Envisioning the Future of Health Informatics and Digital Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI250052

The Relevance of General Intelligence Measurement in Deep Learning for Healthcare

Marko MILETIC[a] and Murat SARIYAR[a,1]
[a]*Bern University of Applied Sciences, Switzerland*
ORCiD ID: Murat Sariyar https://orcid.org/0000-0003-3432-2860

Abstract. The integration of artificial intelligence (AI) into medical informatics presents significant opportunities to enhance healthcare through data-driven diagnostics, predictive analytics, and personalized therapeutic recommendations. This paper examines the role of general intelligence in improving the effectiveness and adaptability of AI systems in complex clinical environments. We explore various levels of generalization – local, broad, and extreme – highlighting their respective contributions and limitations in healthcare. Local generalization provides robust assessments based on well-defined risk factors, while broad generalization allows for nuanced patient stratification across diverse populations. Extreme generalization, however, presents the greatest challenge, requiring AI systems to adapt to entirely new contexts without prior exposure. Despite advancements, existing metrics for assessing generalization difficulty remain inadequate, necessitating the development of new evaluation methodologies.

Keywords. Artificial Intelligence (AI), artificial general intelligence (AGI), Large Language Model (LLMs), predictive modeling

1. Introduction

The integration of artificial intelligence (AI) into healthcare offers valuable opportunities to enhance healthcare through data-driven diagnostics, predictive analytics, and personalized therapeutic recommendations [1, 2]. The effectiveness of these systems largely depends on their "intelligent" processing capacity. General intelligence, often referred to as the "g-factor" [3], describes a system's or individual's ability to process new information, generalize across contexts, and solve unknown problems (unknown unknowns). Leveraging this general intelligence factor in medical applications could significantly improve the effectiveness and adaptability of such systems, enhancing their performance in complex, often unpredictable clinical environments [4].

In this context, two critical questions arise: How can the g-factor be measured, and is its consideration necessary in healthcare? The psychometric literature documents numerous approaches to measuring general intelligence [5]. A seminal paper, "On Measuring Intelligence", synthesizes the relevant discussions on intelligence

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

measurement within the framework of artificial intelligence [6]. It emphasizes that, in practice, benchmark datasets such as SuperGlue are often utilized, alongside extensive human evaluations and white-box analyses, to assess the performance of AI systems [7]. This paper examines the significance general intelligence and associated notions in healthcare, explores existing methodologies for intelligence assessment, and addresses the challenges of implementing these metrics in clinical practice.

In the following section, we will present three forms of intelligence according to [6] and discuss how priors, experience and generalization difficulty factors in. Subsequently, we will discuss whether general intelligence is sufficient for all tasks or, conversely, whether it is unnecessary and that task-specific skills alone are adequate.

2. Methods

Intelligence can be assessed based on the extent to which observed phenomena can be generalized to unobserved ones. The lowest form of this assessment is known as local generalization, which involves successfully performing a well-defined task – such as distinguishing benign from malignant tumors using CT images – on new data (known unknowns). The primary characteristic of local generalization is robustness [8]. The next level is referred to as broad generalization, which demonstrates human-level ability within a single, yet expansive, activity domain. An example of this would be a surgical robot capable of performing a variety of different procedures. At the highest level is extreme generalization, characterized by the ability to adapt to unknown unknowns across a wide array of tasks and domains. The key aspect of this level is skill-acquisition efficiency, which emphasizes the capacity to rapidly acquire new skills without relying on prior knowledge or experience beyond a foundational core knowledge base.

Information processing systems can be conceptualized along a continuum, ranging from fully static systems to highly adaptive, data-driven systems. At one end of this spectrum are static systems that operate based entirely on predefined priors – fixed, hard-coded knowledge derived from established guidelines or expert input [9]. Examples include expert systems in clinical decision-making, which follow rigid algorithms rooted in medical guidelines without modification based on new data. At the other end are systems that function with minimal reliance on priors, instead learning primarily through exposure to large datasets. Such systems, like neural networks trained on extensive patient data, autonomously identify patterns and make predictions based on observed correlations and trends. While some level of prior knowledge is essential for any form of intelligence, true intelligence is characterized by an ability to operate beyond rigid dependence on pre-existing information. Intelligence, therefore, is not merely reflected in the capacity to improve at a skill with accumulated experience, but in the system's ability to adapt and generalize beyond specific, learned patterns.

In addition to priors and experiential knowledge, a third critical factor "generalization difficulty", is essential for assessing intelligence. Higher levels of generalization difficulty necessitate a correspondingly higher level of intelligence, as demonstrated through the successful resolution of tasks that demand adaptive and flexible solutions. A key distinction must be made between the intelligent system (IS) itself and a skill program, which addresses a particular task within a defined problem space. The IS should be capable of generating a skill program that extends beyond the specific situations it was trained on, generalizing across a wider array of scenarios within the broader situation space. This ability to generalize is not merely a function of

increasing training data; rather, it represents a distinct dimension integral to defining intelligence. For fair comparisons of generalization difficulty, it is essential to restrict the model's embedded knowledge to a foundational set, including core concepts such as object permanence, basic physical principles (e.g., cohesion), notions of agency and goal-directed behavior, the natural number system, elementary arithmetic, and basic topology [6]. In this framework, the intelligence of an IS can be considered a measure of its efficiency in acquiring skills across diverse tasks, with respect to its priors, accumulated experience, and the inherent challenge of generalization difficulty.

In the following section, we present a predictive model aimed at assessing the risk of developing diabetes. This model might incorporate established priors, such as recognized risk factors (e.g., age, family history, obesity). We will evaluate the three levels of generalization to determine which approach is most applicable.

3. Results

In the context of our predictive model for assessing the risk of developing diabetes, local generalization is evident in the model's ability to accurately evaluate known unknowns. This includes distinguishing between patients at risk for diabetes and those who are not, based on established criteria. For instance, when the model is applied to new patient data, it effectively utilizes well-defined risk factors to make reliable predictions. This robustness ensures consistent performance across similar scenarios, demonstrating its efficacy in identifying individuals who may benefit from preventive interventions derived from previously learned patterns. This application exemplifies one of the most prevalent use cases of artificial intelligence. Achieving accuracy above a certain threshold on test data generally justifies model implementation. The prior knowledge and experience are well-defined, and generalization is relatively easy, as no novel situations are included in the test data. In such cases, it is essential to ensure the model is not overfitted and appropriately reflects the data's complexity.

Broad generalization is a notable feature of the model’s ability to adapt its risk assessment capabilities across diverse patient populations and associated health conditions. For instance, the predictive model can be trained not only to evaluate the risk of diabetes but also to incorporate comorbidities such as metabolic syndrome and cardiovascular disease. By integrating a variety of data inputs – including lifestyle factors and socio-demographic information – the model significantly enhances its ability to generate nuanced risk profiles. This broader applicability allows healthcare providers to identify at-risk individuals across different demographic groups and tailor interventions accordingly, ultimately improving patient outcomes in various contexts. However, traditional machine learning models such as random forests or boosting are insufficient for this level of complexity; they require either an ensemble of methods or a neural network based on transformer architectures. In this scenario, the focus extends beyond mere accuracy; it necessitates the development of comprehensive metrics that evaluate how many new contexts can be addressed. Weighted accuracy alone is inadequate; instead, metrics should be designed around concepts such as contextual coverage, adaptability, and the balance between exploration and exploitation.

Extreme generalization in diabetes risk prediction differs from broad generalization due to a higher degree of generalization difficulty. This difficulty pertains to the model's ability to integrate new and often entirely unknown contexts or risk factors without having explicitly encountered them during training. While broad generalization refers to

the expansion of knowledge to similar or related domains within a known framework, extreme generalization challenges the model to function effectively in completely unfamiliar and unforeseen situations. A model capable of extreme generalization must be able to recognize unknown patterns and relationships that may lie outside existing medical knowledge, such as newly discovered genetic markers or unexpected environmental factors not included in standard risk assessments. Hence, the model must possess sufficient flexibility to formulate hypotheses and identify patterns based on a small core of knowledge in entirely new domains.

There are no satisfactory measurement methods for such an extreme generalization. Chollet proposed the "Abstraction and Reasoning Corpus (ARC)", which, according to his assertion, cannot be solved purely through example-based learning [6]. This is where we diverge from Chollet's position. General intelligence is not orthogonal to experience. Through experience, genetic predispositions are epigenetically modified, resulting in the incorporation of knowledge that extends beyond direct experience into the structure of intelligence. While Large Language Models (LLMs) have not yet achieved human-level intelligence, they exhibit some astonishing emergent properties arising from exposure to vast amounts of data. Furthermore, human intelligence itself is never entirely independent of prior experiences, as our g factor is a product of ancestral priors and experiences. Returning to the measurement problem, we therefore propose utilizing the same metrics as those employed for broad generalization but with heightened expectations; the specific requirements will depend on the use case.

Although such advanced systems with general intelligence are not yet fully realized, some LLMs have demonstrated capabilities in medical examinations that surpass those of many medical students [10]. Additionally, numerous physicians are utilizing these systems as aids in their daily practice due to the impressive aggregation of information they provide. The challenge of measurement is becoming increasingly urgent, especially as we transition from the traditional medical device context, which typically focuses on specific, narrowly defined applications. Just as an individual cannot be classified as a medical device, a general artificial intelligence (AGI) cannot be treated as one either. The fundamental distinction lies in the fact that AGI is not designed for a single application, but rather aims for broader, flexible problem-solving abilities across various domains. This broad applicability, coupled with the inherent complexity and unpredictability of AGI, makes it incompatible with the rigid, narrowly focused framework of medical device regulation.

In conclusion, while traditional metrics may be sufficient for specific applications such as local generalization, extreme generalization necessitates the development of new evaluation and validation concepts to reliably capture the capabilities of autonomous hypothesis formation and adaptation. The integration of AGI into clinical practice thus requires the creation of novel metrics and an expanded validation model that transcends the conventional framework of medical products, thereby highlighting the full potential of generalized AI applications. Currently, we are still far from illuminating this complex area, and existing approaches to explainable AI may be destined for failure, as the challenge is similar to that of elucidating consciousness, which is not aligned with the practical needs in this field.

4. Discussion

In examining the generalization capabilities of predictive models for diabetes risk, we observe distinct limitations and potentials at various levels of generalization. Local generalization offers a reliable and narrowly defined risk assessment based on established factors; it is constrained to predetermined contexts and familiar patient profiles. In contrast, broad generalization enhances the model's applicability by facilitating risk prediction across diverse populations and associated health conditions. This broader level allows for more refined patient stratification and enhanced preventive measures. However, it still depends on familiar domains, potentially overlooking novel risk factors not considered by the model developers. Often, this is even not intended, as the system's reliability was not validated for such unforeseen factors.

Extreme generalization, which seeks adaptability in unforeseen and complex contexts, represents the most ambitious and least understood application, particularly in healthcare. The primary challenge lies in equipping a model to identify, hypothesize, and integrate new medical insights or environmental changes without prior exposure. This level not only complicates generalization but also extends the limits of AI validation within current frameworks. While Chollet's proposal of the ARC provides valuable insights for testing, its clinical applicability is limited, highlighting the urgent need for new evaluation methodologies specifically designed for dynamic health contexts. To advance the field, interdisciplinary collaboration among healthcare professionals, data scientists, and ethicists is essential to ensure that emerging risk factors are effectively integrated into predictive models.

References

[1] Olawade DB, David-Olawade AC, Wada OZ, et al. Artificial intelligence in healthcare delivery: Prospects and pitfalls. J Med Surg Public Health. 2024 Aug 1;3:100108. Available from: https://www.sciencedirect.com/science/article/pii/S2949916X24000616

[2] Moulaei K, Yadegari A, Baharestani M, et al. Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. Int J Med Inf. 2024 Aug;188:105474. doi: 10.1016/j.ijmedinf.2024.105474.

[3] Hoogdalem A van , Bosman AMT. Intelligence tests and the individual: Unsolvable problems with validity and reliability. Methodological Innovations. 2023 Dec 8;17(1).

[4] Fahad M, Basri T, Hamza MA, et al. The Benefits and Risks of Artificial General Intelligence (AGI). In: El Hajjami S, Kaushik K, Khan IU (eds) Artificial General Intelligence (AGI) Security: Smart Applications and Sustainable Technologies. Singapore: Springer Nature; 2024. p. 27–52.

[5] Pellert M, Lechner CM, Wagner C, et al. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. Perspect Psychol Sci 2024; 19: 808–826.

[6] Chollet F. On the Measure of Intelligence. 2019. DOI: 10.48550/arXiv.1911.01547.

[7] Tatiana S, Valentin M. How not to Lie with a Benchmark: Rearranging NLP Leaderboards. 2021. DOI: 10.48550/arXiv.2112.01342.

[8] Wang H, Keskar NS, Xiong C, et al. Assessing Local Generalization Capability in Deep Models. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2077–2087.

[9] Von Rueden L, Mayer S, Beckh K, et al. Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. IEEE Trans Knowl Data Eng 2023; 35: 614–633.

[10] Abbas A, Rehman MS, Rehman SS. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. Cureus 2024; 16: e55991.