40 years old), and older adults

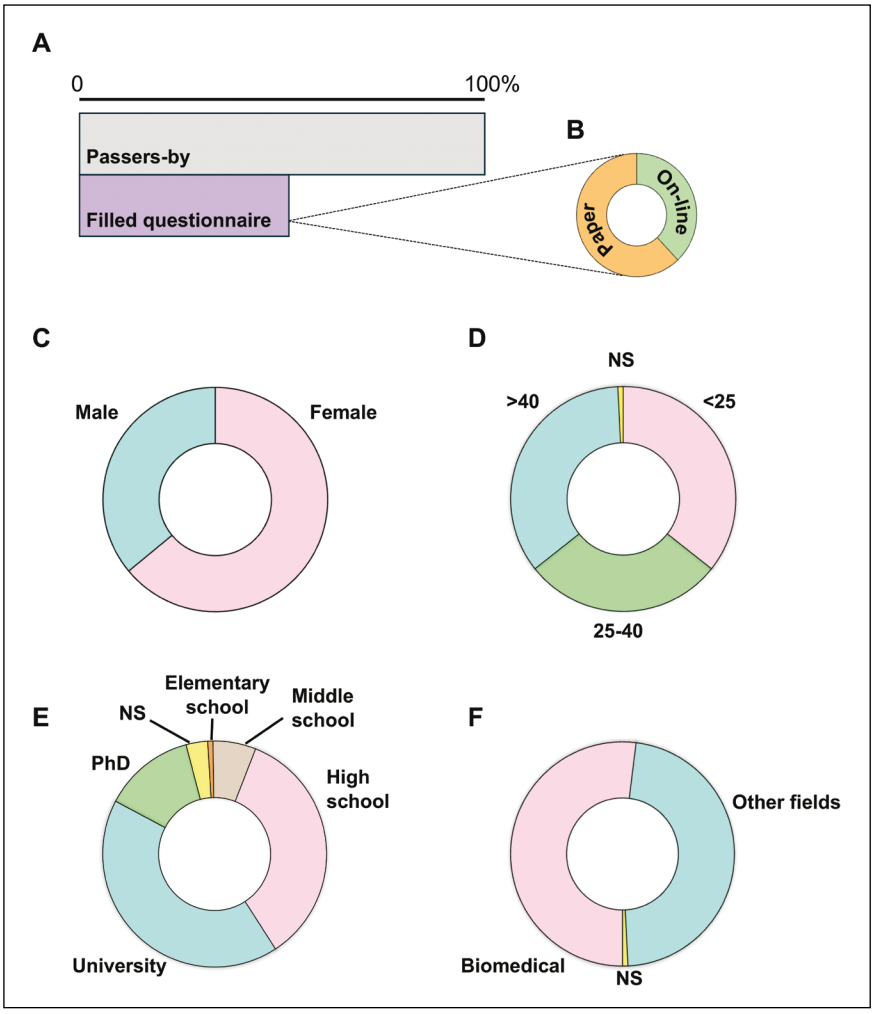


Figure 1. Descriptive analysis of demographic and educational characteristics of the population of respondents. **A**, Distribution of survey respondents over the total of passers-by. **B**, Chosen modality of compilation of the survey. Distribution of the population of respondents according to (**C**) sex, (**D**) age range, (**E**) level of education, and (**F**) professional occupation. NS=Not Specified.

Table 1. Demographic characteristics of the survey participants.

Sex	Males (34%)	Females (66%)				
Age	<25 (36%)	25-40 (28%)	>40 (35%)	NS (1%)		
Education	PhD/ Specialization (12%)	University degree (42%)	High school (35%)	Middle school (6%)	Elementary school (2%)	NS (3%)
Work	Biomedical field (47%)	Others (52%)	NS (1%)			

NS=Not Specified.

(over 40 years old). The 1% did not declare their age (NS, not specified) (Figure 1D). A heterogeneous distribution was observed in terms of educational background: 42% of the participants held a university degree and only 12% had a higher title (PhD or specialization), 35% completed high school, 6% concluded middle school, and 2% of the participants stopped their studies during elementary education. Finally, 3% of respondents did not indicate their level of education (NS) (Figure 1E). The professional occupation of respondents also showed a good level of diversity. Nearly half of them (47%) reported working in the biomedical or healthcare field, while 52% indicated that they did not work in this sector (Figure 1F).

Awareness of Newborn Screening in the General Population and Parental Gaps

The data collected from responses revealed a significant lack of awareness regarding ENBS among the general population. In our survey, in fact, to the question "Did you know about neonatal screening?" 62% of respondents replied "No", demonstrating that they were unaware of ENBS procedures. An additional 5% did not respond to this question (NS), resulting in a minority of individuals (33%) who reported understanding this crucial health initiative (Figure 2A). Considering the heterogeneity of passers-by with relation to their educational and professional background, this lack of knowledge was notably higher among individuals without professional experience in biomedical/healthcare fields (80% of "No" respondents) or among subjects with no high academic qualifications (50% of "No" respondents) (Figure 2B). Finally, people without both a university degree and a biomedical occupation corresponded to 43.3% of "No" respondents (Figure 2B), indicating a positive correlation between educational/professional background and knowledge of ENBS. Among the total interviewees, survey data reveal that 36% of respondents declared to have children (Figure 2C). However, within this group, it emerged that more than half of these parents were unaware that neonatal screening had been performed on their child at birth (Figure 2D).

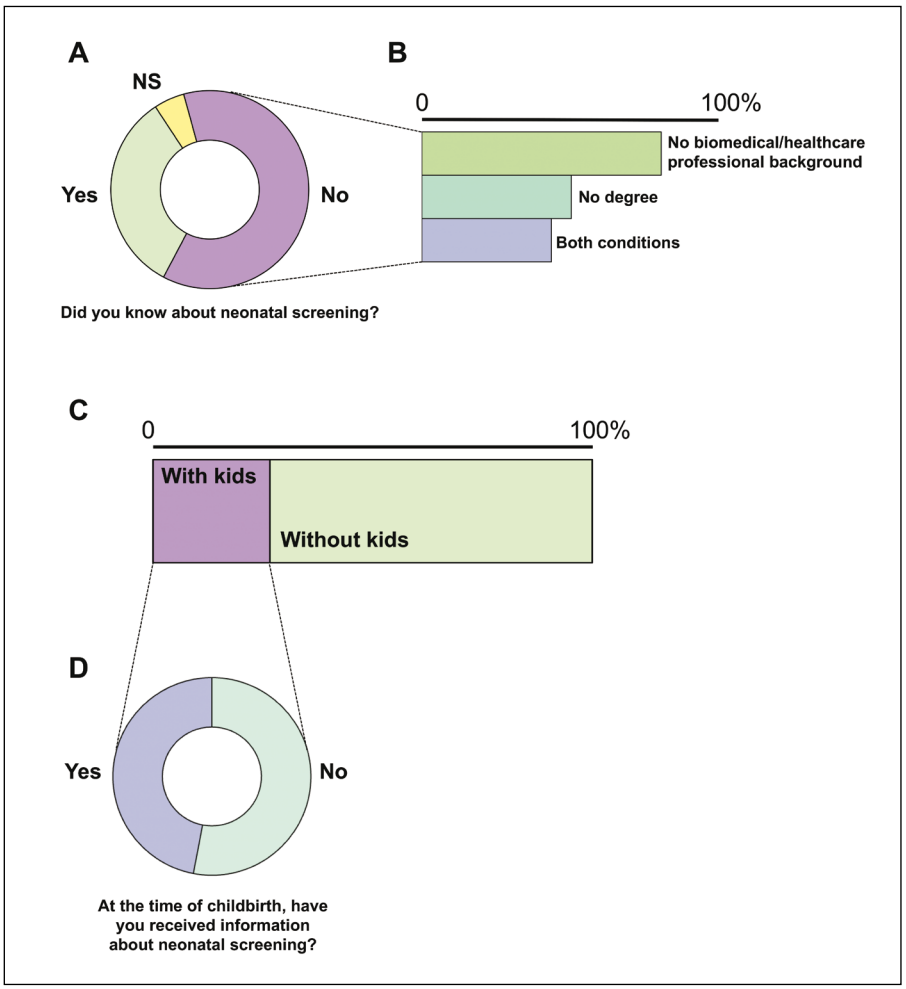


Figure 2. Distribution of respondents to the main questions about ENBS knowledge. **A**, Distribution of respondents to the question "Did you know about neonatal screening?" and (**B**) percentage of "No" respondents according to their professional and educational background. **C**, Respondents reported having or not having children. **D**, Parent awareness about ENBS testing performed at the time of childbirth. NS=Not Specified.

Participant Feedback and Emerging Questions

A specific need to gain more insight into ENBS emerged from questions 11 and 12 regarding appreciation of the outreach activity, and from some emblematic replies collected in the final open-ended questions 13-15. Public feedback gathered during our health outreach activity is summarized in Figure 3. When asked to rate the initiative, most participants gave the highest rating on a 1-to-3 scale, with the vast majority choosing "3" (Figure 3A). Similarly, when asked about the dissemination material provided, most participants selected "5" on a 1-to-5 scale, indicating high satisfaction (Figure 3B). These results suggest that the initiative was well-received and that the materials used were effective at engaging the public. Open-ended responses provided valuable insight into what participants gained from the outreach activity and what information they still want to get. Some subjects discovered the possibility of diagnosis at birth and prevention in early childhood. A few of them showed interest in the coverage of disorders that can be screened, emphasizing how this information had changed their perception of newborn care. Other participants asked for clarification on which tests are mandatory and automatically performed at birth, *vs.* those that are optional or require parental consent.

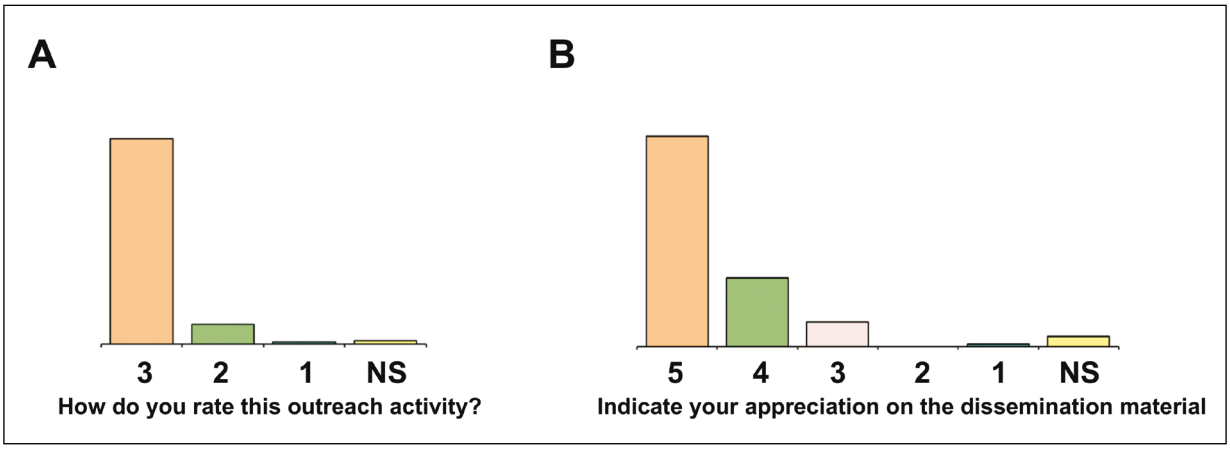


Figure 3. Appreciation rating of participants to the survey about (**A**) the proposed health outreach activity (1-to-3 scale) and (**B**) the dissemination material distributed (1-to-5 scale). NS=Not Specified.

DISCUSSION

The results presented below provide an overview of the surveyed population, highlighting key gaps in awareness, as well as emerging needs for improved information and communication strategies related to ENBS.

This study provides real-world evidence of a substantial gap in public awareness. Despite the high participation rate, around two-thirds of respondents reported no prior knowledge of neonatal screening. While the demographic profile of the surveyed population was relatively heterogeneous, our findings require attention as a considerable proportion was represented by individuals with higher education and professional experience in the biomedical or healthcare field. This aspect, likely related to the voluntary nature of participation and the interest in health-related topics, may have introduced a selection bias and should therefore be considered a limitation of the study. Nevertheless, a substantial lack of awareness regarding ENBS was still observed, suggesting that knowledge within the general population may be even lower than that detected in our cohort.

Literature and historical experience in public health, particularly in preventive screening, clearly show how socio-economic and educational inequalities can influence the understanding and accessibility of health programs, including newborn screening[24,27]. Accordingly, our results corroborate these achievements, highlighting that individuals with lower educational levels or from a different professional occupation are less informed. This underscores the urgent need for more effective and accessible educational strategies to raise awareness about the value of newborn screening and its impact on child health[28,29].

A concerning result that emerged from this survey is the limited awareness among parents of the newborn screening procedures performed on their own children. This gap is alarming as parents' understanding can significantly influence their engagement in follow-up care and interventions necessary for their child's health[30]. Furthermore, adequate information at the time of childbirth or before it may help the communication with the clinical staff in case of recall or notification of a positivity to one of the diseases included in the program. Parents who are insufficiently informed may experience confusion or anxiety in the event of a recall or a positive screening result, potentially compromising trust in healthcare providers and adherence to follow-up pathways. Informed parental engagement is a key component of effective screening programs, as it facilitates timely communications and continuity of care, guaranteeing full engagement in child healthcare[31].

Notwithstanding, an encouraging finding was represented by the high level of appreciation expressed by participants toward the outreach initiative and the dissemination of materials. Participants expressed the need to be better informed, especially before or during hospitalization for childbirth, to be more conscious regarding the health of their newborns or potential risks. The positive feedback indicates that public engagement activities can be effective tools for increasing interest, awareness, and understanding of newborn screening[32]. These reflections highlight the value of knowledge spreading, but also underscore the necessity for clear, accessible, and proactive communication within healthcare settings.

CONCLUSIONS

This study collected data from surveys administered to passers-by in the most frequented areas of the city of Naples with reference to ENBS to quantify the degree of public knowledge. Our initiative included several valuable components: the presence of experts available to answer questions, brochures on ENBS (including lysosomal storage disorders), and informational flyers. These elements offered a comprehensive overview and allowed for personal engagement with the public. The responses to the questionnaire "Extended neonatal screening – how much do you know?" made it clear that more structured communication strategies are needed. Informing the public should not rely solely on events or printed materials. Training on ENBS also needs to be strengthened within healthcare facilities. Nurses and clinical staff involved in the neonatal screening process need to be adequately trained to provide parents with clear and early guidance[33]. Given the crucial importance of these tests for a child's health, early communication must be both informative and empathetic, addressing parents' questions and emotional needs from the very first interaction. This approach is instrumental in preparing families for potential follow-up steps, such as recall notifications or additional testing, to minimize distress and reduce possible harm to children and their families[34]. Indeed, parental reactions are closely linked to the quality of communication they receive throughout the screening process. While the outreach activity was successful in terms of public satisfaction, the persistent knowledge gap highlights the need for ongoing, more integrated communication strategies to guide parents and families through subsequent steps with confidence and clarity.

Acknowledgements:
We thank all participants who took part in the survey for their time and contribution.

Artificial Intelligence-Assisted Technologies:
No artificial intelligence-assisted technologies were used in the production of this article.

Authors' Contributions:
Study conception and design: M. Costanzo, S. Bianco, M. Ruoppolo; collection and interpretation of data: all; statistical analysis: M. Costanzo, S. Bianco; manuscript drafting: all; manuscript editing: all; approval to submit: all.

Availability of Data and Material:
All data generated or analyzed in this study are included in the main article and its Supplementary Material.

Conflict of Interest:
The authors have no conflicts of interest to disclose.

Ethics Approval and Informed Consent:

This study was approved by the CERSUB Ethical Committee of the University of Naples Federico II (project PG/2025/0165206 of the 2nd December 2025). All subjects provided written informed consent for inclusion in the study and treatment of the data.

Funding:

No funding was received for this study.

ORCID ID:

Michele Costanzo: 0000-0002-6386-2009
Sabrina Bianco: 0009-0005-1848-1054
Marianna Caterino: 0000-0001-6035-8213
Armando Cevenini: 0000-0002-1858-2833
Marica Cozzolino: 0009-0005-2628-6005
Chiara Cazzorla: 0000-0002-8081-3997
Marco Bani: 0000-0002-6500-6513
Margherita Ruoppolo: 0000-0002-7364-0602

REFERENCES

1. Lemonde H, Cleary M, Chakrapani A. Newborn screening for inborn errors of metabolism. Paediatr Child Health 2015; 25: 103-107.
2. Sandlers Y. The future perspective: metabolomics in laboratory medicine for inborn errors of metabolism. Transl Res 2017; 189: 65-75.
3. Lund AM, Hougaard DM, Simonsen H, Andresen BS, Christensen M, Dunø M, Skogstrand K, Olsen RK, Jensen UG, Cohen A, Larsen N, Saugmann-Jensen P, Gregersen N, Brandt NJ, Christensen E, Skovby F, Nørgaard-Pedersen B. Biochemical screening of 504,049 newborns in Denmark, the Faroe Islands and Greenland — Experience and development of a routine program for expanded newborn screening. Mol Genet Metab 2012; 107: 281-293.
4. Selim LA, Hassan SAH, Salem F, Orabi A, Hassan FA, El-Mougy F, Mahmoud IG, El-Badawy A, Girgis MY, Elmonem MA, Mehaney D. Selective screening for inborn errors of metabolism by tandem mass spectrometry in Egyptian children: A 5year report. Clin Biochem 2014; 47: 823-828.
5. Gaugler S, Rykl J, Wegner I, von Däniken T, Fingerhut R, Schlotterbeck G. Extended and Fully Automated Newborn Screening Method for Mass Spectrometry Detection. Int J Neonatal Screen 2017; 4: 2.
6. Sahai I, Eaton RB, Hale JE, Mulcahy EA, Comeau AM. Long-term follow-up to ensure quality care of individuals diagnosed with newborn screening conditions: Early experience in New England. Genet Med 2010; 12: S220-S227.
7. Levy HL. Robert Guthrie and the Trials and Tribulations of Newborn Screening. Int J Neonatal Screen 2021; 7: 5.
8. Lehotay DC, Hall P, Lepage J, Eichhorst JC, Etter ML, Greenberg CR. LC-MS/MS progress in newborn screening. Clin Biochem 2011; 44: 21-31.
9. Millington DS. How mass spectrometry revolutionized newborn screening. J Mass Spectrom Adv Clin Lab 2024; 32: 1-10.
10. Ashenden AJ, Chowdhury A, Anastasi LT, Lam K, Rozek T, Ranieri E, Siu CW, King J, Mas E, Kassahn KS. The Multi-Omic Approach to Newborn Screening: Opportunities and Challenges. Int J Neonatal Screen 2024; 10: 42.
11. Nakar-Weinstein A, Ben-Moshe Y, Sutton VR. From PKU to Genome Sequencing: The Past, Present, and Future of Newborn Screening. Clin Chem 2025; 72: 123-132.
12. Ruoppolo M, Malvagia S, Boenzi S, Carducci C, Dionisi-Vici C, Teofoli F, Burlina A, Angeloni A, Aronica T, Bordugo A, Bucci I, Camilot M, Carbone MT, Cardinali R, Carducci C, Cassanello M, Castana C, Cazzorla C, Ciatti R, Ferrari S, Frisso G, Funghini S, Furlan F, Gasperini S, Gragnaniello V, Guzzetti C, La Marca G, La Spina L, Lorè T, Meli C, Messina M, Morrone A, Nardecchia F, Ortolano R, Parenti G, Pavanello E, Pieragostino D, Pillai S, Porta F, Righetti F, Rossi C, Rovelli V, Salina A, Santoro L, Sauro P, Schiaffino MC, Simonetti S, Vincenzi M, Tarsi E, Uccheddu AP. Expanded Newborn Screening in Italy Using Tandem Mass Spectrometry: Two Years of National Experience. Int J Neonatal Screen 2022; 8: 47.
13. Tluczek A, Ersig AL, Lee S. Psychosocial Issues Related to Newborn Screening: A Systematic Review and Synthesis. Int J Neonatal Screen 2022; 8: 53.
14. Hewlett J, Waisbren SE. A review of the psychosocial effects of false-positive results on parents and current communication practices in newborn screening. J Inherit Metab Dis 2006; 29: 677-682.
15. van den Heuvel LM, van der Pal SM, Verschoof-Puite RK, Klapwijk JE, Elsinghorst E, Dekkers E, van der Ploeg CPB, Henneman L. Psychosocial Impact of a True-Positive, False-Positive, or Inconclusive Newborn Bloodspot Screening Result: A Questionnaire Study among Parents. Int J Neonatal Screen 2024; 10: 18.
16. Russo S, Greco B, Carcereri MS, Vissani S, Caviglia S. Long-term psychological support for families receiving communication of positivity for metabolic diseases at newborn screening. JIM 2024; 1: e452.
17. Gurian EA, Kinnamon DD, Henry JJ, Waisbren SE. Expanded Newborn Screening for Biochemical Disorders: The Effect of a False-Positive Result. Pediatrics 2006; 117: 1915-1921.
18. Bani M, Russo S, Gasperini S, Crescitelli V, Menni F, Furlan F, Tagliaferri F, Cefalo G, Paci S, Banderali G, Marchisio P, Biondi A, Strepparava MG. Prevalence and predictors of parental distress at the communication of positivity at newborn screening for metabolic diseases: an Italian longitudinal study. BMJ Paediatr Open 2024; 8: e003103.
19. Farrell MH. Child Health Providers' Precautionary Discussion of Emotions During Communication About Results of Newborn Genetic Screening. Arch Pediatr Adolesc Med 2012; 166: 62.

20. Chudleigh J, Shakespeare L, Holder P, Chinnery H, Hack G, Gill T, Gould R, Southern KW, Olander EK, Morris S, Bonham JR, Simpson A, Moody L. Co-designing Improved Communication of Newborn Bloodspot Screening Results to Parents: Mixed Methods Study. J Particip Med 2022; 14: e33485.
21. Bani M, Caviglia S, Bensi G, Carcereri MS, Greco B, Lastrucci E, Massa P, Vissani S, Cazzorla C. Availability of psychological resources for parents receiving communication of positivity at newborn screening for metabolic diseases in Italy. Eur J Pediatr 2024; 183: 965-969.
22. Franková V, Driscoll RO, Jansen ME, Loeber JG, Kožich V, Bonham J, Borde P, Brincat I, Cheillan D, Dekkers E, Fingerhut R, Kuš IB, Girginoudis P, Groselj U, Hougaard D, Knapková M, la Marca G, Malniece I, Nanu MI, Nennstiel U, Olkhovych N, Oltarzewski M, Pettersen RD, Racz G, Reinson K, Salimbayeva D, Songailiene J, Vilarinho L, Vogazianos M, Zetterström RH, Zeyda M; Members of the European Society of Human Genetics (ESHG)-EuroGentest Quality Sub-Committee. Regulatory landscape of providing information on newborn screening to parents across Europe. Eur J Hum Genet 2021; 29: 67-78.
23. Cazzorla C, Gragnaniello V, Gaiga G, Gueraldi D, Puma A, Loro C, Benetti G, Schiavo R, Porcù E, Burlina AP, Burlina AB. Newborn Screening for Gaucher Disease: Parental Stress and Psychological Burden. Int J Neonatal Screen 2025; 11: 14.
24. Raia MH, Lynch MM, Ward AC, Brown JA, Bonhomme NF, Hunting VL. One Size Does Not Fit All: A Multifaceted Approach to Educate Families about Newborn Screening. Int J Neonatal Screen 2024; 10: 44.
25. Vaidyanathan AK. Basics in statistics: Sample size calculation and descriptive data statistics. J Indian Prosthodont Soc 2023; 23: 207-209.
26. GraphPad Software. GraphPad Prism, version 10. GraphPad Software; 2025. Available at: https://www.graphpad.com
27. Kemper AR, Boyle CA, Brosco JP, Grosse SD. Ensuring the Life-Span Benefits of Newborn Screening. Pediatrics 2019; 144: e20190904.
28. IJzebrink A, van Dijk T, Franková V, Loeber G, Kožich V, Henneman L, Jansen M. Informing Parents about Newborn Screening: A European Comparison Study. Int J Neonatal Screen 2021; 7: 13.
29. Franková V, Dohnalová A, Pešková K, Hermánková R, O'Driscoll R, Ješina P, Kožich V. Factors Influencing Parental Awareness about Newborn Screening. Int J Neonatal Screen 2019; 5: 35.
30. Brower A, Chan K, Williams M, Berry S, Currier R, Rinaldo P, Caggana M, Gaviglio A, Wilcox W, Steiner R, Holm IA, Taylor J, Orsini JJ, Brunelli L, Adelberg J, Bodamer O, Viall S, Scharfe C, Wasserstein M, Chen JY, Escolar M, Goldenberg A, Swoboda K, Ficicioglu C, Matern D, Lee R, Watson M. Population-Based Screening of Newborns: Findings From the NBS Expansion Study (Part One). Front Genet 2022; 13: 867337.
31 DeLuca JM. Public Attitudes Toward Expanded Newborn Screening. J Pediatr Nurs 2018; 38: e19-e23.
32. Picou EM, McAlexander SN, Day BC, Jirik KJ, Morrison AK, Tharpe AM. An evaluation of newborn hearing screening brochures and parental understanding of screening result terminology. Int J Audiol 2023; 62: 541-551.
33. Gigante S, Pozzi K, Sisti VB, Cerizza A, Pretese R, Crescitelli V. Nursing perspectives on inherited metabolic disorders: experience and educational needs in an Italian hospital setting. JIM 2025; 2: e1050.
34 Boychuk NA, Mulrooney NS, Kelly NR, Goldenberg AJ, Silver EJ, Wasserstein MP. Parental Depression and Anxiety Associated with Newborn Bloodspot Screening for Rare and Variable-Onset Disorders. Int J Neonatal Screen 2022; 8: 59.

JIM
JIM 2026; 3 (3): e1211
DOI: 10.61012/JIM_202608_1211
Category: *Original Article*

HOW RELIABLE ARE LARGE LANGUAGE MODELS IN METABOLIC DIETETICS? A CROSS-SECTIONAL BENCHMARKING EVALUATION

M. TOSI[1], M. MUSCARELLA[2], G. GUGELMO[3], A. PINTO[4], M. SUÁREZ GONZÁLEZ[5,6], S. POLONI[7], A. HÖLLER[8,9,10], V. PERICO[11], A. FAVARO[12], A. TAVIAN[13], A. BELVEDERE[14], A. DIANIN[12]

• • •

[1]Metabolic Diseases Unit, Department of Pediatrics, Vittore Buzzi Children's Hospital, Milan, Italy
[2]Dietetic Unit, Meyer Children University Hospital IRCCS, Florence, Italy
[3]Division of Metabolic Diseases, Department of Medicine, University Hospital of Padua, Padua, Italy
[4]Department of Dietetics, Birmingham Women's and Children's Hospital, Birmingham, UK
[5]Department of Pediatrics, Hospital Universitario Central de Asturias, Oviedo, Spain
[6]Pediatric Research, Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo, Spain
[7]Nutrition and Dietetics Service, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
[8]Division of Nutrition and Dietetics, University Hospital Innsbruck, Innsbruck, Austria
[9]Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL-University for Health Sciences and Technology, Hall in Tirol, Austria
[10]Digital Health Information Systems, Center for Health & Bioresources, AIT Austrian Institute of Technology, Innsbruck, Austria
[11]Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy
[12]Regional Center for Newborn Screening, Diagnosis and Treatment of Inherited Metabolic Diseases and Congenital Endocrine Diseases, Pediatric Unit C, University Hospital of Verona, Verona, Italy
[13]SOSD Technical Health Care Professions – Dietitian and Regional Coordinating Center for Rare Diseases, Azienda Sanitaria Universitaria Friuli Centrale, Udine, Italy
[14]Independent AI Researcher, Verona, Italy

Corresponding Author
Martina Tosi, Ph.D; e-mail: martina.tosi@unimi.it

***ABSTRACT* – Objective:** The aim of this study was to evaluate the performance of free-tier large language models (LLMs) in providing dietary recommendations for inherited metabolic disorders (IMDs), with a focus on scientific accuracy, completeness, and response consistency of generated responses across models under real-world conditions.

Materials and Methods: This cross-sectional benchmarking study assessed four LLMs – ChatGPT, GitHub Copilot, Gemini and Claude – using 30 standardized clinical questions on dietary management in IMDs. Models were tested under real-world free-tier conditions, without fine-tuning or prompt engineering. Each question was submitted three times to each model. Responses were independently evaluated by three metabolic dietitians using a 6-point ordinal scale for scientific accuracy and completeness, while response consistency was assessed across repeated outputs. Inter-rater agreement was assessed using the intraclass correlation coefficient (ICC).

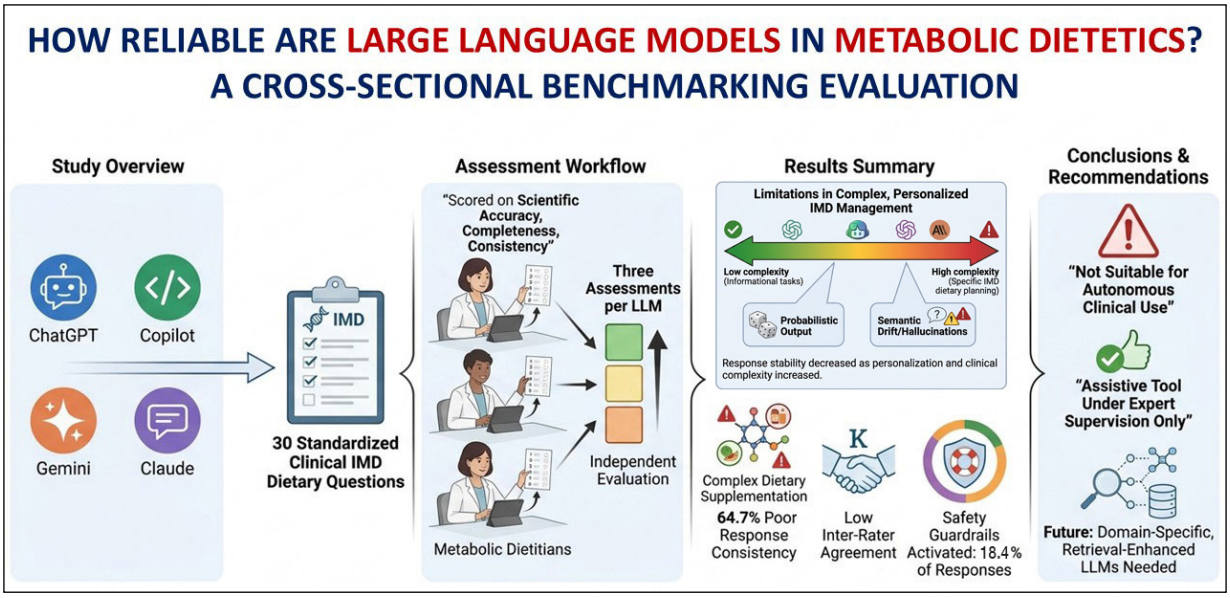
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Results: A total of 359 unique responses were collected for evaluation. Safety guardrails were activated in 18.4% of responses, with substantial variability across models and clinical domains. Of these, 80.3% resulted in a complete block and were excluded from the analysis. Consequently, the final sample consisted of 306 responses. Model performance was heterogeneous: ChatGPT achieved the highest and most consistent scores, with no severe inaccuracies or omissions. Copilot, Gemini and Claude showed intermediate performance, with relevant variability. Across clinical domains, responses related to diet planning and clinical monitoring were generally more stable, whereas dietary supplementation and nutritional requirements showed greater variability and clinically relevant errors. Poor response consistency was observed in 64.7% of responses, indicating limited consistency across repeated identical prompts. Inter-rater agreement is defined as poor for accuracy, completeness, and response consistency.

Conclusions: The performance of free-tier LLMs in IMD dietary management is heterogeneous, with substantial limitations in response consistency and in clinically complex scenarios. These findings indicate that current models are not yet suitable for autonomous clinical use and can be dangerous to use without a healthcare professional's oversight.

Keywords: Artificial intelligence, Large language models, Dietetics, Inherited metabolic disorder, Clinical decision support.

Abbreviations: AI, artificial intelligence; CKD, chronic kidney disease; IMD, inherited metabolic disorder; IMDs, inherited metabolic disorders; ICC, intraclass correlation coefficient; Q1, first quartile; Q3, third quartile; LLM, large language model; LLMs, large language models; MMA, methylmalonic acidemia; RAG, retrieval-augmented generation.



Graphical Abstract. Comparative evaluation of four free-tier large language models in answering patient-like questions on dietary management of inherited metabolic disorders, assessing scientific accuracy, completeness, and response consistency under real-world conditions, and demonstrating heterogeneous performance with limitations for autonomous clinical use.

INTRODUCTION

Internet searches for health-related information represent a rapidly expanding phenomenon in modern healthcare communication. Advances in artificial intelligence (AI) have led to the development of large language models (LLMs), such as ChatGPT, which use natural language processing and machine learning techniques to generate contextually appropriate, human-like responses to user queries[1]. These systems are increasingly able to synthesize information and provide coherent, interactive answers, contributing to growing interest in their use as accessible tools for preliminary medical and health-related guidance[2]. These tools may offer advantages by providing real-time, personalized information and enhancing patient engagement[3]. More broadly, AI and machine learning methods are increasingly being applied in nutri-

tion research to address high-dimensional datasets, enable predictive modeling, and support personalized dietary recommendations[4]. These approaches range from classical statistical learning techniques to advanced deep learning architectures, including models such as LLMs that can process unstructured clinical and nutritional text data[5]. However, despite their rapid adoption, the effectiveness and reliability of LLMs in providing domain-specific guidance remain insufficiently characterized. Recent scoping evidence indicates that, although a growing number of studies have explored applications of LLMs in nutrition and dietetics, most investigations are based on hypothetical scenarios rather than real-world clinical implementations, with limited use of domain-specific or retrieval-augmented approaches[6]. In particular, their utility in nutritional counseling and dietary management has not been fully evaluated, especially in the context of rare and complex conditions such as inherited metabolic disorders (IMDs). Dietary management in IMDs is particularly challenging because recommendations are disease-specific, age-dependent, and frequently require continuous adjustment according to biochemical monitoring and metabolic control. The aim of this study was to evaluate the scientific accuracy, completeness, and response consistency of responses generated by four freely available LLMs (ChatGPT, GitHub Copilot Free, Gemini, and Claude) to standardized patient-oriented questions concerning dietary management in IMDs. Responses were independently assessed by expert metabolic dietitians according to current clinical guidelines.

MATERIALS AND METHODS

Study Design

To address the current lack of evidence regarding the reliability of freely available LLMs in providing dietary recommendations for IMDs, we conducted a cross-sectional comparative study. Four LLMs (ChatGPT, GitHub Copilot Free, Gemini, and Claude) were assessed under real-world conditions using their publicly accessible free versions, without fine-tuning, retrieval-augmented generation (RAG), or domain-specific prompt engineering. The overall study design is summarized in Figure 1. The study consisted of three sequential phases: (i) development of standardized patient-style questions, (ii) response generation by each LLM under identical conditions, and (iii) blinded evaluation of responses by expert metabolic dietitians.

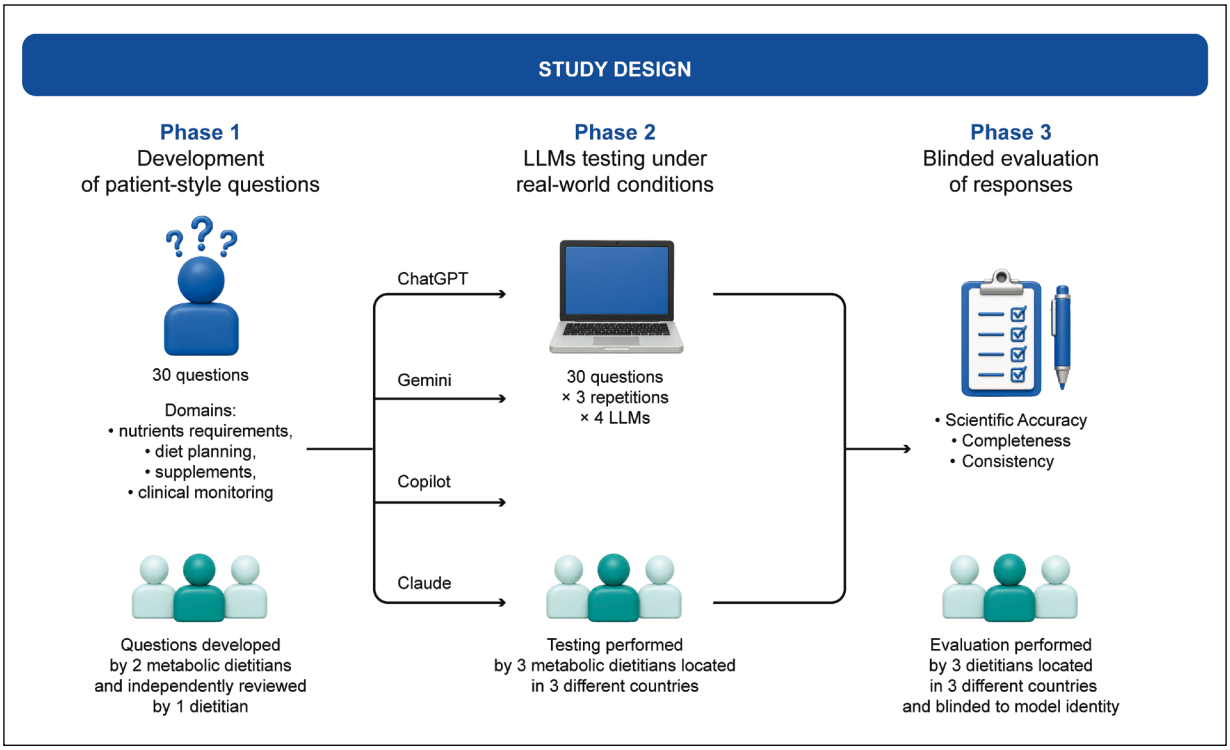


Figure 1. Study design illustrating the evaluation of free-tier LLM responses to patient-style dietary management queries for IMDs.

Questions Development

A standardized set of 30 clinically relevant questions was developed to cover the main domains of dietary management in IMDs (Appendix 1): nutrient requirements and restrictions (n = 10), diet planning (n = 5), dietary supplements (n = 5), and clinical considerations and monitoring (n = 10).

The complete set of questions is available in Appendix 1, while Table 1 provides an overview of the investigated dietary management domains together with representative questions submitted to the LLMs. Questions were developed by two experienced metabolic dietitians based on current dietary guidelines for IMDs and independently reviewed by a third metabolic dietitian. As the study involved experts from different countries, all questions were standardized and submitted to the LLMs in English. Content validity was established through alignment with current international guideline recommendations[7-18] and representative real-world clinical scenarios. The question set was designed to encompass different levels of complexity, ranging from basic nutritional knowledge to advanced clinical decision-making. Any disagreements regarding the wording or content of the questions were resolved by consensus. Investigators involved in question development did not participate in LLM testing or response evaluation.

Table 1. Main investigated topics and categories with representative questions submitted to the LLMs.

	Nutrient requirements and restrictions	Diet planning	Dietary supplements	Clinical considerations and monitoring
Disorders of protein or amino acid metabolism GA I, TYR II, UCD, MSUD, PKU, MMA, CTLN 1	• Low-protein foods • High in lysine foods • Low in Phe foods • Breastfeeding	• Meals at a restaurant	• PS recommended dose • Gastrointestinal symptoms following PS use	• Night-time DBS collection • DBS results • Pregnancy • Emergency feeds (vomiting)
Disorders of carbohydrate metabolism GAL, HFI, GSD I, GSD III, GSD V	• Lactose-free dairy products • Ketogenic diet • Fructose in vegetables	• Ketones levels • Daily snacks • Ideas for breakfast	• Multivitamin supplements	• Obesity and weight loss
Defects of fatty acid beta-oxidation VLCAD, LCHAD, CPT 2	• Low-fat foods	• Number of meals during the day and night	• Maltodextrin use and physical activity	• Sports • Hiking

CPT II, carnitine palmitoyltransferase II deficiency; CTLN1, citrullinemia type I; DBS, dried blood spot; GA I, glutaric aciduria type I; GAL, galactosemia; GSD I, glycogen storage disease type I; GSD III, glycogen storage disease type III; GSD V, glycogen storage disease type V; HFI, hereditary fructose intolerance; LCHAD, long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency; MMA, methylmalonic acidemia/aciduria; MSUD, maple syrup urine disease; Phe, phenylalanine; PKU, phenylketonuria; PS, protein substitute; TYR II, tyrosinemia type II; UCD, urea cycle disorders; VLCAD, very-long-chain acyl-CoA dehydrogenase deficiency.

LLMs Selection

Four freely available LLMs were evaluated: 1. ChatGPT (OpenAI, free version), accessed through OpenAI's publicly available web interface. The exact model version and inference configuration were not disclosed by the provider. 2. GitHub Copilot Free (GitHub/Microsoft), accessed through the GitHub Copilot interface. This platform dynamically routes user requests to different LLMs depending on system availability, without disclosing the specific model used for each interaction. 3. Gemini (Google, free

version), accessed through Google's publicly available web interface. Although access was provided to the Gemini family of models, the exact model variant and configuration were not specified. 4. Claude (Anthropic, free version), accessed through Anthropic's publicly available web interface. The precise model version and inference parameters were not disclosed. All models were evaluated using their publicly accessible free versions without subscription plans, custom instructions, fine-tuning, or external knowledge integration. Testing was conducted under routine user conditions, with the understanding that free-tier services may be subject to usage limits and dynamic model allocation policies.

Model Testing Procedure

All LLMs were evaluated using their default configurations without fine-tuning, custom instructions, or additional prompt engineering. Each model was accessed through a newly created, neutral free-tier account with no prior interaction history. To minimize potential personalization effects, each query was submitted in a separate chat session, thereby preventing conversational memory from influencing subsequent responses. Identical, clinically phrased prompts were submitted to all models. Each question was entered three times independently for each LLM to assess response consistency. Model testing was performed by three metabolic dietitians from three different countries (Brazil, Austria, and Italy), who were responsible for submitting the standardized questions and collecting the generated responses. These investigators were not involved in either question development or response evaluation. All interactions with the LLMs were conducted through their respective publicly available web interfaces between February and March 2026. Testing sessions were distributed across different days to reduce the potential influence of transient server load or temporary model updates. The exact version of the model and its inference configuration were not made available by the provider. Because LLM providers implement silent updates and automatic changes without notifying users, it was not possible for users of the free versions to determine or verify the specific model version generating the response. No follow-up questions or clarifications were provided after the initial prompt, and responses were recorded without modification.

Response Evaluation

Responses were independently evaluated by three metabolic dietitians from three different countries (Italy, the United Kingdom, and Spain), all with extensive experience in the clinical management of IMDs and a peer-reviewed publication record in the field. To minimize assessment bias, all responses were anonymized, randomized, and evaluated in a blinded manner with respect to model identity. Each evaluator independently assessed all responses using predefined scoring criteria. As all prompts and responses were generated in English, evaluations were also performed in English. All evaluators were proficient in English and had more than 10 years of experience in the clinical management of IMDs, ensuring reliable and consistent assessment.

Outcome Measures

- Scientific accuracy was defined as the extent to which each response was consistent with current evidence-based dietary guidelines for IMDs. Each response was rated using a 6-point ordinal scale: 0 = serious scientific inaccuracies; 1 = major inaccuracies; 2 = some inaccuracies; 3 = mostly accurate with minor errors; 4 = accurate with only very minor or negligible errors; and 5 = completely accurate according to current evidence-based guidelines. Higher scores indicated greater scientific accuracy.
- Completeness was defined as the extent to which each response addressed all clinically relevant elements required to provide a comprehensive answer to the question. Responses were rated using a 6-point ordinal scale: 0 = fails to address the question; 1 = addresses very few relevant aspects; 2 = addresses some relevant aspects; 3 = addresses most relevant aspects; 4 = addresses all relevant aspects with minor omissions or limited depth; and 5 = fully addresses all relevant aspects of the question. Higher scores indicated greater completeness.
- Response consistency: Response reproducibility was defined as the consistency of responses generated by the same LLM when presented with identical prompts across repeated independent runs. For each question–model pair, three independently generated responses were compared descriptively to assess within-model response consistency. Responses were classified as fully consistent (score =

2), partially consistent (score = 1), or inconsistent (score = 0). A score of 2 indicated that all three responses conveyed essentially the same information with only negligible wording differences; a score of 1 indicated minor variations in wording or emphasis without changes in the overall content; and a score of 0 indicated substantial differences in the information or recommendations provided.

- Inter-rater agreement. Agreement among the three independent evaluators was assessed for each outcome domain (scientific accuracy, completeness, and response consistency) using the intraclass correlation coefficient (ICC). A two-way random-effects model with absolute agreement was used, as all responses were independently evaluated by the same three raters, who were considered representative of a broader population of expert metabolic dietitians. ICC values were interpreted according to commonly accepted criteria: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent agreement. In addition, the percentage agreement among evaluators was calculated as a descriptive measure.

Statistical Analysis

All outcomes were analyzed using descriptive statistics. For scientific accuracy and completeness, scores assigned by the three independent evaluators were aggregated at the response level using two complementary approaches. First, the mean score across evaluators was calculated for each response, and the resulting values were summarized using the median and first and third quartiles (Q1-Q3). Second, the median score across evaluators was calculated for each response and subsequently summarized using the overall median and Q1-Q3. These two approaches were reported to provide complementary estimates of central tendency and variability, with the median-based aggregation providing a more robust measure less influenced by potential outlier ratings among evaluators. Response-level median scores were used to identify extreme performance outcomes. Severe inaccuracy and severe incompleteness were defined as median scores ≤1.5 on the 6-point ordinal scale. For response reproducibility, consistency scores derived from repeated runs of each LLM were summarized descriptively using median values and Q1-Q3. Poor response reproducibility was defined as a median consistency score <2. Inter-rater agreement among the three independent evaluators was assessed for each outcome domain (scientific accuracy, completeness, and response consistency) using the ICC, as described above. A two-way random-effects model with absolute agreement and average measures [ICC(2,*k*)] was applied. ICC values were interpreted according to commonly accepted criteria: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent agreement. Safety guardrails identified in LLM-generated responses were characterized using descriptive analyses. No comparative inferential analyses between LLMs were performed.

Ethical Considerations

This study evaluated only AI-generated outputs; no human participants or patient data were included. Expert participation was voluntary.

RESULTS

A total of 360 AI-generated responses were initially collected across four free-tier LLMs. Before the blinded evaluation phase, one response was excluded due to a data collection error, as a duplicated response had been inadvertently recorded for a different query (diet planning domain, Question 2, Gemini – Model C). The final dataset, subjected to expert evaluation and statistical analysis, therefore consisted of 359 unique responses. When LLMs were unable to fully address specific medical queries, safety guardrails were activated in 18.4% of cases (66/359). Of these, 80.3% (53/66) resulted in a complete block with no scientific content and were excluded from the analysis. Consequently, the final sample consisted of 306 responses.

Overall Model Performance

Overall performance varied across platforms. ChatGPT (Model A) showed the highest overall performance, with a consensus median scientific accuracy score of 4.0 (4.0-4.0) and a median of averaged scores of 4.0 (3.3-4.3). Completeness was also high, with a consensus median score of 4.0 (4.0-4.0) and

a median of averaged scores of 4.0 (3.7-4.3). In contrast, GitHub Copilot (Model B) showed low performance in scientific accuracy, with a consensus median scientific accuracy score of 3.0 (2.0-3.0) and a median of averaged scores of 3.3 (3.0-3.7). Completeness followed a different pattern, with a consensus median score of 4.0 (3.0-4.0) and a median of averaged scores of 3.3 (3.0-3.7). Gemini (Model C) and Claude (Model D) showed comparable performance. Gemini had a consensus median scientific accuracy score of 3.0 (2.0-4.0) and a median of averaged scores of 3.3 (2.7-3.7); completeness was similar, with a consensus median score of 3.0 (3.0-3.0) and a median of averaged scores of 3.3 (3.0-3.7). Claude had a consensus median scientific accuracy score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (2.7-3.7). Completeness was slightly lower, with a consensus median score of 3.0 (3.0-3.0) and a median of averaged scores of 3.0 (2.7-3.3). Table 2 summarizes the performance of the evaluated models.

Table 2. Summary of the overall model performance, reporting scientific accuracy and completeness, together with the number of severe inaccuracies and instances of severe incompleteness for each model.

Model	Scientific accuracy*	Completeness*	Severe inaccuracies % (n/total number of responses)	Severe incompleteness % (n/total number of responses)
Model A (ChatGPT)	4.0 (4.0-4.0)/ 4.0 (3.3-4.3)	4.0 (4.0-4.0)/ 4.0 (3.7-4.3)	0.0% (0/90)	0.0% (0/90)
Model B (Copilot)	3.0 (2.0-3.0)/ 3.3 (3.0-3.7)	4.0 (3.0-4.0)/ 3.3 (3.0-3.7)	2.4% (1/42)	0.0% (0/42)
Model C (Gemini)	3.0 (2.0-4.0)/ 3.3 (2.7-3.7)	3.0 (3.0-3.0)/ 3.3 (3.0-3.7)	6.0% (5/84)	6.0% (5/84)
Model D (Claude)	3.0 (3.0-4.0)/ 3.3 (2.7-3.7)	3.0 (3.0-3.0)/ 3.0 (2.7-3.3)	8.9% (8/90)	0.0% (0/90)
Total	–	–	**4.6% (14/306)**	**1.6% (5/306)**

*Data are presented as consensus median (Q1-Q3)/median of averaged scores (Q1-Q3). Severe inaccuracies and incompleteness are defined as a consensus median score ≤1.5. For Scientific Accuracy: 0 = serious scientific inaccuracies; 1 = major inaccuracies; 2 = some inaccuracies; 3 = mostly accurate with minor errors; 4 = accurate with only very minor or negligible errors; 5 = completely accurate based on current evidence-based guidelines. For Completeness: 0 = fails to address the question; 1 = addresses very few aspects; 2 = addresses some aspects; 3 = addresses most aspects; 4 = addresses all aspects with minor omissions or limited depth; 5 = fully addresses all aspects of the question.

Performance by Clinical Domain

Performance varied across clinical domains, reflecting differences in task complexity. In the domain of "nutrient requirements and restrictions", models achieved a consensus median scientific accuracy score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (2.7-3.7). Completeness followed a similar distribution, with a consensus median score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (3.0-3.7). "Diet planning" showed similar performance, reaching a consensus median score of 3.0 (3.0-3.0) and a median of averaged scores of 3.3 (3.3-3.3) for both scientific accuracy and completeness, reflecting absolute consistency in the averaged metrics.

The highest scientific accuracy was observed in the "dietary supplement" domain, which reached a consensus median score of 4.0 (4.0-4.0) and a median of averaged scores of 3.8 (3.8-3.8). Completeness in this domain presented a median score of 4.0 (4.0-4.0) and a median of averaged scores of 3.7 (3.7-3.7).

In contrast, the "clinical considerations and monitoring" domain showed comparable performance, with a slightly higher average accuracy. For scientific accuracy, this domain achieved a consensus median score of 3.0 (3.0-4.0) and a median of averaged scores of 3.7 (3.0-4.0). For completeness, it reached a consensus median score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (3.3-3.7). Table 3 summarizes the performance across the clinical domain.

Table 3. Summary of the performance across clinical domains, reporting clinical accuracy and completeness, together with the number of severe inaccuracies and instances of severe incompleteness for each domain.

Clinical domain	Scientific accuracy*	Completeness*	Severe inaccuracies % (n/total number of responses)	Severe incompleteness % (n/total number of responses)
Nutrient requirements and restrictions	3.0 (3.0-4.0)/ 3.3 (2.7-3.7)	3.0 (3.0-4.0)/ 3.3 (3.0-3.7)	4.6% (5/108)	1.9% (2/108)
Diet planning	3.0 (3.0-3.0)/ 3.3 (3.3-3.3)	3.0 (3.0-3.0)/ 3.3 (3.3-3.3)	2.1% (1/47)	0.0% (0/47)
Dietary supplements	4.0 (4.0-4.0)/ 3.8 (3.8-3.8)	4.0 (4.0-4.0)/ 3.7 (3.7-3.7)	10.0% (5/50)	4.0% (2/50)
Clinical considerations and monitoring	3.0 (3.0-4.0)/ 3.7 (3.0-4.0)	3.0 (3.0-4.0)/ 3.3 (3.3-3.7)	3.0% (3/101)	1.0% (1/101)
Total	–	–	**4.6% (14/306)**	**1.6% (5/306)**

*Data are presented as consensus median (Q1-Q3)/median of averaged scores (Q1-Q3). Severe inaccuracies and incompleteness are defined as a consensus median score ≤ 1.5. For Scientific Accuracy: 0 = serious scientific inaccuracies; 1 = major inaccuracies; 2 = some inaccuracies; 3 = mostly accurate with minor errors; 4 = accurate with only very minor or negligible errors; 5 = completely accurate based on current evidence-based guidelines. For Completeness: 0 = fails to address the question; 1 = addresses very few aspects; 2 = addresses some aspects; 3 = addresses most aspects; 4 = addresses all aspects with minor omissions or limited depth; 5 = fully addresses all aspects of the question.

Critical Failures: Severe Inaccuracies and Incompleteness

Overall, severe scientific inaccuracies were identified in 4.6% of responses (14/306), while severe incompleteness was observed in 1.6% (5/306). When stratified by LLM, the distribution of critical failures was uneven. Claude accounted for the highest rate of severe inaccuracies (8.9%, 8/90). Gemini and GitHub Copilot showed lower rates, observed in 6.0% (5/84) and 2.4% (1/42) of responses, respectively. Gemini was the only model presenting severe incompleteness (6.0%, 5/84). ChatGPT did not generate any responses classified as severely inaccurate or severely incomplete based on the consensus median score.

When the 14 severe scientific inaccuracies were stratified by clinical domain, a notable pattern emerged in the "dietary supplement" domain: despite having the highest overall median accuracy, this domain contained the highest number of severe errors (10.0%, 5/50). The remaining severe inaccuracies were distributed across "nutrient requirements and restrictions" (4.6%, 5/108), "clinical considerations and monitoring" (3.0%, 3/101), and "diet planning" (2.1%, 1/47).

Regarding severe incompleteness, the five identified cases were concentrated within specific clinical domains. The highest number of critical omissions occurred in the "dietary supplement" domain (4.0%, 2/50), followed by "nutrient requirements and restrictions" (1.9%, 2/108) and "clinical considerations and monitoring" (1.0%, 1/101). Table 2 summarizes the performance of the evaluated models with respect to scientific accuracy, completeness, number of severe inaccuracies, and severe incompleteness. Table 3 presents an analogous analysis within the clinical domain.

Response Consistency

The models showed limited consistency when the same query was submitted multiple times. Overall, 64.7% of responses (198/306) were classified as inconsistent or only partially consistent across the three repetitions. Detailed results of the response consistency analysis, stratified by model and clinical domain, are presented in Tables 4 and 5.

Table 4. Response consistency across models, reporting the proportion of responses with poor response consistency (median consistency score <2) and those with full consistency (consistency score=2).

Model	Poor consistency (score <2), % (n/total number of responses)	Fully consistent (score=2), % (n/total number of responses)
Model A (ChatGPT)	60.0% (54/90)	40.0% (36/90)
Model B (Copilot)	92.9% (39/42)	7.1% (3/42)
Model C (Gemini)	82.1% (69/84)	17.9% (15/84)
Model D (Claude)	40.0% (36/90)	60.0% (54/90)
Total	**64.7% (198/306)**	**35.3% (108/306)**

Table 5. Response consistency across clinical domains, reporting the proportion of responses with poor response consistency (median consistency score <2) and those with full consistency (consistency score=2).

Clinical Domain	Poor consistency, (score <2)% (n)	Fully consistency (score = 2), % (n)
Nutrient requirements and restrictions	66.7% (72/108)	33.3% (36/108)
Diet planning	74.5% (35/47)	25.5% (12/47)
Dietary supplements	58.0% (29/50)	42.0% (21/50)
Clinical considerations and monitoring	66.4% (62/101)	33.6% (39/101)
Total	**64.7% (198/306)**	**35.3% (108/306)**

Inter-rater Agreement

Inter-rater agreement was poor across all outcome domains. For scientific accuracy, the ICC(2,*k*) was 0.468 (95% CI: 0.36-0.56; $p<0.001$), indicating poor agreement among evaluators. Similarly, completeness showed poor agreement, with an ICC(2,*k*) of 0.351 (95% CI: 0.20-0.47; $p<0.001$). Finally, response reproducibility also demonstrated poor agreement, with an ICC(2,*k*) of 0.312 (95% CI: -0.01-0.53; $p<0.001$) (Table 6).

Table 6. Inter-rater agreement for each evaluation metric, reporting ICC(2,*k*) values and corresponding 95% confidence intervals (CI) for scientific accuracy, completeness, and response reproducibility.

Metric	ICC(2,*k*)	95% CI	*p*-value	Interpretation
Scientific accuracy	0.468	0.36, 0.56	<0.001	Poor
Completeness	0.351	0.2, 0.47	<0.001	Poor
Response consistency	0.312	-0.01, 0.53	<0.001	Poor

Data interpretation: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent agreement.

Safety Guardrails

Safety guardrails were activated in 18.4% of cases (66/359) when LLMs were unable to fully address specific medical queries. Among these, 80.3% (53/66) resulted in a complete response block with no scientific content, whereas 19.7% (13/66) generated a partial response that included a recommendation to seek clinical consultation while still providing scientific information.

Considering the median scientific accuracy and completeness scores across all 359 responses, an important association with the activation of safety filters emerged. Specifically, a median scientific accuracy score of 0.0 was observed in 7 of the 359 responses, and all of these cases were associated with the activation of safety mechanisms. Of these seven responses, 1/7 belonged to the "diet planning" domain and 6/7 to the "dietary supplement" domain. Looking at the model, 4/7 responses were generated by Copilot (Model B) and 3/7 by Gemini (Model C), with most cases occurring during Testing 3 (5/7) rather than Testing 2 (2/7). Similarly, 8 responses received a median completeness score of 0.0, all associated with safety filter activation. These cases involved the "diet planning" (2/8), "dietary supplement" (5/8), and "clinical considerations and monitoring" (1/8) domains. Copilot (Model B) and Gemini (Model C) each accounted for four responses, with most cases occurring during Testing 3 (6/8) rather than Testing 2 (2/8).

Excluding responses with a median score of 0.0 (i.e., fully blocked outputs), 59 responses remained for the analysis of scientific accuracy and 58 for completeness. For the 59 responses assessed for scientific accuracy, safety mechanisms led to complete blocks in 39 cases, partial blocks in 13 cases, and brief pathology-only explanations in seven cases. A similar distribution was observed for the 58 responses assessed for completeness, with 39 cases resulting in a complete block, 12 in a partial block, and seven providing only a brief pathological overview before termination.

For scientific accuracy, safety mechanism activation was highest in GitHub Copilot (Model B; 53.3%, 48/90), followed by Gemini (Model C; 20.2%, 18/89), while no activations were observed for ChatGPT (Model A) or Claude (Model D). A qualitative divergence was observed in the implementation of safety filters across LLMs. Model B (GitHub Copilot) exhibited a more rigid safety architecture, with restrictions manifesting almost exclusively as complete blocks (87.5%, 42/48) or, in a minority of cases, as brief pathology-only explanations (12.5%, 6/48). No mid-generation terminations were observed. In contrast, Model C (Gemini) showed a more reactive filtering behavior, with 72.2% (13/18) of safety interventions resulting in partial blocks characterized by mid-output interruption. Notably, Model C accounted for all partial blocks observed in the study, indicating a distinct approach to clinical content restriction compared with Model B.

Activation varied across clinical domains, peaking in "dietary supplement" (21.7%, 13/120) and "diet planning" (20.3%, 12/59), and being lowest in "nutrient requirements and restrictions" (15.0%, 23/120). Safety restriction patterns varied by clinical domain. Clinical considerations and monitoring elicited the most rigid responses, with 82.6% (19/23) resulting in complete blocks. In contrast, nutrient requirements and restrictions showed a more heterogeneous distribution (seven complete blocks, six partial blocks, and five brief pathology-only responses).

Substantial geographic and session-based variability was also observed, with activation occurring during Testing 2 in Brazil (18.3%, 22/120) and peaking during Testing 3 in Austria (37.0%, 44/129), while no activations were observed during Testing 1 in Italy.

Responses in which the safety mechanism was applied were not evaluated uniformly by the three evaluators. Regarding scientific accuracy (0.0-5.0 scale), 32.3% (64/198) received a score of 0.0, 13.1% (26/198) a score of 1.0, 41.9% (83/198) a score of 2.0, 7.1% (14/198) a score of 3.0, 4.0% (8/198) a score of 4.0, and only 1.5% (3/198) achieved the maximum score of 5.0. Evaluator 2 applied a stricter scoring approach, assigning 0.0 in 86.4% of cases (57/66), whereas Evaluator 1 and Evaluator 3 predominantly assigned scores of 2.0 (54.5%, 36/66 and 71.2%, 47/66, respectively). Out of 1,077 total evaluations, only 8 (0.7%) received a scientific accuracy score of 0.0 without safety mechanism activation, reflecting incorrect rather than safety-restricted responses.

DISCUSSION

In this study, the performance of free-tier LLMs applied to dietary management in IMDs varied considerably across all models in terms of accuracy, completeness, and consistency of responses.

Among the investigated models, ChatGPT (Model A) showed the most consistent performance, suggesting greater robustness than other systems assessed, whereas Claude (Model D) accounted for the most severe inaccuracies. Copilot (Model B), Gemini (Model C), and Claude (Model D) achieved acceptable median scores but

showed relevant variability, especially across repeated or more complex clinical queries. The models generally demonstrated a consistent ability to handle basic nutritional information. Across clinical domains, "dietary supplements" received higher and more stable scores, but also the most severe inaccuracies. "Nutrient requirements and restrictions", "diet planning" " and "clinical considerations and monitoring" showed moderate performance, lower consistency and greater variability. This suggests that domains requiring more nuanced or guideline-dependent reasoning remain challenging for these models. The uneven distribution of severe errors also indicates that summary measures alone may mask high-risk outliers, and the broader distribution in the interquartile ranges, particularly in decision-sensitive clinical scenarios, highlights the inherent ambiguity and higher risk of inconsistent or incomplete advice. Response consistency analyses further showed inconsistency across prompt repetitions, indicating limited stability of free-tier models in generating clinical recommendations. Finally, the poor inter-rater agreement among expert evaluators reflects the inherent ambiguity of AI-generated clinical text and the difficulty of its consistent interpretation in specialized clinical contexts.

The results of this study show that safety guardrails play a central role in shaping both the completeness and the scientific accuracy of LLM outputs in clinical contexts. The strong parallelism observed between these two metrics suggests that when safety mechanisms are activated, they consistently limit not only the presence of clinical content but also its depth and correctness. Safety mechanism activation was not uniform across models, with Model B and Model C accounting for all observed interventions, while no activations were recorded for Model A and Model D. This highlights substantial heterogeneity in how different LLMs implement safety constraints in response to similar clinical queries. The nature of the safety intervention differed across models, ranging from complete response blocks to partial outputs or brief pathology-only explanations. In particular, Model B predominantly produced full blocks, whereas Model C more frequently generated partial responses, suggesting different operationalizations of content restriction. These patterns were also reflected in the variability observed across clinical domains, where more sensitive or high-risk topics elicited stronger safety responses. Taken together, these findings indicate that safety guardrails are not merely binary filters but rather structured mechanisms that significantly influence both the availability and the form of clinical information provided by LLMs. The variability observed across LLMs, geographic locations, and testing periods may also be partially influenced by platform-level safety and control mechanisms, rather than by intrinsic model differences alone. Potential contributing factors include iterative system updates, regional policy differences, and variations in medical safety filtering configurations across platforms[19].

Evidence from nutrition and clinical settings supports both the potential and the limitations of disease-specific LLM applications. In diabetes care, ChatGPT has generally shown acceptable performance when addressing common patient questions[2,20]; however, occasional inaccuracies and excessive or mixed-quality information raise concerns about response reliability[20,21]. In type 2 diabetes and metabolic syndrome, ChatGPT outputs were often clear and coherent but showed gaps in key components, including weight-loss strategies, energy-deficit recommendations, nutrient-specific guidance, and monitoring[22]. Although LLMs can generate plausible nutritional advice, they may deviate from established clinical guidelines or dietitian-designed references, particularly in more complex and multi-step interventions[23].

In chronic kidney disease (CKD), ChatGPT and other models achieved moderate-to-high accuracy in structured tasks such as identifying potassium and phosphorus content in foods; however, performance appears to depend on task complexity and is less reliable when generating personalized, stage-specific dietary recommendations, which require the integration of guidelines and patient-specific factors[24,25]. In addition, guideline-based evaluations of CKD dietary scenarios showed inconsistent energy, macro- and micronutrient targets, marked inter-model variability and no full compliance with established nutritional recommendations[26]. Expert assessments further confirm discrepancies in perceived accuracy, comprehensiveness, and clinical applicability across models, with performance depending on task formulation and model type[26].

In celiac disease, LLMs showed generally high accuracy and clarity; however, the non-negligible rates of misinformation (ranging from approximately 13% to 24%) have raised concerns about unsupervised patient use[27]. Mean quality scores of LLMs' responses may be high, around 4.3/5, yet variability persists across models and question types, especially for diagnostic content[28].

In IMDs such as methylmalonic acidemia (MMA), where treatment requires precise, individualized, and tightly controlled dietary interventions, ChatGPT displayed low appropriateness, with many dietary recommendations judged as inadequate[29].

In general nutrition advice, ChatGPT was described as an effective tool[1] and as providing answers similar to or even superior to those of dietitians, in terms of scientific correctness, practical applicability, and clarity[3]. However, limitations emerge with complex individualized nutritional care and in domains characterized by scientific uncertainty and heterogeneous evidence, such as dietary supplements. In-

deed, accuracy remains considerably lower than that observed in structured nutrition tasks, and vulnerability to misinformation increases along with inconsistent or controversial evidence[30].

Compared with more prevalent chronic conditions, IMDs involve highly specialized and strictly individualized dietary management, characterized by precise quantitative restrictions and disorder-specific metabolic requirements. This level of complexity may challenge LLMs, as it demands highly detailed, condition-specific metabolic knowledge that is less frequently encountered in general clinical nutrition literature, potentially limiting the consistency and completeness of generated recommendations.

Overall, LLM performance appears stronger in low-complexity informational tasks and progressively less reliable with increasing clinical specificity and personalization. This pattern may reflect the probabilistic architecture of LLMs, which generate outputs through statistical inference over learned representations rather than explicit rule-based reasoning[31]. Because knowledge is distributed across high-dimensional semantic spaces, model performance depends on the density and coherence of the activated semantic region and on the constraints provided by the prompt[32,33]. General queries usually align with well-represented semantic regions, whereas highly specialized clinical inputs may activate sparse or overlapping domains, increasing the risk of semantic drift and hallucinations[34,35]. Transparent reporting of model conditions is methodologically essential, as generation settings can introduce variability in model outputs, potentially affecting response consistency and comparability across runs. Thus, identical inputs may yield different responses across runs, especially in complex clinical scenarios where prompt ambiguity interacts with probabilistic decoding[36].

In real-world settings, the LLM response variability may be further amplified by underspecified or inconsistently formulated user queries. Additionally, professional output interpretation may differ, with clinicians prioritizing medical plausibility and dietitians emphasizing nutritional accuracy and practical applicability.

Collectively, these findings highlight the limited reliability of current LLMs for clinical use, with patient safety remaining critical, especially when inaccurate recommendations may have direct clinical consequences. While LLMs may offer advantages in accessibility, scalability, and patient engagement, they should be considered as assistive tools for clinical nutrition, operating under human oversight rather than autonomous decision-making systems[37]. A cautious and structured integration of LLMs into clinical nutrition, primarily for information access and educational support, rather than as standalone systems for dietary decision-making, is therefore desirable.

The development of domain-specific, retrieval-enhanced systems grounded in validated clinical databases represents a necessary step toward safer deployment in medical contexts[38]. Although generative AI systems are rapidly expanding in personalized dietary applications, most current implementations remain largely preference-driven rather than biologically grounded[39]. Clinically reliable AI-driven nutrition will require improved model architectures, standardized reporting frameworks, and stronger integration of biological data sources. General LLMs, which are not systematically trained on structured food-nutrition mappings or ontology-aligned datasets, have shown reduced performance in nutrient estimation, food classification, and semantic entity linking, whereas domain-specific models fine-tuned on large-scale, task-aligned datasets achieved higher accuracy and semantic consistency across nutritional profiling and structured classification tasks[40]. This could be achieved through supervised fine-tuning on curated clinical nutrition datasets, integration of structured food composition databases, and alignment with biomedical ontologies to improve consistency in nutrient representation and clinical reasoning.

This study has several limitations that should be acknowledged. First, the number of evaluators was limited, as this was a preliminary exploratory study designed to assess feasibility and generate initial insights rather than to provide definitive conclusions. The large number of evaluated responses also imposed a substantial cognitive and time burden, which may have contributed to inter-evaluator variability. Second, the study employed a predominantly descriptive and qualitative analytical approach, which, while appropriate for exploratory purposes, does not allow for formal inferential comparisons. Third, only four LLMs were included, which limits the generalizability of the findings across the rapidly evolving landscape of AI systems. In addition, given the fast pace of development in this field, model performance and configurations may change over time, meaning that results may not fully reflect future or updated versions of the same systems. Finally, all LLMs were accessed *via* free-tier interfaces with quotas, rate limits, and restricted availability, representing a further limitation. Exact model versions, configurations, and routing procedures were often undisclosed, limiting response consistency and direct comparison across platforms. Inter-evaluator variability may also have been influenced by heterogeneous response formats, including partial or safety-filtered outputs, and by varying levels of evaluator familiarity with specific clinical domains, which could have led to differences in scoring interpretation across conditions.

Future research should also define the levels of dietary education and counseling that can be safely supported by LLMs in IMDs. Transversal educational content may be more amenable to AI assistance, whereas counseling involving treatment complexity, metabolic control, clinical monitoring, or patient-specific dietary adaptation should remain highly personalized and supervised by specialized professionals.

CONCLUSIONS

Free-tier LLMs can generate clinically plausible responses in metabolic dietetics but remain insufficiently accurate, complete, and consistent for autonomous use in IMD dietary management. Their performance appears acceptable for general information retrieval and potentially for standardized educational support; however, it declines in domains requiring individualized interpretation, guideline integration, monitoring, and risk-sensitive decision-making. LLMs should therefore be considered assistive tools and used only under specialist supervision, as current models are not yet suitable for autonomous clinical use and can be dangerous to use without a healthcare professional's oversight. Future work should define which levels of dietary education and counseling may be safely delegated to AI and should prioritize domain-specific, retrieval-enhanced systems grounded in validated metabolic nutrition resources.

Acknowledgements:
We thank the Dietetic Working Group of the Italian Society for the Study of Hereditary Metabolic Diseases and Newborn Screening (SIMMESN).

Artificial Intelligence-Assisted Technologies:
Artificial intelligence-assisted technologies were used for language editing purposes, including grammar and syntax correction and improvement of readability.

Authors' Contributions:
Study conception and design: M.T., A.B., A.D.; acquisition, collection and interpretation of data: M.M., G.G., A.P., M.S.G., S.P., A.H., V.P., A.F., A.T.; manuscript drafting: M.T. and M.M.; manuscript editing: M.T., M.M., G.G., A.P., M.S.G., S.P., A.H., V.P., A.F., A.T., A.B. and A.D.; supervision: M.T. and A.D.; approval to submit: all the authors have approved the manuscript.

Availability of Data and Material:
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest:
The authors declare that they have no conflict of interest to disclose.

Ethics approval:
Not applicable.

Funding:
No funding was received for this study.

Informed Consent:
Not applicable.

ORCID ID:
Martina Tosi: 0000-0003-4859-5288
Marta Muscarella: 0009-0005-6203-5940
Giorgia Gugelmo: 0000-0002-3557-8078
Alex Pinto: 0000-0002-8206-0829
Marta Suárez González: 0000-0001-7552-4894
Soraia Poloni: 0000-0003-4506-6487
Alexander Höller: 0000-0002-8541-5904
Veronica Perico: 0000-0001-7741-0874
Aurora Favaro: 0009-0002-0624-6418
Alessandra Tavian: 0009-0006-0005-9661
Andrea Belvedere: 0009-0005-2539-8187
Alice Dianin: 0009-0000-9539-2649

REFERENCES

1. Ponzo V, Goitre I, Favaro E, Merlo FD, Mancino MV, Riso S, Bo S. Is ChatGPT an Effective Tool for Providing Dietary Advice? Nutrients 2024; 16: 469.
2. Bean AM, Payne RE, Parsons G, Kirk HR, Ciro J, Mosquera-Gómez R, Hincapié M S, Ekanayaka AS, Tarassenko L, Rocher L, Mahdi A. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nat Med 2026; 32: 609-615.
3. Kirk D, van Eijnatten E, Camps G. Comparison of Answers between ChatGPT and Human Dieticians to Common Nutrition Questions. J Nutr Metab 2023; 2023: 5548684.
4. Gugelmo G, Chesini F, Lenzini L, Vitturi N, Girelli D, Marchi G. Potential applications of digital health technologies in adults with inherited metabolic disorders. JIM 2025; 2: e1003.
5. Southey NL, Zhu R, Holscher HD. Machine Learning and Artificial Intelligence in Nutrition Research: Analytical Methods, Applications, and Key Considerations. J Nutr 2026; 156: 101528.
6. O'Hara C, Kent G, Leydon CL, Walsh NM, Gibney ER, Skoutas D, Flynn AC, Timon CM. Language models in nutrition and dietetics: a scoping review. Am J Clin Nutr 2026; 123: 101127.
7. van Wegberg AMJ, MacDonald A, Ahring K, Bélanger-Quintana A, Beblo S, Blau N, Bosch AM, Burlina A, Campistol J, Coşkun T, Feillet F, Giżewska M, Huijbregts SC, Leuzzi V, Maillot F, Muntau AC, Rocha JC, Romani C, Trefz F, van Spronsen FJ. European guidelines on diagnosis and treatment of phenylketonuria: First revision. Mol Genet Metab 2025; 145: 109125.
8. Chinsky JM, Singh R, Ficicioglu C, van Karnebeek CDM, Grompe M, Mitchell G, Waisbren SE, Gucsavas-Calikoglu M, Wasserstein MP, Coakley K, Scott CR. Diagnosis and treatment of tyrosinemia type I: a US and Canadian consensus group review and recommendations. Genet Med 2017; 19.
9. Boy N, Mühlhausen C, Maier EM, Heringer J, Assmann B, Burgard P, Dixon M, Fleissner S, Greenberg CR, Harting I, Hoffmann GF, Karall D, Koeller DM, Krawinkel MB, Okun JG, Opladen T, Posset R, Sahm K, Zschocke J, Kölker S; Additional individual contributors. Proposed recommendations for diagnosing and managing individuals with glutaric aciduria type I: second revision. J Inherit Metab Dis 2017; 40: 75-101.
10. Frazier DM, Allgeier C, Homer C, Marriage BJ, Ogata B, Rohr F, Splett PL, Stembridge A, Singh RH. Nutrition management guideline for maple syrup urine disease: an evidence- and consensus-based approach. Mol Genet Metab 2014; 112: 210-217.
11. Häberle J, Burlina A, Chakrapani A, Dixon M, Karall D, Lindner M, Mandel H, Martinelli D, Pintos-Morell G, Santer R, Skouma A, Servais A, Tal G, Rubio V, Huemer M, Dionisi-Vici C. Suggested guidelines for the diagnosis and management of urea cycle disorders: First revision. J Inherit Metab Dis 2019; 42: 1192-1230.
12. Forny P, Hörster F, Ballhausen D, Chakrapani A, Chapman KA, Dionisi-Vici C, Dixon M, Grünert SC, Grunewald S, Haliloglu G, Hochuli M, Honzik T, Karall D, Martinelli D, Molema F, Sass JO, Scholl-Bürgi S, Tal G, Williams M, Huemer M, Baumgartner MR. Guidelines for the diagnosis and management of methylmalonic acidaemia and propionic acidaemia: First revision. J Inherit Metab Dis 2021; 44: 566-592.
13. Welling L, Bernstein LE, Berry GT, Burlina AB, Eyskens F, Gautschi M, Grünewald S, Gubbels CS, Knerr I, Labrune P, van der Lee JH, MacDonald A, Murphy E, Portnoi PA, Õunap K, Potter NL, Rubio-Gozalbo ME, Spencer JB, Timmers I, Treacy EP, Van Calcar SC, Waisbren SE, Bosch AM; Galactosemia Network (GalNet). International clinical guideline for the management of classical galactosemia: diagnosis, treatment, and follow-up. J Inherit Metab Dis 2017; 40: 171-176.
14. Úbeda F, Santander S, Luesma MJ. Clinical Practice Guidelines for the Diagnosis and Management of Hereditary Fructose Intolerance. Diseases 2024; 12: 44.
15. John TA, Anastasopoulou C. Glycogen Storage Disease. [Updated 2025 Jan 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459277/
16. LoPiccolo MK, Vockley J. Carnitine Palmitoyltransferase II Deficiency. 2004 Aug 27 [Updated 2026 Apr 9]. In: Adam MP, Bick S, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1253/
17. Van Calcar SC, Sowa M, Rohr F, Beazer J, Setlock T, Weihe TU, Pendyal S, Wallace LS, Hansen JG, Stembridge A, Splett P, Singh RH. Nutrition management guideline for very-long chain acyl-CoA dehydrogenase deficiency (VLCAD): An evidence- and consensus-based approach. Mol Genet Metab 2020; 131: 23-37.
18. Prasun P, LoPiccolo MK, Ginevic I. Long-Chain Hydroxyacyl-CoA Dehydrogenase Deficiency/Trifunctional Protein Deficiency. 2022. In: Adam MP, Bick S, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583531/
19. Shvetsova O, Katalshov D, Lee SK. Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment. Appl Sci 2025; 15: 7298.
20. Hong J, Kikuta NT, Simos A, Tsai S, Lin B, Rodriguez F, Palaniappan L. Testing the Appropriateness of Diabetes Prevention and Care Information Given by the Online Conversational AI ChatGPT. Clin Diabetes 2023; 41: 549-552.
21. Huang C, Chen L, Huang H, Cai Q, Lin R, Wu X, Zhuang Y, Jiang Z. Evaluate the accuracy of ChatGPT's responses to diabetes questions and misconceptions. J Transl Med. 2023; 21: 502.
22. Naja F, Taktouk M, Matbouli D, Khaleel S, Maher A, Uzun B, Alameddine M, Nasreddine L. Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome. Eur J Clin Nutr 2024; 78: 887-896.
23. Karakas PE, Calik A, Bilen AB, Kandemir K, Alphan ME. Large Language Models as Clinical Nutrition Decision Tools: Quantitative Bias and Guideline Deviation in Type 2 Diabetes Meal Planning. Healthcare (Basel) 2026; 14: 739.
24. Qarajeh A, Tangpanithandee S, Thongprayoon C, Suppadungsuk S, Krisanapan P, Aiumtrakul N, Garcia Valencia OA, Miao J, Qureshi F, Cheungpasitporn W. AI-Powered Renal Diet Support: Performance of ChatGPT, Bard AI, and Bing Chat. Clin Pract 2023; 13: 1160-1172.
25. Kairat M, Adilmetova G, Ibraimova I, Gaipov A, Varol HA, Chan MY. Benchmarking ChatGPT and Other Large Language Models for Personalized Stage-Specific Dietary Recommendations in Chronic Kidney Disease. J Clin Med 2025; 14: 8033.
26. Aydin Cil M, Karatas N, Mustafaoglu O, Sevinc C. Comparison of AI-generated renal diets by different large language models: a guideline-based evaluation with expert input. BMC Nephrol 2026; 27: 133.
27. Bertin L, Branchi F, Ciacci C, Lee AR, Sanders DS, Trott N, Zingone F. Efficacy of Large Language Models in Providing Evidence-Based Patient Education for Celiac Disease: A Comparative Analysis. Nutrients 2025; 17: 3828.

28. Peled N, Shouval DS, Gillett P, Szajewska H, Valitutti F, Shamir R, Guz-Mark A. Performance of large language models in answering frequently-asked questions on celiac disease. J Pediatr Gastroenterol Nutr 2026; 82: 1072-1077.
29. Yolcu F, Koç Yekedüz M, Köse E, Gülhan Samur F, Eminoğlu FT. Evaluation of scientific validity and appropriateness of artificial intelligence-assisted ChatGPT advices in dietary treatment of methylmalonic acidemia. Nutr Hosp 2025; 42: 910-915.
30. Liu M, Okuhara T, Shirabe R, Nishiie Y, Chang X, Okada H, Kiuchi T. Validity and reliability of ChatGPT's responses on dietary supplements in Japan: A quality assessment and content analysis. PEC Innov 2026; 8: 100461.
31. Ferraris AF, Audrito D, Di Caro L, Poncibò C. The architecture of language: Understanding the mechanics behind LLMs. Cambridge Forum on AI: Law and Governance 2025; 1: e11.
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv [preprint]. 2017. Available from: https://doi.org/10.48550/arXiv.1706.03762
33. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. Engineering 2023; 25: 51-65.
34. Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive mirage: A review of hallucinations in large language models. arXiv [preprint]. 2023. Available from: https://doi.org/10.48550/arXiv.2309.06794
35. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023; 27: 120.
36. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering in large language models: A comprehensive review. Patterns 2025; 6.
37. Wang D, Zhang S. Large language models in medical and healthcare fields: Applications, advances, and challenges. Artif Intell Rev 2024; 57: 299.
38. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, Hu S, Chen Y, Chan CM, Chen W, Yi J, Zhao W, Wang X, Liu Z, Zheng HT, Chen J, Liu Y, Tang J, Li J, Sun M. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell 2023; 5: 220-235.
39. Abdur Rahman L, Dedousis V, Papathanail I, Poursoleymani R, Kafyra M, Kalafati IP, Mougiakakou SG. Generative AI in Precision Nutrition: A Review of Current Developments and Future Directions. Nutrients 2026; 18: 938.
40. Gjorgjevikj A, Martinc M, Cenikj G, Stojanov R, Drole J, Ispirova G, Menichetti G, Ogrinc N, Trajanov D, Džeroski S, Koroušić Seljak B, Eftimov T. Large language models in food and nutrition science: Opportunities, challenges, and the case of FoodyLLM. Curr Res Food Sci 2026; 12: 101351.

JIM
JIM 2026; 3 (3): e1215
DOI: 10.61012/JIM_202608_1215
Category: *Review*

MIND THE BIAS: MAPPING AND MITIGATING COGNITIVE ERRORS ACROSS THE EXPANDED NEWBORN SCREENING PATHWAY

S. RUSSO[1], S. GASPERINI[2], V. CRESCITELLI[2], M.C. FARAGUNA[1,2], R. PRETESE[2], C. BONFANTI[2], S. GIGANTE[2], K. POZZI[2], A. CERIZZA[2], G. LA MARCA[3], M. BANI[1]

• • •

[1]School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy
[2]Pediatrics, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy
[3]Newborn Screening, Clinical Biochemistry and Clinical Pharmacy Laboratory, Meyer Children's Hospital IRCCS, Florence, Italy

Corresponding Author
Marco Bani, Ph.D; e-mail: marco.bani1@unimib.it

ABSTRACT – Expanded newborn screening (ENBS) has substantially improved the early identification of inherited metabolic diseases, enabling timely interventions that have transformed the prognosis of many affected infants. However, the effectiveness of ENBS relies on a complex multidisciplinary pathway extending well beyond laboratory analysis, involving healthcare professionals with diverse roles as well as families who must understand, interpret, and respond to screening information. At every stage of this process, cognitive biases and heuristics may influence perception, judgment, communication, and decision-making, potentially affecting diagnostic accuracy, follow-up, and family experience. This narrative case-based work examines the role of cognitive biases throughout the ENBS pathway, from parental information provided during pregnancy and dried blood spot collection to laboratory interpretation, communication of positive or uncertain results, confirmatory testing, dietary management, and long-term clinical follow-up. Particular attention is devoted to the different professionals involved – including nurses, midwives, laboratory scientists, metabolic physicians and dietitians – as well as to the cognitive processes that may shape parental understanding, information seeking, and healthcare decisions after positive screening results. The work also discusses practical examples of common cognitive biases encountered in ENBS and explores evidence-based strategies to mitigate their impact, including reflective practice, standardized communication, multidisciplinary collaboration, and cognitive debiasing interventions. Recognizing the influence of cognitive biases represents an important opportunity to strengthen diagnostic reasoning, improve communication with families, reduce avoidable errors, and promote more equitable and patient-centered newborn screening programs.

Keywords: Cognitive bias, ENBS, Medical decision making, Families, Metabolist, Dietitian, Nurse.

INTRODUCTION

Expanded Newborn Screening (ENBS) is a comprehensive, integrated, and multidisciplinary secondary prevention program aimed at the population-wide early detection of rare congenital disorders, including metabolic and genetic conditions, in newborns. By integrating screening, confirmatory diagnosis, timely therapeutic intervention, and long-term follow-up, ENBS seeks to improve clinical

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

outcomes and reduce disease-related morbidity and mortality[1,2]. ENBS involves a complex multistep process requiring the coordinated efforts of numerous healthcare professionals, including pediatricians, nurses, psychologists, dietitians, and laboratory specialists, as well as multiple interactions with families throughout the screening pathway. These interactions occur during the provision of information about ENBS, blood spot sample collection, recall procedures, confirmatory testing, and subsequent clinical follow-up. Each stage of the screening process involves decisions and judgments by both healthcare professionals and families, which may be influenced by cognitive biases and heuristics.

Cognitive biases are systematic and predictable deviations from rational judgment arising from heuristic information processing[3,4]. Widely studied across psychology and medicine, these biases are known to affect diagnostic reasoning, risk assessment, and clinical communication[5].

In healthcare, these biases can significantly influence clinical reasoning, with potential significant consequences on diagnostic accuracy, patient communication, and healthcare outcomes.

When discussing this pedagogical theory, one cannot fail to mention Daniel Kahneman and, in particular, his seminal 2011 work 'Thinking, Fast and Slow'[6]. According to Dual-Process Theory, clinical reasoning relies on two interacting systems: a fast, intuitive, heuristic-based system (System 1) and a slower, analytic, effortful system (System 2)[7,8]. While System 1 enables efficient pattern recognition, it is also particularly vulnerable to cognitive biases, especially under conditions of uncertainty and time pressure.

Applying the Dual-Process Theory to the medical context, some authors[9] have argued that System 1 is more likely involved in generating multiple diagnostic hypotheses and pattern recognition of clustered symptoms; conversely, System 2 plays a central role in the systematic search for additional information from a history and physical or lab exams.

More than 100 biases have been identified and described in the literature, but only some are specifically related to the medical field. The majority pertain to patients' or clinicians' therapeutic choices, whereas only a few focus on diagnostic reasoning[10].

In healthcare, evidence indicates that cognitive biases are a major contributor to diagnostic failures; they can distort diagnosis, risk assessment, and communication, leading to mismanagement or delays[11,12] and cause patient harm in healthcare facilities[13,14]. Cognitive biases seem to be implicated in over 50% of diagnostic errors in ambulatory care[15] and up to 83% of physician-reported cases, with failures of judgment (rather than lack of knowledge) being the primary cause in most malpractice claims related to misdiagnosis[16-18]. Healthcare professionals often underestimate how much such mental shortcuts shape their decisions, which helps explain the persistence of cognitive-process failures despite growing medical expertise and technology.

To date, however, no studies have systematically explored how, when, and for whom cognitive biases may affect the Expanded Newborn Screening (ENBS) pathway.

Awareness of cognitive biases is critical in the context of ENBS for metabolic and genetic disorders because true positives and false positives often cause parental anxiety[19-21], while false negatives can delay treatment; furthermore, cognitive biases can influence each actor's decisions.

Previous studies have examined cognitive biases in several clinical contexts, including the emergency department[17,22], intensive care unit[23], oncology[24], internal medicine[25], chronic diseases[26], vaccination hesitancy[27], laboratory results interpretation[28], and child physical abuse[29].

The ENBS represents a particularly compelling setting in which to examine these mechanisms. It combines a highly standardized diagnostic workflow with emotionally salient decision points while engaging multiple stakeholders – including pediatricians, laboratory biochemists, metabolic specialists, nurses, and families – who collectively contribute to a process of distributed cognition[30,31]. This complex sociotechnical environment creates multiple opportunities for cognitive biases to influence clinical reasoning, communication, and decision-making across the screening pathway, from the interpretation of biochemical results to confirmatory testing, diagnostic disclosure, and long-term follow-up.

This study employs a narrative, case-based qualitative design to map the primary cognitive biases that may affect decision-making and communication across the ENBS pathway. The objective was not to estimate prevalence or causal effects, but to develop a clinically grounded conceptual framework that illustrates how cognitive biases operate in real-world ENBS scenarios and how their impact may be mitigated.

METHODS

Cognitive biases were identified through a narrative review of the psychological and medical literature on heuristics and biases in clinical decision-making. Biases were selected based on three criteria: (a) relevance to healthcare contexts, (b) plausibility of occurrence within ENBS workflows, and (c) potential impact on diagnostic reasoning, communication with families, or clinical management. The final set of biases reflects those most frequently discussed in the literature and most applicable to ENBS, as identified by the authors' multidisciplinary team.

For each cognitive bias, one vignette was developed to illustrate realistic scenarios occurring at different stages of the ENBS pathway and involving different stakeholders (metabolic specialists, nurses, laboratory scientists, dietitians, and families). Vignettes were constructed by the authors based on their clinical and research experience in neonatal screening, metabolic medicine, and health psychology. All cases are fictional but are intended to reflect common or plausible clinical situations.

All clinical vignettes were independently reviewed by a multidisciplinary panel of expert clinicians, including metabolic specialists, dietitians, nurses, laboratory scientists, and clinical psychologists. Experts assessed each vignette for clinical plausibility, internal coherence, and relevance to real ENBS practice. Revisions were made iteratively until consensus was reached regarding the accuracy and educational value of each scenario.

The analytic process consisted of mapping each vignette to specific cognitive biases and examining their potential impact on clinical reasoning, communication, and workflow. The analysis was conducted collaboratively by the authors and resulted in stakeholder-specific thematic groupings presented in the following sections.

In Table 1, we list the selected cognitive biases that may affect the ENBS process, and in the following sections, we present stakeholder-specific scenarios – metabolic specialists, nurses, clinical biochemists, dietitians, and families – in which these biases and reasoning errors are illustrated and their potential impacts analyzed.

Table 1. Cognitive biases and heuristics involved in the ENBS process.

Representativeness heuristic	A mental shortcut whereby the likelihood of an event is judged by how closely it resembles the typical features of its parent category or prototype.
Framing effect bias	The way information is framed affects how individuals (clinicians and parents alike) receive and process it.
Negativity bias	It is the tendency to accord greater emotional and cognitive weight to negative events than to neutral or positive events of equal intensity.
Anchoring bias	The tendency to rely too heavily on an initial reference point (whether numeric or contextual) when making subsequent judgments or decisions, even if that anchor is irrelevant.
Availability heuristic	The tendency to judge the probability or importance of an event by how readily examples of it come to mind, often leading to over-weighting recent, frequent, or emotionally salient instances.
Bandwagon effect	The tendency to adopt certain beliefs, behaviors, or styles because many others are doing so.
Confirmation bias	The tendency to seek out, interpret, and remember information in a way that confirms one's pre-existing beliefs or expectations, while overlooking or downplaying contradictory evidence.
Overconfidence effect	The tendency to display confidence in your own judgements that exceeds the actual accuracy of those judgements, manifesting as overestimation of performance, overplacement relative to others, or unwarranted precision in one's beliefs.
Attribution bias (fundamental attribution error)	The tendency to explain others' behaviour or outcomes chiefly in terms of their dispositions or intentions, while underestimating situational or contextual factors.

Continued

Table 1 *(continued).* Cognitive biases and heuristics involved in the ENBS process.

Routine bias (automation)	The tendency to follow familiar routines automatically, even when changing circumstances call for a different approach. This is not a classic bias but reflects the neglect of vigilance after many similar tasks.
Hindsight bias	The tendency, after an event has occurred, to overestimate the extent to which the outcome could have been foreseen. It is colloquially known as the "I knew it all along phenomenon."
Premature closure	The tendency to stop considering alternative diagnoses once an initial plausible explanation is reached, leading to incomplete evaluation and potential diagnostic errors.
Outcome bias	The tendency to judge the quality of a decision by its eventual outcome rather than by the information and reasoning available when it was made.
Status quo bias	Preference for maintaining existing practices, even when new information suggests change is warranted.
Certainty effect (zero-risk bias)	The tendency to prefer options that completely eliminate a specific risk, overweighting outcomes that are certain relative to those that are merely probable, even when the probabilistic option would produce a greater overall reduction in expected harm.
Curse of knowledge bias	Tendency to assume that others have the same level of expertise or know-how as we do. We forget what it was like to be new to something and how difficult it was to learn.
Dunning-Kruger effect	The tendency of clinicians with insufficient knowledge or experience to overestimate their diagnostic competence, leading to excessive confidence in incorrect clinical judgments

METABOLIC SPECIALISTS

Metabolic specialists are physicians – typically pediatricians with specialized training in inherited metabolic diseases and metabolic genetics – who receive and interpret abnormal ENBS results. They are responsible for the timely clinical assessment of the newborn, the initiation and interpretation of confirmatory investigations, coordination of the diagnostic and therapeutic pathway within a preventive medicine framework, and the clear communication of diagnostic findings and management plans to both the multidisciplinary clinical team and the infant's family. As the primary communicators of true positives, false positives, or variants of uncertain significance (VUS), metabolic specialists guide families through each subsequent step. These tasks demand the management and interpretation of complex biochemical data, occasionally nonspecific clinical signs, and evolving prognoses, all of which contribute to a high cognitive load. Consequently, metabolic specialists must remain vigilant to a range of cognitive biases, including anchoring bias, confirmation bias, framing effect, availability bias, overconfidence, and attribution bias (see Supplementary Table 1 for clinical examples), especially when communicating uncertain results to parents.

Clinical reasoning may also be compromised by implicit gender-related biases[32,33]. Recently, Lenzini et al[34] described a woman's perspective on inherited metabolic diseases, suggesting that some disparities in approach may be related to clinicians' or patients' gender, including symptom interpretation, diagnostic process, and the prioritization of further testing. Some of these differences can be explained by cognitive biases, specifically stereotyping, attribution, and anchoring biases.

Owing to stereotyping bias, clinicians may implicitly associate particular symptoms with a given gender, leading to under- or over-investigation. For example, an adolescent girl with a urea cycle disorder with subtle neurocognitive or somatic symptoms might have those signs hastily attributed to anxiety or an eating disorder, delaying appropriate metabolic investigations and timely treatment. By contrast, owing to attribution bias, the same clinical picture may be explained differently depending on the patient's sex: symptoms in males are more readily ascribed to biological causes, while similar symptoms in females are more likely to be viewed as psychosocial or emotional. For example, mild neurological signs in a male child may prompt immediate neurological testing, whereas similar signs in a female child may

first prompt assessment of family dynamics or emotional health, thereby delaying metabolic work-up. Owing to anchoring bias, initial gendered assumptions can lock clinicians' thinking in place and make them less responsive to new or contradictory evidence. For example, developmental delays in girls may be normalized or dismissed for longer, whereas similar concerns in boys might prompt earlier investigation, a pattern that delays diagnosis and intervention for affected girls.

NURSES AND MIDWIVES

Nurses and midwives play a pivotal role in ENBS by collecting newborn dried blood spot samples and often serving as the first point of contact for parents, whether to explain the screening procedure or to communicate the need for repeat sampling or preliminary positive results. Although empirical evidence specifically addressing cognitive biases among nurses and midwives in the ENBS setting remains limited, a recent scoping review[35] of nursing decision-making found that nurses are susceptible to cognitive biases, particularly confirmation and anchoring biases, similarly to physicians. In addition, the highly standardized and repetitive nature of newborn screening may increase susceptibility to complacency bias, while several other cognitive biases may also influence clinical judgment and communication in this professional group (Supplementary Table 2).

BIOCHEMISTS, GENETICISTS (LABORATORY SPECIALISTS)

Clinical biochemists, geneticists, and laboratory scientists perform the biochemical and molecular analyses underpinning ENBS and conduct confirmatory investigations when required. Their responsibilities include selecting appropriate confirmatory tests and interpreting borderline biochemical findings or genetic variants, both of which are essential for establishing an accurate diagnosis and guiding subsequent clinical management. As with other professionals involved in the screening pathway, these judgments are susceptible to cognitive biases (Supplementary Table 3), which may influence test selection, data interpretation, and the reporting of laboratory findings.

DIETITIANS

Within ENBS, dietitians play a pivotal role in translating biochemical and clinical findings into individualized dietary management plans, particularly for disorders such as phenylketonuria (PKU). However, clinical decision-making in this setting often occurs under conditions of diagnostic uncertainty and emotional pressure from families, making dietitians susceptible to cognitive biases that may influence nutritional assessment, treatment recommendations, and patient counseling[36]. Several cognitive biases relevant to dietetic practice in ENBS are summarized in Supplementary Table 4.

FAMILIES RECEIVING POSITIVE OR FALSE-POSITIVE ENBS RESULTS

Parents and family members are central stakeholders in the ENBS pathway, and the service should be grounded in their informational and emotional needs. On receiving a positive or borderline screening result, families undergo rapid emotional and cognitive processing; even when a result later proves to be a false positive, the experience can remain intensely distressing[20,21,37]. In such high-stress, uncertain moments, fast, intuitive System 1 thinking tends to dominate, making parents particularly vulnerable to heuristics and biases (Supplementary Table 5). These processes influence how families seek, interpret, respond to, and later remember information, so recognizing them is essential for clinicians and other health professionals who must tailor clear, empathetic communication and decision-support.

A related and frequently overlooked consequence is the psychosocial and logistical burden of caring for a child with an inherited metabolic disorder, which often falls disproportionately on mothers[38,39]. Clinicians' assumptions about caregiving roles can implicitly reinforce this imbalance and obscure the needs of other family members. For example, owing to gender-role expectations bias, healthcare professionals may default to addressing mothers more directly, providing instructions and follow-up recommendations primarily to them, even when fathers or other caregivers are actively involved. Similarly,

affective bias can lead clinicians to offer more emotional support to mothers, thereby under-recognizing paternal stress or involvement. Fathers who show signs of emotional fatigue may be less likely to be offered psychological support and may feel less authorized to voice their difficulties. Finally, confirmation bias may lead clinicians to interpret maternal fatigue in a way that confirms their expectation that such symptoms are a normal consequence of caregiving, rather than considering the possibility of significant psychological distress or caregiver burden. As a result, reports of persistent sleep deprivation may be acknowledged without further assessment or referral for supportive services.

MITIGATION STRATEGIES: TRAINING AND EDUCATIONAL INTERVENTIONS

ENBS is a clinical context in which integrated collaboration among professionals can help develop a broader perspective on decision-making processes, thereby correcting most cognitive biases. Although evidence regarding the effectiveness is limited, instruction in metacognition, reflective practice, cognitive bias awareness, and interdisciplinary collaboration has been proven to improve diagnostic accuracy and emotional support for families in healthcare settings[12,14,40,41]. At the individual level, teaching healthcare staff to slow down, reflect, and use decision aids or checklists can reduce the impact of bias[42]. Cognitive bias awareness is the first step to mitigating them[42]. Education and training programs can help clinicians (and families) become aware of common heuristics and adopt strategies to counteract them. For clinicians, curricula increasingly include modules on clinical reasoning and cognitive bias. A recent scoping review found that targeted training can improve healthcare providers' recognition of biases[43]. Existing training focuses on increasing awareness and self-reflection about explicit and implicit biases and their harmful effects on patient care[44], as well as on acquiring strategies to manage biases in real-time, promoting perspective-taking to understand patients' unique viewpoints and challenges and to regulate emotions that can contribute to biased decisions[12].

Training should be tailored to the clinician's level of experience. For junior or less experienced clinicians, educational interventions should emphasize structured decision-making frameworks to reduce inappropriate deference to authority and promote confident, evidence-based decision-making. For senior clinicians, training should focus on peer case review, reflective practice, and deliberate reconsideration of diagnostic conclusions despite extensive experience, while reinforcing the consistent use of evidence-based tools and clinical algorithms even when intuitive judgment suggests an alternative course of action.

Educational interventions for healthcare professionals should be designed to fit the demands of busy clinical practice. Short, case-based workshops on common heuristics and cognitive biases, complemented by microlearning modules, can increase awareness while providing practical strategies to recognize and mitigate cognitive traps. Simulation-based training, particularly for communicating uncertain or distressing results and managing challenging specimen collection scenarios (e.g., dehydrated or very small neonates), allows practitioners to develop calibrated responses in realistic clinical settings. In addition, confidence-calibration exercises – where clinicians record their diagnostic confidence and subsequently compare it with actual outcomes – can improve metacognitive awareness, reduce overconfidence, and enhance clinical judgment over time.

Mitigating cognitive errors across the ENBS pathway requires a deliberate shift from ad hoc, intuition-led practice to predictable, data-driven processes. The imperative for system-level solutions was set out in the seminal patient-safety reports and has been reinforced by work on diagnostic error: moving from individual heroics to reliable systems reduces avoidable harm[45,46]. At a service level, this means designing workflows so that routine judgments are supported by objective rules and decision aids rather than left to memory or momentary impressions. Simple, standardized procedures that set out the steps from initial screen to confirmatory testing and, where appropriate, treatment, make practice repeatable and reduce scope for individual judgment to be distorted by recent experiences or emotional states.

Equally important is the principle of "designing for the human factor": accept that people are fallible and build prompts, checks, and rapid feedback into everyday tasks. This can be achieved, for example, through structured reporting templates that require clinicians and laboratory staff to record alternative explanations, the degree of certainty, and explicit next steps, or through Laboratory Information Management System/Electronic Health Record prompts that flag equivocal results and require a short justification before escalation[47]. By making the expected steps visible and unavoidable in the record, multidisciplinary teams reduce the chance that a single striking data point will dominate decision-making.

Furthermore, low-cost usability interventions and "nudges" can sustain behavioral change[48]. Visual reminders in collection areas, brief on-screen prompts in the Electronic Health Record, default follow-up

schedules, and ready-to-use reporting templates reduce cognitive load and make the right action the easy action[49]. Implementation should start small: one protocol, one teach-back requirement, one audit loop, and scale up iteratively, engaging frontline staff in co-design so that measures fit real workflows rather than adding burdensome tasks.

System changes may include deliberate procedural "pause points" and short checklists that serve as cognitive forcing functions, encouraging reflective, analytical review before irreversible actions[50-52]. Short time-outs (introduced before irreversible actions such as admission or major dietary changes) create space for verification, to check whether other causes have been considered, and to seek a second opinion when appropriate. These pause points work best when coupled with mandatory quality checks at critical junctures (for example, an objective checklist for blood-spot adequacy at collection and a pre-sign-off review in the laboratory). In practice, a brief, structured pause codified into the pathway will interrupt premature closure and anchoring and prompt a short multidisciplinary check for borderline or high-consequence findings.

Communicating with families should be regarded and delivered as a clinical intervention: how information is transmitted directly affects parental understanding, emotional response, and subsequent health actions, not merely the "soft" side of care[19,21,53].

When clinicians communicate unexpected or provisional results to families, they should begin with an empathetic acknowledgment of the family's likely distress, then provide balanced context and a clear, time-bound plan of next steps (what will be tested, when, and who will contact them). Protocols for delivering difficult news, such as the SPIKES framework, support opening with empathy and following a structured sequence to reduce harm and improve comprehension[54].

The way in which screening results are framed has an important influence on parental understanding and decision-making. Clinicians should avoid isolated or alarmist wording and instead emphasize the provisional nature of the findings while clearly explaining the subsequent diagnostic pathway. Risk communication should rely on simple absolute risks or natural frequencies, supported where appropriate by visual aids, as these formats improve comprehension and reduce misinterpretation compared with vague or relative expressions[55].

Because patients and families commonly forget much of what they are told in a single consultation, providing a one-page written or visual summary that recapitulates the result, the planned next steps, and a named contact improves retention and allows families to review information when they are less emotionally aroused. Empirical work documents poor recall of medical information and supports the routine use of written takeaways to reinforce verbal explanations[56]. Services should provide multimodal, clinician-vetted information such as brief leaflets, icon arrays, translated materials, and short videos.

Finally, the teach-back method – in which caregivers are asked to explain or demonstrate the care plan in their own words – is an evidence-based communication strategy recommended by major healthcare organizations and supported by systematic reviews[57] demonstrating improvements in patient comprehension, adherence, and safety. Documenting the use of teach-back can also help verify that information has been accurately understood rather than merely acknowledged[57].

More broadly, adopting a family-centered approach can help reduce the likelihood that parents' understandable anxiety will contribute to unnecessary or disproportionate interventions[58]. For families experiencing high levels of distress, timely psychosocial support from a clinical psychologist or other qualified mental health professional may provide additional benefit by facilitating coping, improving communication, and supporting informed decision-making.

The goal is not to eliminate heuristics (necessary and efficient tools for navigating a complex and overloading environment), but to normalize the idea that healthcare staff, like other individuals, are cognitively biased and to promote a culture of cognitive transparency. By combining awareness, education, structured checks, teamwork, and effective communication strategies, healthcare staff and families can reduce reasoning and decision-making errors and emotional distress during the ENBS process.

CONCLUSIONS

Cognitive biases have the potential to influence every stage of Expanded Newborn Screening, from specimen collection and laboratory interpretation to clinical decision-making and communication with families. Although clinical experience and pattern recognition remain indispensable in the diagnosis of rare diseases, they do not protect clinicians from systematic errors in reasoning. On the contrary, experience may reinforce predictable cognitive vulnerabilities – including anchoring, availability, and confirmation bias – that can contribute to over- or under-investigation, unnecessary interventions, avoidable parental distress, and inef-

ficient use of healthcare resources. Reducing the impact of cognitive bias requires a multifaceted approach that combines evidence-based communication, structured clinical workflows, and targeted educational interventions. Balanced communication of provisional screening results, teach-back techniques, standardized decision-support tools, diagnostic pause points, second-opinion review of high-impact or borderline cases, and regular audit with feedback represent complementary strategies that strengthen clinical reasoning while reducing reliance on intuition alone. Educational initiatives – including case-based learning, simulation, microlearning, and confidence-calibration exercises – can further promote metacognitive awareness and reflective practice. To maximize effectiveness, these interventions should be pragmatic, integrated into routine clinical workflows, and co-designed with frontline healthcare professionals. This work is based on fictional but expert-validated clinical vignettes and therefore provides a conceptual and educational framework rather than empirical evidence. Future research should evaluate the effectiveness of these mitigation strategies within real ENBS programs, investigate how families understand and respond to risk communication, and determine whether behavioral interventions – such as framing, salience, or priming – can improve comprehension, reduce distress, and support shared decision-making[59,60]. Ultimately, recognizing cognitive bias as an inherent feature of human decision-making rather than an individual failing is fundamental to improving the quality and safety of Expanded Newborn Screening. By integrating behavioral science with clinical expertise, ENBS programs can strengthen diagnostic accuracy, improve the family experience, and enhance the effectiveness and reliability of early detection of inherited metabolic disorders.

Expanded newborn screening has transformed the early detection of inherited metabolic diseases; reducing cognitive bias represents the next step toward improving the quality, consistency, and safety of this complex diagnostic pathway.

Artificial Intelligence-Assisted Technologies:

An artificial intelligence (AI)-based large language model (ChatGPT) was used solely to assist in drafting selected clinical vignettes and improving the language and readability of the manuscript. The AI system was not used for study design, data collection, data analysis, interpretation of results, or scientific decision-making. All AI-generated content was critically reviewed, revised, and verified by the authors, who assume full responsibility for the accuracy and integrity of the final manuscript.

Authors' Contributions:

Study conception and design: M.B., S.R.; writing – original draft: M.B., S.R.; writing – review and editing, manuscript preparation: S.R., S.Ga., V.C., M.C.F., R.P., C.B., S.Gi., K.P., A.C., G.L., M.B.

Availability of Data and Material:

Data sharing is not applicable to this article, as no datasets were generated or analyzed during the current study.

Conflicts of Interest:

The authors declare that they have no conflicts of interest to disclose.

Ethics Approval:

Ethical review and approval were not applicable due to the design of the study.

Funding:

No funding was received for this study.

Informed Consent:

Written informed consent was not obtained as the study does not include patient data.

ORCID ID:

Selena Russo: 0000-0002-3024-3316
Serena Gasperini: 0000-0001-8743-4407
Roberta Pretese: 0009-0006-6713-8030
Martha Caterina Faraguna: 0000-0003-3426-2749
Cristina Bonfanti: 0009-0000-4415-0324
Silvia Gigante: 0009-0007-4035-9017
Katia Pozzi: 0009-0000-7244-7272
Alessia Cerizza: 0009-0007-7902-7968
Giancarlo La Marca: 0000-0003-3319-7260
Marco Bani: 0000-0002-6500-6513

REFERENCES

1. Mak CM, Lee HC, Chan AYW, Lam CWK. Inborn errors of metabolism and expanded newborn screening: review and update. Crit Rev Clin Lab Sci 2013; 50: 142-162.
2. Ministero della Salute. Screening neonatale esteso per la prevenzione e la diagnosi precoce di malattie metaboliche ereditarie. Roma: Ministero della Salute; 2017. Available from: https://www.salute.gov.it/new/it/tema/salute-del-bambino-e-dell-adolescente/screening-neonatali/?paragraph=2 (Accessed July 1, 2026).
3. Kahneman D, Tversky A. Judgment under uncertainty: heuristics and biases. Science 1974; 185: 1124-1131.
4. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science 1981; 211: 453-458.
5. Gilovich T, Griffin D, Kahneman D, editors. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge: Cambridge University Press; 2002.
6. Kahneman D. Thinking, Fast and Slow. Macmillan, 2011.
7. Evans SB. Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 2008; 59: 255-278.
8. Evans SB, Stanovich KE. Dual-process theories of higher cognition: advancing the debate. Perspect Psychol Sci 2013; 8: 223-241.
9. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017; 92: 23-30.
10. Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making 2015; 35: 539-557.
11. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016; 16: 138.
12. Daniel M, Carney M, Khandelwal S, Merritt C, Cole M, Malone M, Hemphill RR, Peterson W, Burkhardt J, Hopson L, Santen SA. Cognitive debiasing strategies: a faculty development workshop for clinical teachers in emergency medicine. MedEdPORTAL 2017; 13: 10646.
13. Rogers JE, Hilgers TR, Keebler JR, Looke T, Lazzara EH. How to mitigate the effects of cognitive biases during patient safety incident investigations. Jt Comm J Qual Patient Saf 2022; 48: 612-616.
14. O'Sullivan ED, Schofield SJ. A cognitive forcing tool to mitigate cognitive bias: a randomized controlled trial. BMC Med Educ 2019; 19: 12.
15. Gandhi TK, Kachalia A, Thomas EJ, Puopolo AL, Yoon C, Brennan TA, Studdert DM. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. Ann Intern Med 2006; 145: 488-496.
16. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013; 173: 418-425.
17. Schiff GD, Hasan O, Kim S, Abrams R, Cosby K, Lambert BL, Elstein AS, Hasler S, Kabongo ML, Krosnjar N, Odwazny R, Wisniewski MF, McNutt RA. Diagnostic error in medicine: analysis of 583 physician-reported errors. Arch Intern Med 2009; 169: 1881-1887.
18. Okafor N, Payne VL, Chathampally Y, Miller S, Doshi P, Singh H. Using voluntary reports from physicians to learn from diagnostic errors in emergency medicine. Emerg Med J 2016; 33: 245-252.
19. Bani M, Russo S, Raggi E, Gasperini S, Motta S, Menni F, Furlan F, Cefalo G, Paci S, Banderali G, Marchisio P, Biondi A, Strepparava MG. Parents' experience of the communication process of positivity at newborn screening for metabolic diseases: a qualitative study. Child Care Health Dev 2023; 49: 961-971.
20. Bani M, Russo S, Gasperini S, Crescitelli V, Menni F, Furlan F, Tagliaferri F, Cefalo G, Paci S, Banderali G, Marchisio P, Biondi A, Strepparava MG. Prevalence and predictors of parental distress at the communication of positivity at newborn screening for metabolic diseases: an Italian longitudinal study. BMJ Paediatr Open 2024; 8: e003103.
21. Russo S, Greco B, Carcereri MS, Vissani S, Caviglia S. Long-term psychological support for families receiving communication of positivity for metabolic diseases at newborn screening. J Innate Metab 2024; 1: e452.
22. Ng M, Wong E, Sim GG, Heng PJ, Terry G, Yann FY. Dropping the baton: cognitive biases in emergency physicians. PLoS One 2025; 20: e0316361.
23. Smith JR. Cognitive bias: a potential threat to clinical decision-making in the neonatal intensive care unit. J Perinat Neonatal Nurs 2017; 31: 294-296.
24. Evans SB, Cain D, Kapur A, Brown D, Pawlicki T. Why smart oncology clinicians do dumb things: a review of cognitive bias in radiation oncology. Pract Radiat Oncol 2019; 9: e347-e355.
25. Loncharich MF, Robbins RC, Durning SJ, Soh M, Merkebu J. Cognitive biases in internal medicine: a scoping review. Diagnosis (Berl) 2023; 10: 205-214.
26. Savioni L, Triberti S. Cognitive biases in chronic illness and their impact on patients' commitment. Front Psychol 2020; 11: 579455.
27. Casigliani V, Menicagli D, Fornili M, Lippi V, Chinelli A, Stacchini L, Arzilli G, Scardina G, Baglietto L, Lopalco P, Tavoschi L. Vaccine hesitancy and cognitive biases: evidence for tailored communication with parents. Vaccine X 2022; 11: 100191.
28. Klatt EC. Cognitive factors impacting patient understanding of laboratory test information. J Pathol Inform 2023; 15: 100349.
29. Harmon KA, Chang TP, Imagawa KK, Schmidt AR, Pham PK, Nager AL. Expert consensus on cognitive biases affecting child physical abuse evaluations in pediatric emergency medicine: A modified delphi study. Child Prot Prac 2025; 4: 100080.
30. Boyle JG, Walters MR, Jamieson S, Durning SJ. Distributed cognition: theoretical insights and practical applications to health professions education. Medical Teacher 2023; 45: 1323-1333.
31. Wilson E, Daniel M, Rao A, Torre D, Durning S, Anderson C, Goldhaber NH, Townsend W, Seifert CM. A scoping review of distributed cognition in acute care clinical decision-making. Diagnosis (Berl) 2022; 10: 68-88.
32. Berner AM, Lund J, Saunders CL. Tackling the complexity of gender bias in primary care. Br J Gen Pract 2021; 71: 296-297.
33. Champagne-Langabeer T, Hedges AL. Physician gender as a source of implicit bias affecting clinical decision-making processes: a scoping review. BMC Med Educ 2021; 21: 171.

34. Lenzini L, Bianconi S, Gugelmo G, Gragnaniello V, Messerotti Benvenuti S, Fadini GP, Vitturi N. The multifaceted challenges faced by women in the field of inherited metabolic disorders. Orphanet J Rare Dis 2025; 20: 104.
35. Thirsk LM, Panchuk JT, Stahlke S, Hagtvedt R. Cognitive and implicit biases in nurses' judgment and decision-making: a scoping review. Int J Nurs Stud 2022; 133: 104284.
36. Hamersley ML, Mullan BA. A dual-process model to predict dietitians' intention to use evidence-based guidelines. J Hum Nutr Diet 2015; 28: 573-581.
37. Bani M, Russo S, Gasperini S, Crescitelli V, Menni F, Furlan F, Tagliaferri F, Cefalo G, Paci S, Banderali G, Marchisio P, Balduzzi A, Strepparava MG. Prevalence and predictors of parental distress at the communication of positivity at newborn screening for metabolic diseases: an Italian longitudinal study. Eur J Pediatr 2026; 185: 315.
38. Chiarotti F, Kodra Y, De Santis M, Bellenghi M, Taruscio D, Carè A, Petrini M. Gender and burden differences in family caregivers of patients affected by ten rare diseases. Ann Ist Super Sanita 2023; 59: 122-131.
39. Xu J, Bao H, Qi X, Wang J, Yin H, Shang C, Tan RL, Wu Q, Huang W. Family caregivers of rare disease: a survey on health-related quality of life in family caregivers for Gaucher disease patients in China. Mol Genet Genomic Med 2021; 9: e1760.
40. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ Qual Saf 2016; 25: 808-820.
41. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: still no easy answers. West J Emerg Med 2020; 21: 125-131.
42. Royce CS, Hayes MM, Schwartzstein RM. Teaching critical thinking: a case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. Acad Med 2019; 94: 187-194.
43. Thompson J, Bujalka H, McKeever S, Lipscomb A, Moore S, Hill N, Kinney S, Meng Cham K, Martin J, Bowers P, Gerdtz M. Educational strategies in the health professions to mitigate cognitive and implicit bias impact on decision making: a scoping review. BMC Med Educ 2023; 23: 455.
44. Fricke J, Siddique SM, Aysola J, Cohen ME, Mull NK. Healthcare Worker Implicit Bias Training and Education: Rapid Review. In: Making Healthcare Safer IV: A Continuous Updating of Patient Safety Harms and Practices. Rockville (MD): Agency for Healthcare Research and Quality (US); January 2024.
45. Balogh EP, Miller BT, Ball JR. Improving Diagnosis in Health Care. Washington (DC): National Academies Press; 2015.
46. Kohn LT, Corrigan JM, Donaldson MS, editors. To Err Is Human: Building a Safer Health System. Washington (DC): National Academies Press; 2000.
47. Rotter T, Kinsman LD, Alsius A, Scott SD, Lawal A, Ronellenfitsch U, Plishka C, Groot G, Woods P, Coulson C, Bakel LA, Sears K, Ross-White A, Machotta A, Schultz TJ. Clinical pathways for secondary care and the effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database Syst Rev 2025; 5: CD006632.
48. Wolf A, Sant'Anna A, Vilhelmsson A. Using nudges to promote clinical decision making of healthcare professionals: a scoping review. Prev Med 2022; 164: 107320.
49. Raban MZ, Gates PJ, Gamboa S, Gonzalez G, Westbrook JI. Effectiveness of non-interruptive nudge interventions in electronic health records to improve the delivery of care in hospitals: a systematic review. J Am Med Inform Assoc 2023; 30: 1313-1322.
50. Yale S, Cohen S, Bordini BJ. Diagnostic time-outs to improve diagnosis. Crit Care Clin 2022; 38: 185-194.
51. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, Schmidt HG. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. JAMA 2010; 304: 1198-1203.
52. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. Acad Med 2011; 86: 307-313.
53. Bucknall TK, Hutchinson AM, Botti M, McTier L, Rawson H, Hitch D, Hewitt N, Digby R, Fossum M, McMurray A, Marshall AP, Gillespie BM, Chaboyer W. Engaging patients and families in communication across transitions of care: an integrative review. Patient Educ Couns 2020; 103: 1104-1117.
54. Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP. SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. Oncologist 2000; 5: 302-311.
55. Wegwarth O, Gigerenzer G. The barrier to informed choice in cancer screening: statistical illiteracy in physicians and patients. Recent Results Cancer Res 2018; 210: 207-221.
56. Kessels RP. Patients' memory for medical information. J R Soc Med 2003; 96: 219-222.
57. Talevski J, Wong Shee A, Rasmussen B, Kemp G, Beauchamp A. Teach-back: a systematic review of implementation and impacts. PLoS One 2020; 15: e0231350.
58. Loutfy A, Zoromba MA, Mohamed MA, El-Gazar HE, Andargeery SY, El-Monshed AH, Van Belkum C, Ali AS. Family-centred care as a mediator in the relationship between parental nurse support and parental stress in neonatal intensive care units. BMC Nurs 2024; 23: 572.
59. Yoong SL, Hall A, Stacey F, Grady A, Sutherland R, Wyse R, Anderson A, Nathan N, Wolfendenet L. Nudge strategies to improve healthcare providers' implementation of evidence-based guidelines, policies and practices: a systematic review of trials included within Cochrane systematic reviews. Implement Sci 2020; 15: 50.
60. Sant'Anna A, Vilhelmsson A, Wolf A. Nudging healthcare professionals in clinical settings: a scoping review of the literature. BMC Health Serv Res 2021; 21: 543.