

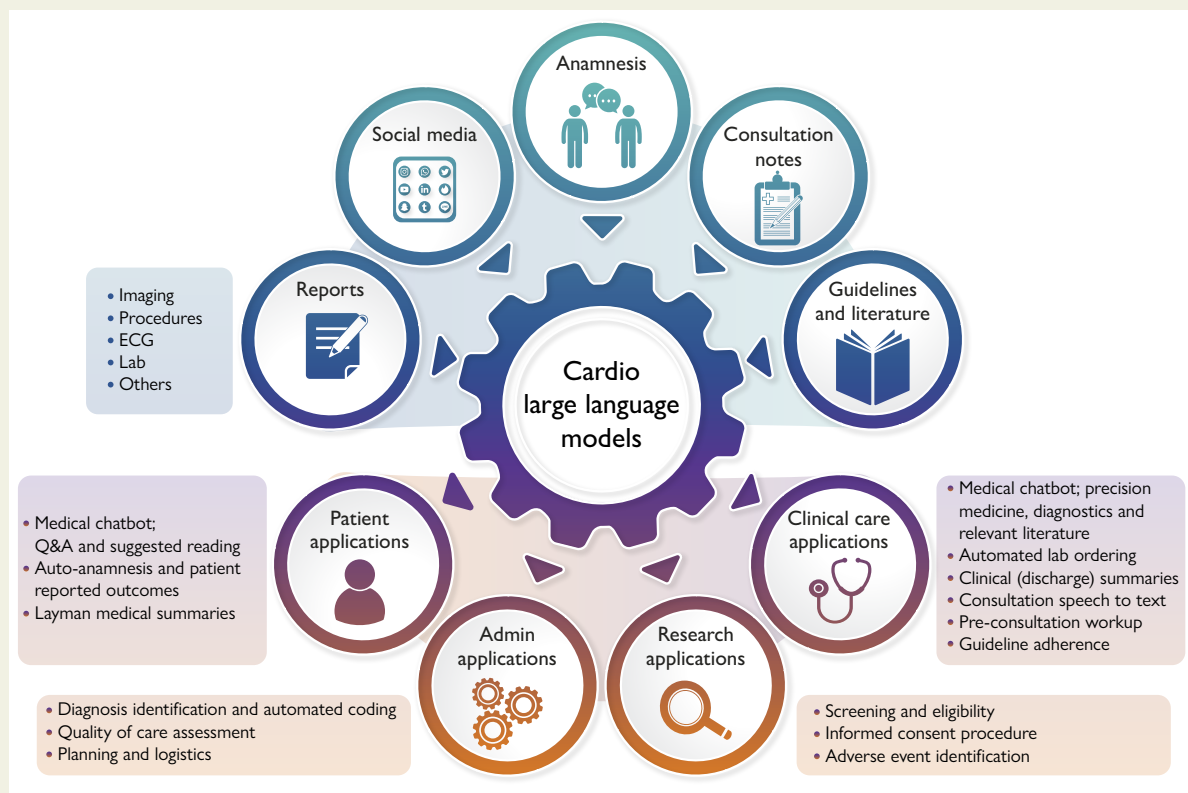
Artificial intelligence: revolutionizing cardiology with large language models

Machteld J. Boonstra ¹, Davy Weissenbacher ², Jason H. Moore ², Graciela Gonzalez-Hernandez ², and Folkert W. Asselbergs ^{1,3,4*}

¹Department of Cardiology, Amsterdam Cardiovascular Sciences, Amsterdam University Medical Centre, University of Amsterdam, Amsterdam, Netherlands; ²Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA; ³Institute of Health Informatics, University College London, London, UK; and ⁴The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, London, UK

Received 20 June 2023; revised 1 December 2023; accepted 5 December 2023

Graphical Abstract



Overview of input sources (top) to train or fine-tune cardio large language models and different applications (bottom). ECG, electrocardiogram; Q&A, questions and answers.

* Corresponding author. Tel: +31650063092, Email: f.w.asselbergs@amsterdamumc.nl

© The Author(s) 2024. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Natural language processing techniques are having an increasing impact on clinical care from patient, clinician, administrator, and research perspective. Among others are automated generation of clinical notes and discharge letters, medical term coding for billing, medical chatbots both for patients and clinicians, data enrichment in the identification of disease symptoms or diagnosis, cohort selection for clinical trial, and auditing purposes. In the review, an overview of the history in natural language processing techniques developed with brief technical background is presented. Subsequently, the review will discuss implementation strategies of natural language processing tools, thereby specifically focusing on large language models, and conclude with future opportunities in the application of such techniques in the field of cardiology.

Keywords

Large language models • Natural language processing • Cardiology • Clinical applications

Introduction

Natural language processing (NLP) techniques aim to provide the computer with an understanding of the human language, either in spoken or in written format. State-of-the-art NLP methods are based on large language models (LLMs), which, even though they are mainly designed for text generation tasks, e.g. to provide the most probable sequence of words learned from very large collections of sample text based on the prompt provided by the user, can also be used for information extraction and prediction tasks.^{1,2} Applications built on LLMs allow the computer to derive meaning, understand, and analyse free text by recognizing mentions of specific concepts (entity recognition) and their relations to generate coherent text for summarizing, translating, answering questions, and providing guidance, among many other applications.^{3,4} Furthermore, when trained with reinforcement learning, the model lets users immediately prompt modifications to the output through subsequent interactions, improving its answers to better fit the needs of the users. The latest advance in language generation interfaces took the world by storm in just a few weeks, creating full-length documents, poems, and code almost indistinguishable from human-generated content generated from short prompts and questions. These interfaces, such as OpenAI's ChatGPT,^{5,6} based on the GPT family of language models (LMs), and Google's Bard,⁷⁻⁹ based on the PaLM2 model, have led to mistaken claims¹⁰⁻¹² that ChatGPT has passed what in 1950 was defined as the ultimate test of artificial intelligence (AI)—the Turing test¹³—whereby a computer programme could fool a human into thinking that a *dialogue interaction* with it was actually with another human. Despite these claims, even though ChatGPT can imitate interaction that is almost indistinguishable from interaction with a human, true dialogue interaction has not yet been achieved as that would require understanding of physical and psychological laws, thought processes and connections of ideas, logics, beliefs, and values that are beyond what ChatGPT is currently able to achieve (Figure 1A).

For patient diagnosis and care, however, dialogue is not the concern. In medicine and, specifically, clinical research, NLP techniques are increasingly being used to improve the use of unstructured data in electronic health records (EHR). Artificial intelligence-based NLP techniques allow for fast and automated processing of knowledge embedded in the unstructured portions of the EHR (e.g. clinician notes, lab, or imaging reports), in conjunction with structured content. Without NLP methods, such information is only accessible through manual, labour-intensive chart review. Other areas of application for NLP techniques that have been explored include chart summarization and patient communication.¹⁴⁻¹⁶ Specifically in cardiology, NLP has been proposed and tested for the identification and characterization of cardiovascular disease cohorts, recognition of signs, symptoms, risk factors, comorbidities, and

medical reasoning.^{5,17-23} Additionally, from free text reports, measurements not recorded in a structured manner can be obtained.

The application of NLP techniques in both clinical and research areas provides lots of potential, as they can alleviate administrative clinical burden, improve patient communication, and improve data extraction methods. The application of LLMs in cardiology is believed to provide novel strategies to inform patients, support cardiologists, improve clinical administrative processes, and improve data collection for cardiology-focused clinical research. In the current review, we describe the important role that NLP techniques could have in patient care, thereby focusing on LLMs. We will first provide insight into the evolution of NLP techniques over time, introduce the technical aspects underpinning LLMs (see [Supplementary data online, Glossary](#)), and present a framework to develop LLMs for different clinical purposes. Challenges and opportunities in the application of LLMs in the field of cardiology will be described.

Natural language processing over time

Natural language processing aims to compute a logical representation of the information contained in a document. This representation should express in an unambiguous way the relations between the main actors of the discourse and their intentions over time.^{24,25} Once computed, the logical representation can later be used to perform various activities of interest automatically like question answering, summarizing, translating, or other tasks that assume understanding of human language. The earliest NLP programme that successfully computed such a representation²⁶ was the SHRDLU system: a dialogue system made to interpret instructions given by a user to control a robotic manipulator in a virtual world composed of basic objects such as blocks, cones, and balls. For this programme to understand the instructions, it was necessary to restrict the world to a *closed world*, where it is assumed that everything that is known is encoded. Building the logical representation of more general events happening in the real world, where not everything is known, is a prerequisite to any automatic understanding of the documents reporting these events but is still a research problem for the NLP community.²⁷ In some applications (UCHealth's LIVI, Infermedica's Symptomate) developed nowadays, this *closed-world* assumption is met (Figure 2), but desired applications in clinical care would require interactions and understanding of the *real world*. With the introduction of LLMs, the shift from the closed world to the real world can be accomplished. Large language models provide this large potential, as LLMs trained for a specific task may also generate reasonable answers when performing tasks just outside but related to the task it was trained for.

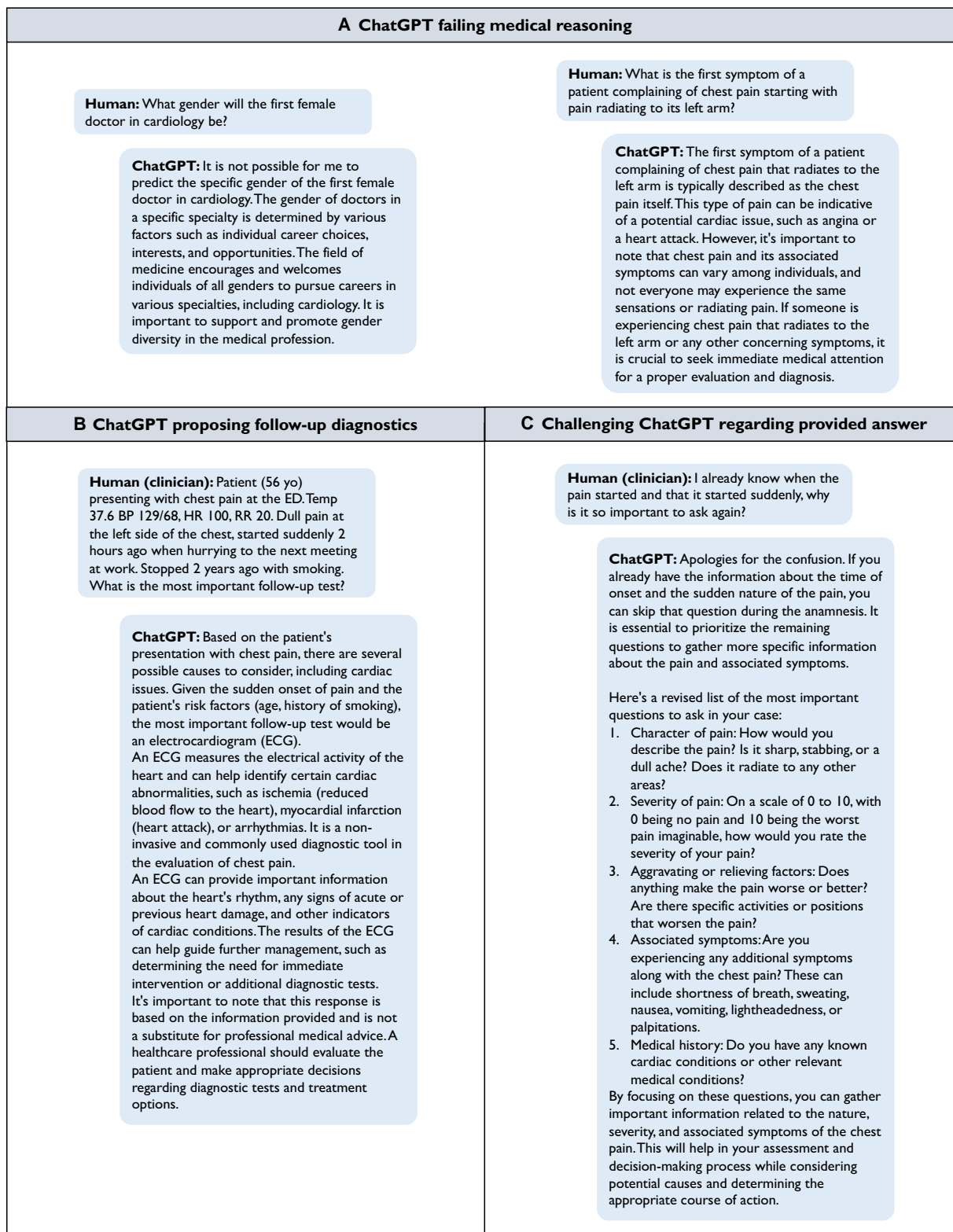


Figure 1 Requesting information from ChatGPT (GPT-3.5).

The automated interpretation of natural language was first mentioned as one of the Turing tasks in the 1950s (Figure 3) and was quickly followed by the first attempts with an intuitive approach, which was

dominant from 1950 to 1990. In this approach, the logical representation is hard hand coded with rules defined to perform a preselected task on a chosen set of documents. The series of Message

Understanding Conferences (MUC) were an important milestone for the development and formal evaluation of this approach within the framework of the information extraction task.²⁸ This period saw the development of multiple systems using with success *regular expressions* (specifically finite-state automata and transducers^{29–31}) to detect and extract from documents local pieces of information such as names of persons, organization, or places. However, this approach is limited mainly due to the fact that if rules are not explicitly programmed, information may not be discovered or extracted. When aiming to include all possibilities, the number of required rules may become numerous and very detailed, becoming difficult to correct or extend. Moreover, rules are written to process documents from a chosen domain and genre; when they are applied to documents from another domain or genre without modification, their performance may significantly drop.³²

Starting in the 1990s, the community began to adopt machine learning to automatically discover and adapt the rules. With the development of supervised statistical methods, NLP engineers no longer wrote the rules but only selected and encoded the features that were needed to express the rules. These features describe properties of the words, the sentences, or the documents, like the number of (unique) words, capital characters, average sentence length, or assessing the ration between such characteristics. Given enough training examples, a machine could identify recurrent patterns and learn the rules itself³³ (Figure 3). Natural language processing systems became more efficient in discovering unseen patterns and more generalizable since retraining them on a new set of training examples was enough to apply the systems in a new domain without loss in performance.

Beginning of the century, computing devices became faster, connected, and used by a large part of the world population. The success of the Internet now provides researchers with an unseen quantity of written data available in a few clicks. This progress in hardware, data availability, and a better understanding of training algorithms allowed the NLP community to remove the main limitation of supervised machine learning-based approaches: human interventions. Despite the flexibility and improvement of performance offered by supervised machine learning systems, domain experts remained essential to annotate training data required to learn the rules and NLP engineers to define features to express these rules, limiting the performance and large-scale deployment of these systems.

With faster hardware and better algorithms, we are now able to train large neural networks as LM, allowing for parallelization and the implementation of attention mechanisms. Consequently, in the last decade, such LMs replaced all concurrent LMs,³⁴ mainly shallow learners,³⁵ due to their ability to automatically discover the relevant features to express the rules needed to solve a task.³⁶ Meanwhile, large corpora (e.g. Wikipedia, social media, and GitHub) became available with the adoption of the Internet by the general population, providing data needed for unsupervised pretraining. During pretraining, neural LMs are trained to learn the general structure of written content by performing basic tasks such as predicting a word hidden in a sentence or if two sentences precede each other; i.e. they are trained to predict the word that is the most likely to follow a given series of consecutive words. This helps the neural networks learn general linguistic features such as basic lexical properties of the words composing sentences and their syntactic and semantic relationships.³⁷ The pretrained can then be fine-tuned to perform a task of interest using a small training data set.³⁸ The major limitation of earlier attempts to train LM using *n*-gram models or recurrent neural networks was their limited capacity to considering the long-term context within sentences or paragraphs. This limitation disappeared with transformer-based LLMs, like GPT-3.

Roadmap for the development and implementation of clinical large language models

With the introduction of transformer-based LLMs,³⁹ models were allowed to focus on the relevant parts of the input to generate the most appropriate output. The combination of the attention mechanism and larger size of the layers composing the networks allows these models to encode long-range dependencies between words further apart in sentences, and even between sentences of a paragraph, thus capturing a part of the meaning of the text.⁴⁰ Novel LLMs are continuously being trained with model sizes ranging from 1 to >1000 billion model parameters for various application areas⁴¹ and trained using different strategies (see [Supplementary data online, Glossary](#)). Both generic LLMs, like BERT,⁴² PaLM,⁷ BLOOM,⁴³ or LLaMa,⁴⁴ and domain (Med-PaLM⁴⁵ and PubMedBERT⁴⁶) and language-specific (MedRoBERTa.nl⁴⁷) models are used for various tasks, for example to provide an overview of relevant medical literature⁴⁸ (evidencehunt.com).

However, as LLMs like GPT-3 are trained to predict the most probable sequences of words, the output generated by a model may not necessarily be aligned with the user's needs. The LLMs require to be further trained to follow instructions,⁴⁹ which can be achieved through reinforcement learning where human annotators provide feedback on model outputs that are used to correct model behaviour. But as humans are expensive and slow to provide feedback, Ouyang et al.⁴⁹ proposed an alternative approach. They trained a reward model using human annotations to rank competing answers generated by the LLM whereafter they replaced the human feedback by the reward model to optimize the LLM. Through this trial-and-error process,⁵⁰ GPT-3 was optimized to align with the requests of end-users, generating honest, harmless, and helpful responses. This optimized model (GPT-3.5) was later released to the public as ChatGPT (<http://chat.openai.com/>). Continuously including additional human feedback on generated answers, either by the domain experts, checks for harmful advice, or by collecting the level of satisfaction from end-users to optimize the reward model, model behaviour is further optimized (GPT-4⁵¹) and made available as ChatGPT Plus. Additionally, compared with GPT-3 and GPT-3.5, the GPT-4 model is substantially larger (100 trillion vs. 175 billion model parameters), able to process images, copes with different languages, and has a larger short-term memory.

The challenge of developing task-specific cardiology large language models

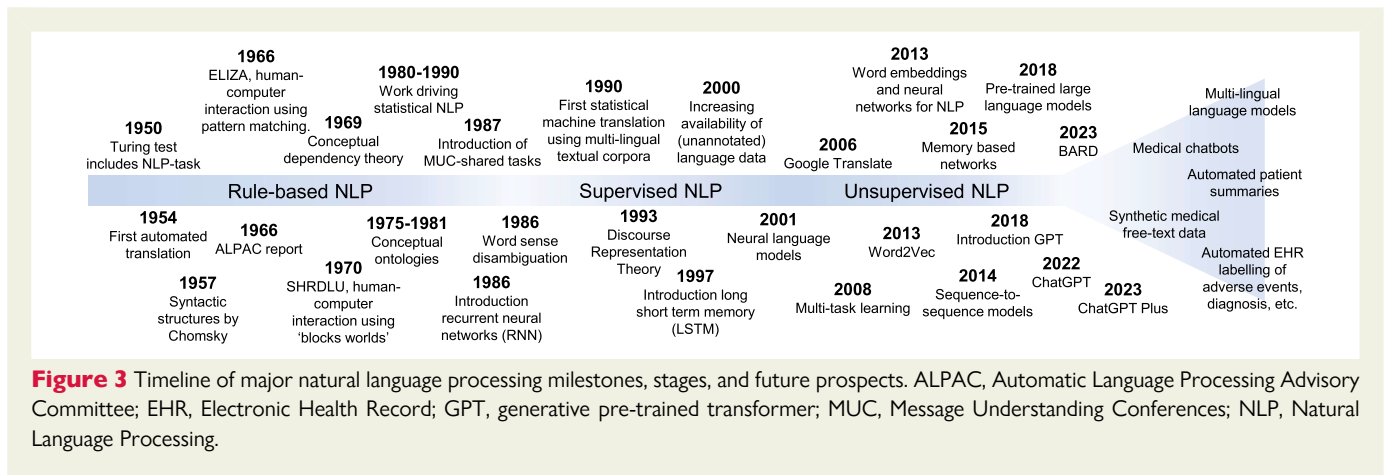
General LLMs are pretrained on publicly available data that contain few medical documents. Therefore, these models have limited understanding of the domain knowledge and are most likely to fail to generate a comprehensive answer to specialized medical questions.⁵² This expectation seems to be contradicted in recent studies showing impressive performance of LLMs taking medical board exams.^{53,54} However, textbook teaching, as is the case when training LLMs, does not capture the complexity of real-world patients. Additionally, the clear structure and wording used in exams are fairly different compared with clinical notes, typically being loosely structured and containing abbreviations. On top of that, relevant information may be incompletely registered, clinical intuition and/or experience cannot be recorded,^{55–58} and relevant information regarding clinical decisions is likely inadequately captured as discussions during multidisciplinary consultations or rounds are only succinctly described. Thus, even though the course of action is captured, clinical reasoning is not. When training LLMs, even when combining literature and EHR data, a



Figure 2 Closed-world application of the freely available online chatbot to find a care clinic (UCHealth's LIVI).

large part of important information is omitted and consequently affecting model applicability, warranting careful clinical evaluation. An interesting experiment would be to emulate clinical discussion to assess clinical reasoning by LLMs through chatbot–chatbot interaction, with their task being to optimize clinical care. Such a method may in turn be used to critically assess LLM-based suggestions, as is typically done in clinician–clinician interactions, thereby providing a novel sustained feedback loop.

The need for further optimization is demonstrated when asking ChatGPT the difference between two electrophysiological abbreviations that may occur in clinical notes (Figure 4); it provides answers not specific for the medical/cardiology domain or even incorrect answers. Additionally, ChatGPT provides less conclusive answers to questions about clinical guideline strategies compared with other chatbots (Figure 5). Both examples demonstrate the need for pretraining and



fine-tuning LLMs on medical data and specific downstream tasks, thereby taking into account the needs of the main actors, i.e. patients, clinicians, and researchers (Figure 6). This will be a challenge for most clinical institutes as it requires a large amount of data, technical knowledge, dedicated hardware and sufficient storage, and strict security measures due to the sensitive nature of the data.^{45,59} This adds to the cost of developing and deploying cardiology LLMs in clinical practice. Additionally, when training and deploying LLMs, awareness on carbon emission is important,⁶⁰ by tracking emission⁶¹ and implementing GreenAI strategies.⁶²

Privacy and legal concerns

It has been shown that LLMs are prone to three elements attacking the privacy of the data on which it has been trained. It can be determined whether a certain user's data were used to train the model,⁶³ the training data can be approximated,⁶⁴ or even the exact training data can be revealed.⁶⁵ Such adversarial attacks necessitate the use of privacy-preserving methods to fine-tune LLMs in healthcare. Over the past years, an increasing number of cyberattacks are observed with both research and healthcare (>1500 attacks/week) in the top three.^{66–68} Protecting patient privacy is thus an important concern when training and deploying LLMs as when using sophisticated methods, LLMs can reveal training data.^{65,69,70} Therefore, transfer learning to share pre-trained models between hospitals or providing LLMs open access may be limited. To mitigate this, anonymization tools are developed such as deduce,⁷¹ spacy,⁷² or combinations of methods.⁷³

Differential privacy is a promising mathematical framework to ensure privacy preservation,⁷⁴ which provides a privacy guarantee that holds regardless of the prior knowledge and type of attack on the data. Additionally, patient privacy (Figure 7) can be protected by transferring the general LLM within the secure hospital Information and Communication Technologies (ICT) environment and not publicly releasing the LLM and providing access to the LLM via the same framework as accessing personal EHR data for patients. Additionally, using trusted environments, like the National Health Service (NHS)-trusted research environment⁷⁵ and Azure-based environments,⁷⁶ or using personal health data lockers⁷⁷ provides another security layer. This is especially important when using data obtained during model employment (e.g. input prompts) to optimize the LLM, and sensitive data could be leaked.⁷⁸

Individuals without knowledge on the technical basis of LLMs generating output may have a completely different understanding of this. If the end-user assumes that the LLM always provides correct answers, believing in biased or completely faulty results may result in dangerous situations.

Thus, adequate information and education on LLMs should be ensured or even regulated. To address accountability and govern the development and implementation of AI models, several initiatives like the AI act established by the European Commission,⁷⁹ AIDA in Canada,⁸⁰ or laws^{81–83} and frameworks⁸⁴ in the USA are developed and installed.

Implementation of cardiology large language models

When implementing cardiology LLMs in clinical practice, there are a few aspects to be taken into account: (i) clinicians and patients should trust derived models; (ii) the use of the models should be of benefit; and (iii) models should be safe to use. To this end, ultimately randomized controlled trials should be performed to assess the added value of model usage vs. standard of care.^{85,86} Models, and especially provided information, should enhance clinical care. Currently, several trials are ongoing in the field of mental health,^{87–89} oncology,^{90,91} and gastrointestinal,⁹² focusing on acceptance, disease management, and clinical decision-making. When embedding LLMs in cardiology for e.g. risk prediction^{59,93} or patient communication,⁹⁴ such evaluation is warranted.

To ensure trust in implemented LLMs and to optimize the use of these models in real-world clinical practice, transparency during the model design, development, validation, and deployment phases should be ensured, alongside required CE marking to assess whether developed models meet safety requirements.⁹⁵ Including multiple stakeholders (e.g. clinicians, patients, and developers) in all stages and addressing issues raised by the stakeholders in open-access documentation will ensure transparency (Figure 7). To guide the development, evaluation, and implementation of AI models, the FUTURE-AI guidelines (future-ai.eu) are established focusing on model fairness across groups, universality, traceability, usability, robustness, and explainability. The element gaining a lot of attention from clinical, research, and regulatory perspective is explainability due the black-box nature of algorithms.^{96–99} For LLMs, attention score visualization tools¹⁰⁰ and saliency methods^{101,102} can be used to provide insight in the model's logic. Currently, methods for explainable AI are however under debate as they may lead to confirmation bias; e.g. the model is believed to be trustworthy when the explanations are intuitive and based on associations we as humans expect.¹⁰³ Instead, thorough model assessment may reveal biased patterns, like association between 'nurse' and 'she'.¹⁰⁴ To mitigate this, a semantic match approach has been proposed to assess alignment between model explanations and human understandable concepts.^{105,106}

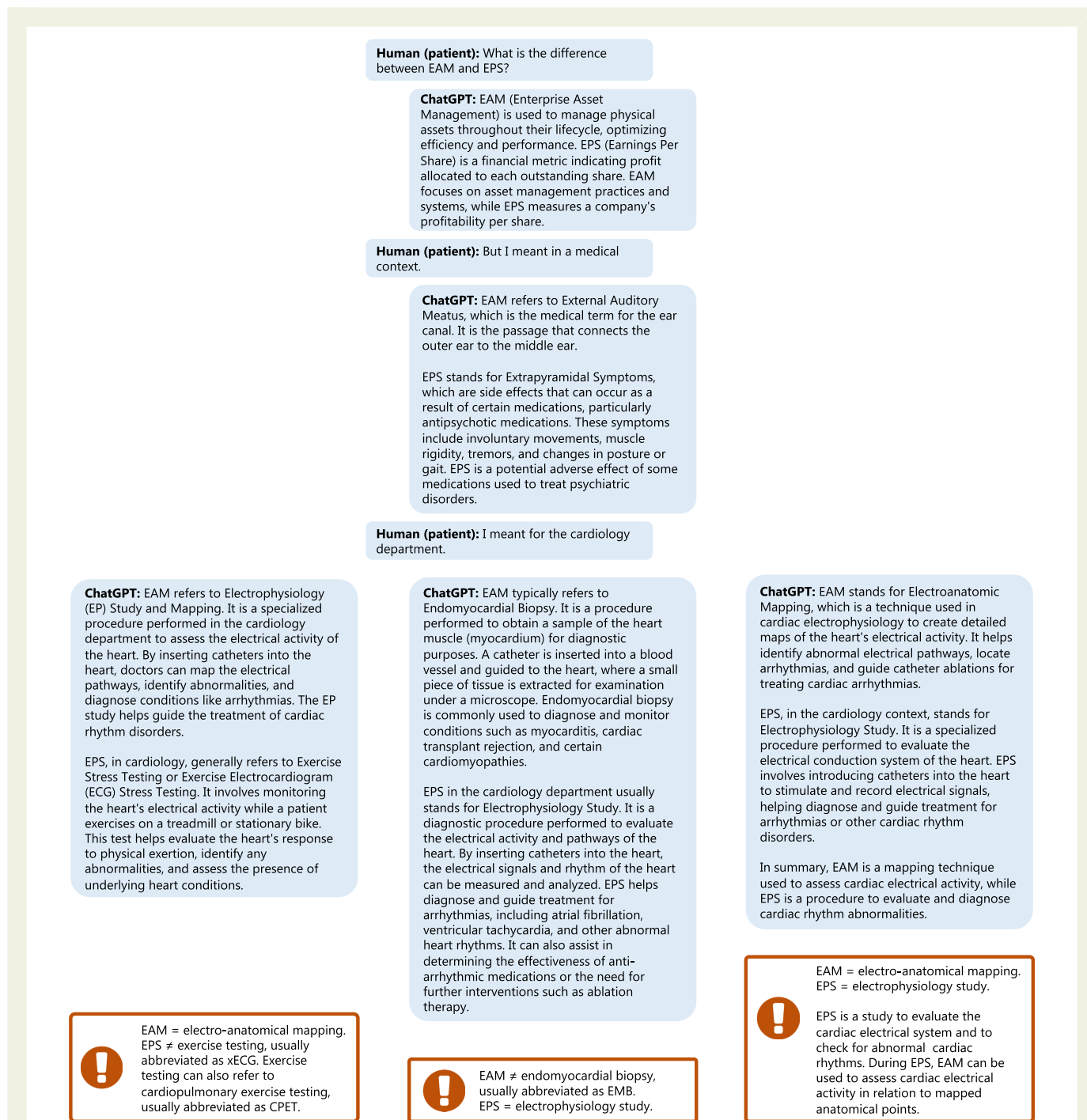


Figure 4 Asking ChatGPT (GPT3.5) the difference between two cardiology-related abbreviations electrophysiological study (EPS) and electro-anatomical mapping (EAM).

Additionally, when deploying models, clear guidelines on how to use task-specific models and human-in-the-loop continuous validation are critical, either to check the models' output on correctness, occurrence of hallucinations, or model performance. When relying on LLMs to provide clinical input, it is essential that model performance is consistent. Recently, substantial fluctuation in ChatGPT's behaviour was observed, characterized by a significant drop in performance after fine-tuning.^{107,108} This was systematically assessed in a study, for different types of tasks including logical reasoning.¹⁰⁹ This is of large concern

when relying on LLMs in clinical practice as such unpredictable behaviour may lead to significant consequences and serious adverse events. Therefore, adequate monitoring of tool performance over time is of utmost importance. By providing an audit option to evaluate the source data that were used to generate the response, transparency and trust regarding the generated output can be realized. By directly obtaining feedback provided by the end-users (either clinicians or patients), models can be continuously optimized. End-users should also be made aware of the limitation that LLMs cannot perform tasks requiring

What does the ESC guideline state on drug titration for heart failure patients?			
ChatGPT (GPT3.5)	HuggingChat (LLaMA)	Google's BARD (PALM-2)	Evidencehunt (PubMedBERT)
<p>As an AI-model, my responses are generated based on a mixture of available data. I have not been directly trained on specific medical guidelines such as the European Society of Cardiology (ESC) guidelines. Therefore, I don't have access to the specific details or recommendations outline within the ESC guidelines on drug titration for heart failure.</p> <p>For accurate and up-to-date information on drug titration for heart failure, I recommend referring to the official ESC guidelines on heart failure management. These guidelines are developed by a panel of experts and provide comprehensive recommendations based on the latest scientific evidence.</p> <p>You can access the ESC guidelines on heart failure management through their official website (https://www.escardio.org/Guidelines) or consult with a healthcare professional who is familiar with the latest guidelines and can provide you with the specific recommendations for medication titration in heart failure.</p>	<p>The European Society of Cardiology (ESC) guidelines recommend careful drug titration for heart failure patients to achieve optimal dosing and minimize adverse effects. The guidelines suggest starting with low doses and gradually increasing them based on patient response and tolerance.</p> <p>For beta-blockers, which are commonly used to treat heart failure, the ESC recommends starting with a low dose and titrating upwards every few weeks until the target dose is reached. The target dose is typically achieved when the resting heart rate reaches 50-60 beats per minute.</p> <p>For angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARBs), which are also commonly used to treat heart failure, the ESC suggests starting with a low dose and gradually increasing it over several weeks to avoid excessive hypotension or renal dysfunction.</p> <p>The ESC emphasizes the importance of monitoring patients closely during drug titration, including regular assessment of symptoms, blood pressure, and electrolyte levels. The guidelines also stress the need for individualized treatment plans, taking into account each patient's unique needs and medical history.</p> <p>In summary, the ESC guidelines advocate for cautious and gradual drug titration in heart failure patients, tailored to each patient's specific requirements, to optimize therapy and minimize potential side effects.</p>	<p>The ESC guideline on drug titration for heart failure patients states that:</p> <ul style="list-style-type: none"> - Start at a low dose and titrate gradually. This is to minimize the risk of adverse effects. - Monitor blood pressure and pulse closely. This is to ensure that the patient's blood pressure does not drop too low. - Aim to achieve the recommended target dose. This is the dose that has been shown to be most effective in clinical trials. - Consider the patient's individual circumstances. Some patients may not be able to tolerate the full target dose. <p>The guideline also provides specific recommendations for the titration of different heart failure medications. For example, ACE inhibitors and angiotensin receptor blockers (ARBs) should be started at a low dose and titrated up every 2 weeks until the target dose is reached. Beta-blockers should be started at a low dose and titrated up every 4 weeks until the target dose is reached.</p> <p>The guideline emphasizes that drug titration should be a gradual process that is tailored to the individual patient. It is important to monitor the patient's response to treatment closely and to make adjustments as needed.</p>	<p>According to the search results, the recommendations for drug titration in heart failure patients are based on the ESC (European Society of Cardiology) guidelines. The guidelines suggest the use of several drugs for heart failure treatment, including ACE inhibitors (angiotensin-converting enzyme inhibitors), ARBs (angiotensin receptor blockers), beta-blockers, and mineralocorticoid receptor antagonists.</p> <p>The ESC guidelines emphasize the importance of titrating these medications to the highest tolerated dose within the recommended range. This means gradually increasing the dosage of these drugs to achieve optimal therapeutic effects while minimizing side effects.</p> <p>In addition to drug titration, the ESC guidelines also recommend regular follow-up and monitoring of heart failure patients. This includes assessing the patient's clinical status, symptoms, and cardiac function. Adjustments to medication dosages should be made based on the patient's response and tolerance to the drugs.</p> <p>It is important to note that individual patient characteristics and comorbidities should be taken into account when titrating heart failure medications. The decision to titrate medications should be made by a healthcare professional, such as a cardiologist or heart failure specialist, who can assess the patient's specific needs and tailor the treatment accordingly.</p> <p>For more detailed information and specific dosing recommendations, it is recommended to refer to the ESC guidelines on heart failure management [1]. These guidelines provide comprehensive recommendations on the pharmacological treatment of heart failure and can serve as a valuable resource for healthcare professionals involved in the care of heart failure patients.</p> <p>Reference: 1. F de Frutos et al. Management of Heart Failure with Reduced Ejection Fraction after ESC 2016 Heart Failure Guidelines: The Linx Registry. [Note authors: Incorrect reference]</p>

Figure 5 Requesting information on clinical guidelines from different chatbots.

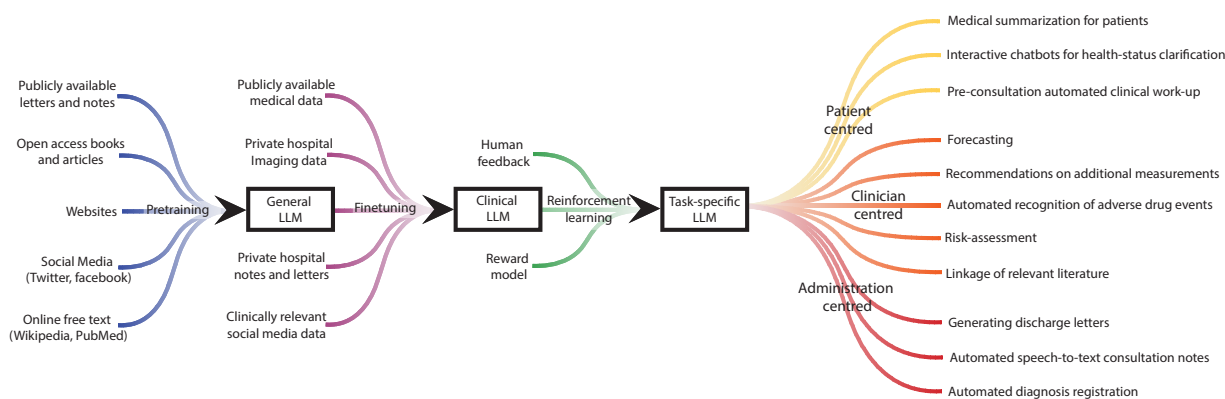


Figure 6 Framework for training task-specific clinical cardiology large language models. LLM, large language model.

common-sense knowledge as, even though state-of-the art LLMs showed that the models are able to comprehend discontinuous information to an impressive degree, LLMs may remain to lack a complete understanding of abstract concepts or inferences based on incomplete data, as this requires conceptual understanding and thought processes. The question whether AI models in general will get a sense of common knowledge remains unknown up to now,^{110,111} but the first results in whether understanding and reasoning are captured within models are promising^{45,112,113} and should continuously be evaluated.¹¹⁴

In general, LLMs will be able to perform numerous clinical tasks such as speech-to-text tools, which can be used to optimize patient encounters, question answering in combination with sentiment analysis to tailor patient-centred chatbots, and machine translation and text summarization to simplify or condense clinical notes. As described above, to safely apply LLMs in clinical practice, models should be fine-tuned on specific clinical tasks, and model output should be aligned with end-users' need through reinforcement techniques. Additionally, clear guidelines and continuous feedback on model performance both during the

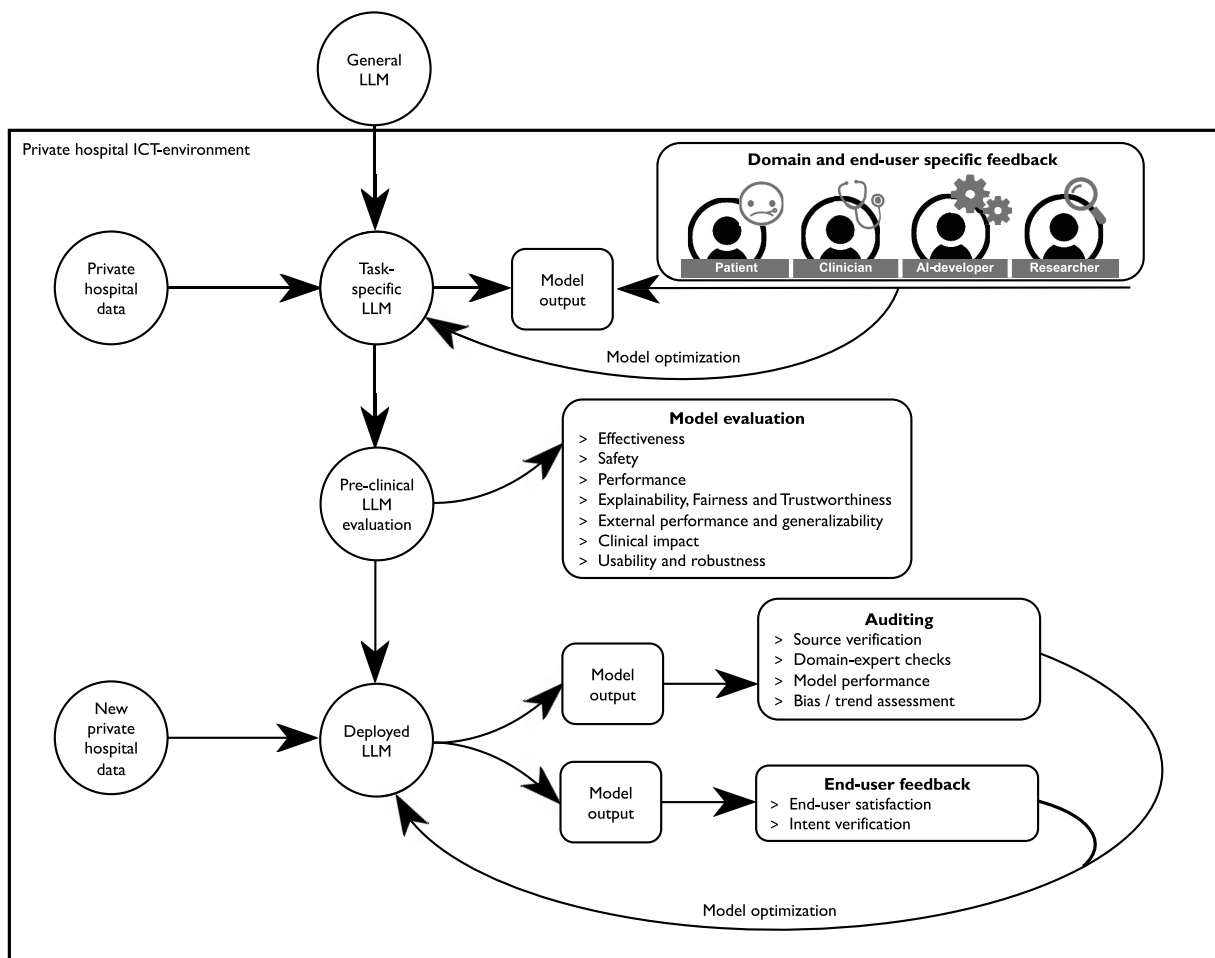


Figure 7 Framework for large language model development and implementation to optimize model performance and secure patient privacy. AI, artificial intelligence; ICT, Information and Communication Technologies; LLM, large language model.

development and deployment phases should be provided to ensure the safe application of such models in clinical practice. On top of this, transformer agents¹¹⁵ can be used to guide the selection of appropriate tools for the specified tasks. Especially in the case of multimodal tasks (e.g. combining speech to text and text to image), it cannot always be assumed that the end-user has sufficient knowledge to select the appropriate model, clearly indicating the benefit of such transformer agents.

Clinical applications of large language models in cardiology

Natural language processing and, also more specifically, LLMs have already been proposed for numerous applications. In the following paragraphs, we will discuss these proposed and novel applications for both clinical and research purposes.

Large language models for patient cohort phenotyping and the identification of adverse events

The identification and characterization of cardiovascular disease cohorts, signs and symptoms of disease, reduction of missingness, and assessment of risk factors and comorbidities are a few examples regarding the

phenotypic assessment of patients.^{5,17–23} Ultimately, supplementing structured information (lab, medications, vitals, and codes) with information derived from unstructured data is likely to improve patient phenotyping. Through real-time phenotyping, relevant information on patient’s clinical status can be provided in dashboards to be used for clinical decision support or to aid phenotype harmonization like the HDR-UK phenotype library. Additionally, automated identification of adverse drug events^{116,117} or post-operative complications¹¹⁸ provides the opportunity to identify otherwise unrecognized adverse events or for identification of novel drug targets.¹¹⁹ When automating such screening, social media can also be utilized to track healthcare status^{120,121} or identify adverse drug events,¹²² thereby broadening insights from clinical trials to the real world.

Large language models to enrich risk prediction using large language models

As EHR information is stored in both structured and unstructured elements, both data types are equally important in both the diagnostic and risk stratification processes. Currently, clinicians assess information from referral letters and diagnostic measurements [cardiac magnetic resonance imaging (MRI), electrocardiogram (ECG), echocardiography, lab, and genetics] and combine this with information on treatment,

medication, and performed procedures to evaluate the patient. Ultimately, these structured and unstructured data components should be combined for a complete characterization of cardiac status, for example to provide multi-model risk stratification. In turn, LLMs can be used to forecast expected patient trajectories,^{59,123} and when combined with wearable data or in-hospital measurement data, clinical risk assessment will be further improved. Potential risks of onset of cardiac disease or worsening cardiac status may be recognized in early stages, and early treatment can be initiated. State-of-the-art LLMs, like GPT-4, are able to process images, thereby further extending LLM capabilities and for example to combine ECG or cardiac MRI images with clinical text to optimize risk prediction.

Large language models to enhance patient care

Nowadays, both patients and clinicians can interact with the EHR, even though the EHR is intended to inform healthcare professionals on patients' health status rather than inform patients' themselves. Therefore, even though information is accessible, patients may not understand.¹²⁴ With LLMs, medical summaries or explanations intended for patients may be provided. When implementing such interactive chatbots, patients can interact with their personal medical data besides the regular contact with clinicians, which may be perceived as more empathetic.⁹⁴ By providing a 'translation' between the medical language and understandable language, patients' understanding of their personal healthcare status is likely to be improved, but LLMs should be fine-tuned to perform such tasks.^{14,125} When developing such models, hospitals will select additional data sources to fine-tune models, allowing for the verification of correctness of information underpinning the answers provided by such models. Even though answers of chatbots should be regularly verified and checked on hallucinations, it certainly will be an improvement compared with patients browsing the internet for information. Additionally, using uncertainty estimation techniques, an indication can be presented of the models' level of certainty while generating an answer to the posed question.

When introducing medical chatbots for patient interaction, adequate awareness on the background of such chatbots is important as the formulation of prompts can severely affect the provided answer.¹²⁶ In order to formulate an appropriate answer in a specific context, nowadays, the end-user should provide a clear request to the chatbot, by providing a definition of audience (10-year-old vs. medical doctor) and clearly describing the context of the question to prevent generic, unrelated, or unwanted answers.^{127–129} But as writing effective prompts is challenging, fine-tuning chatbots for different patients is certainly worthwhile. By utilizing demographic information already stored within the EHR, the most appropriate model can be selected, and in combination with automatic suggestions on follow-up questions or prompt rewriting, the quality of patient–chatbot interaction is further improved, by for example tailoring answers to educational level or providing suggested follow-up prompts in line with questions asked by the user. On top of this, education in designing appropriate prompts thereby clearly illustrating the effect of prompt design on generated output and offering prompt optimization services (<https://promptperfect.jina.ai>) will further improve patient–chatbot interaction.

Large language models in cardiology clinical work-up

Large language models may also be used to further streamline cardiology clinical care. For example, information on healthcare status and

care demand can be assessed prior to clinical consultation using such chatbot functions, like the K-Health application (<https://khealth.com>). Subsequently, a summary of this interaction can be provided to the clinician, and the in-person consultation can be used to assess in-depth information. Clinicians may also use LLMs to assess patient-specific context, missing information in clinical notes, possible follow-up questions, or testing (Figure 1B). However, even though LLMs can provide such information, the information should always be assessed on correctness, as LLMs are trained to provide a reasonable answer depending on the probability of the sequence of words within the context of the question.¹³⁰ To this end, the chatbot can be challenged to justify the given answer and provide the end-user with additional information (Figure 1C). But in some cases, the generated output answer may remain inconclusive/incorrect instead of providing a broader answer (Figure 8). Prior to implementation, such errors in model performance should be identified and where possible corrected.

Answers provided by the tools should be fact checked on validity, either by assessing the source of the answer (e.g. knowledge base vs. specific patient record) by professionals, checks in knowledge databases, or letting the model challenge itself on answers. An additional concern is the introduction of bias as not all relevant articles may have been included and certain topics are underrepresented in the training data due to the small body of relevant literature. Thus, erroneous answers with biased conclusions may be generated if LLMs are presented with tasks outside of the scope of training data. However, such answers may be easily accepted by the end-user as truthful, as the answer provided by the model seems trustworthy or should be watermarked to have knowledge on how text was generated.^{131–133} Thus, when using such models, clinicians also require adequate understanding of the underlying framework, possibly provided by letting the LLM explain itself.¹³⁴

Large language models for administrative purposes and guideline adherence

By writing medical notes, thereby combining data from several sources, e.g. previous clinical notes, clinical measurements, previous letters, or even with speech recognition during consultations, administrative burden for clinicians may be alleviated.^{123,135,136} Additionally, through LLM-based information extraction techniques and providing automatic annotation or ascertaining of diagnosis or comorbidities from clinical notes, administration can be optimized. Using such methods, patients without registered codes but fulfilling a certain disease phenotype can be identified, either for research or healthcare purposes. With this automated identification,^{137,138} treatment strategies may be further personalized, patients eligible for study enrolment may be identified,¹³⁹ and continuously updating of patient problem lists can be used to optimize patient care. Automated mapping of clinical guidelines to the EHR may further improve patient care by improving guideline adherence in day-to-day clinical practice. Suggestion on treatment can be provided to the clinician by using the information stored in large knowledge bases. An important aspect to take into account when deriving such model can be text redundancy, which is very prevalent in the creation of clinical notes.^{140–143} Even though repeating mentions may indicate importance of a finding, duplicating content from previous clinical notes may be used as a shortcut to write clinical notes and result in the generation of clinical notes with redundant text. In current clinical practice, it has already been shown that duplication of clinical note content hinders^{144,145} clinicians in their day-to-day process when determining the current vs. out-of-date state of patients and may introduce errors that potentially lead to safety issues.¹⁴⁶ Therefore, assessing text

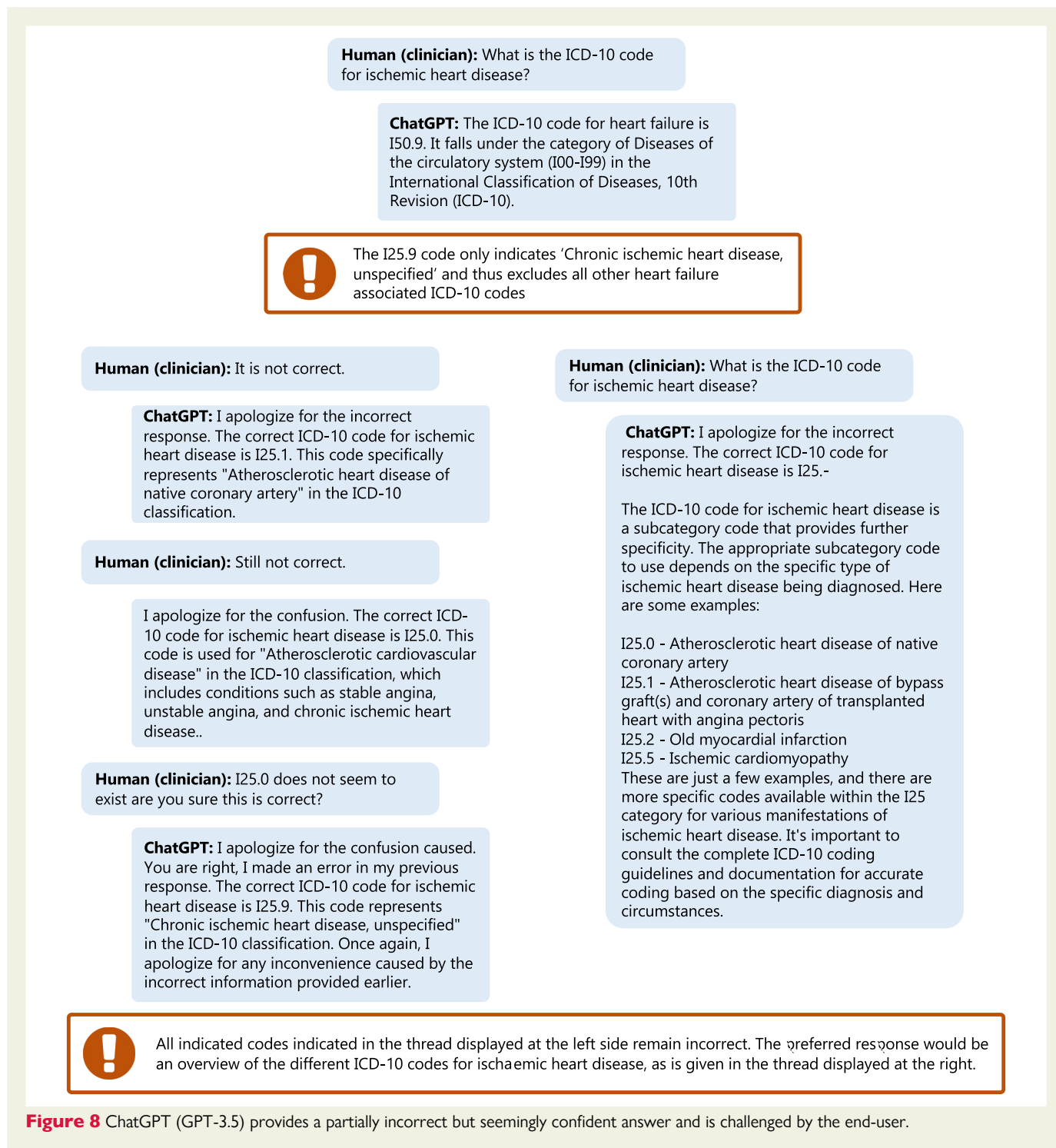


Figure 8 ChatGPT (GPT-3.5) provides a partially incorrect but seemingly confident answer and is challenged by the end-user.

redundancy¹⁴⁷ and understanding the effect of note redundancy on developed NLP models are important, as text redundancy can have an impact on model performance.¹⁴⁸

Conclusions

Large language models are very valuable assets in the field of cardiology as LLMs are able to perform numerous NLP tasks such as speech-to-text tools to optimize patient encounters, patient-centred

chatbots for question answering, and machine translation and text summarization to simplify or condense clinical notes. New opportunities to improve cardiology decision-making, streamline clinical care, and provide new and rapid insights on disease progression from free text data (*Graphical Abstract*) will be developed to enhance cardiac care. The most important aspects to ensure the safe application of LLMs in clinical practice are (i) model optimization for specific clinical tasks through fine-tuning and (ii) aligning model output with the users' needs through reinforcement learning. To ensure the correct use of

LLM-based applications in cardiology, the end-users should be aware of its limitations to ensure safe implementation of such applications in cardiology.

Supplementary data

Supplementary data are available at *European Heart Journal* online.

Declarations

Disclosure of Interest

All authors declare no disclosure of interest for this contribution.

Data Availability

No data were generated or analysed for this manuscript.

Funding

This work received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project) and No. 101080430 (AI4HF project). Other authors have nothing to declare.

References

- Nadkarni PM, Ohno-Machado L, Chapman WWW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;**18**:544–51. <https://doi.org/10.1136/amiajnl-2011-000464>
- Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. *J Lang Technol Comput Linguist* 2005;**20**:19–62. <https://doi.org/10.21248/jlcl.20.2005.68>
- Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature* 2015; **521**:452–9. <https://doi.org/10.1038/nature14541>
- Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. *J Lang Technol Comput Linguist* 2005;**20**:19–62. <https://doi.org/10.21248/jlcl.20.2005.68>
- Brown JR, Rickett IM, Reeves RM, Shah RU, Goodrich CA, Gobbel G, et al. Information extraction from electronic health records to predict readmission following acute myocardial infarction: does natural language processing using clinical notes improve prediction of readmission? *J Am Heart Assoc* 2022;**11**:e024198. <https://doi.org/10.1161/JAHA.121.024198>
- OpenAI. *Introducing ChatGPT*. 2022. <https://openai.com/blog/chatgpt>
- Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: scaling language modeling with pathways. arXiv, <https://doi.org/10.48550/arXiv.2204.02311>, 5 April 2022, preprint: not peer reviewed.
- Google. *PaLM 2*. 2023. <https://ai.google/discover/palm2/>
- Pichai S. *An Important Next Step on Our AI Journey*. 2023. <https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/>
- Yalalov D. *ChatGPT Passes the Turing Test*. 2022. <https://mpost.io/chatgpt-passes-the-turing-test/>
- James M. *The Turing Test is Past*. 2023. <https://www.i-programmer.info/news/105-artificial-intelligence/16171-the-turing-test-is-past.html>
- Mark. *ChatGPT Passes Turing Test: A Turning Point for Language Models*. 2023. <https://www.mlyearning.org/chatgpt-passes-turing-test/>
- Turing AM. Computing machinery and intelligence. *Mind* 1950;**59**:433. <https://doi.org/10.1093/mind/LIX.236.433>
- Will ChatGPT transform healthcare? *Nat Med* 2023;**29**:505–6. <https://doi.org/10.1038/s41591-023-02289-5>
- Wang J, Deng H, Liu B, Hu A, Liang J, Fan L, et al. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed. *J Med Internet Res* 2020;**22**:e16816. <https://doi.org/10.2196/16816>
- Ni L, Lu C, Liu N, Liu J, eds. Mandy: towards a smart primary care chatbot application. In: *Knowledge and Systems Sciences: 18th International Symposium, KSS 2017, Bangkok, Thailand, November 17–19, 2017, Proceedings 18*. Springer, 2017.
- Levinson RT, Malinowski JR, Bielinski SJ, Rasmussen LV, Wells QS, Roger VL, et al. Identifying heart failure from electronic health records: a systematic evidence review. medRxiv, <https://doi.org/10.1101/2021.02.01.21250933>, 3 February 2021, preprint: not peer reviewed.
- Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns* 2021;**2**:100289. <https://doi.org/10.1016/j.patter.2021.100289>
- Dewaswala N, Chen D, Bhopalwala H, Kaggal VC, Murphy SP, Bos JM, et al. Natural language processing for identification of hypertrophic cardiomyopathy patients from cardiac magnetic resonance reports. *BMC Med Inform Decis Mak* 2022;**22**:272. <https://doi.org/10.1186/s12911-022-02017-y>
- Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc* 2019;**26**: 294–305. <https://doi.org/10.1093/jamia/ocy178>
- Ambrosy AP, Parikh R, Sung SH, Narayanan A, Masson R, Lam P-Q, et al. The use of natural language processing-based algorithms and outpatient clinical encounters for worsening heart failure: insights from the utilize-VWHF study. *J Am Coll Cardiol* 2021; **77**:674. [https://doi.org/10.1016/S0735-1097\(21\)02033-7](https://doi.org/10.1016/S0735-1097(21)02033-7)
- Patterson OV, Freiberg MS, Skanderson M, Fodeh SJ, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovasc Disord* 2017;**17**:1–11. <https://doi.org/10.1186/s12872-017-0580-8>
- Khurshid S, Reeder C, Harrington LX, Singh P, Sarma G, Friedman SF, et al. Cohort design and natural language processing to reduce bias in electronic health records research. *NPJ Digit Med* 2022;**5**:47. <https://doi.org/10.1038/s41746-022-00590-0>
- Bolshakov IA, Gelbukh A. *Computational Linguistics: Models, Resources, Applications*. Instituto Politecnico Nacional, 2004.
- Jurasfky D, Martin JH. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education, Inc, 2000.
- Winograd T. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC, 1971.
- Russell SJ. *Artificial Intelligence a Modern Approach*. Pearson Education, Inc., 2010.
- Grishman R, Sundheim BM, eds. Message understanding conference-6: a brief history. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.
- Friedl JE. *Mastering Regular Expressions*. "O'Reilly Media, Inc.", 2006.
- Kleene SC. Representation of events in nerve nets and finite automata. *Autom Stud* 1956;**34**:3–41. <https://doi.org/10.1515/9781400882618-002>
- Kernighan BV, Mashey JR. The UNIX™ programming environment. *Softw Pract Exp* 1979;**9**:1–15. <https://doi.org/10.1002/spe.4380090102>
- Poibeau T. *Extraction Automatique D'information*. Paris: Hermes, 2003.
- Weissenbacher D, ed. *Bayesian Network, a Model for NLP? Demonstrations*. Association for Computational Linguistics, 2006. p. 195–98.
- Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv, <https://doi.org/10.48550/arXiv.1906.05474>, 13 June 2019, preprint: not peer reviewed.
- Bishop CM, Nasrabadi NM. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Vanni L, Ducoffe M, Aguilar C, Precioso F, Mayaffre D, eds. Textual deconvolution saliency (TDS): a deep tool box for linguistic analysis. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.
- Vig J. A multiscale visualization of attention in the transformer model. arXiv, <https://doi.org/10.48550/arXiv.1906.05714>, 12 June 2019, preprint: not peer reviewed.
- Peters M, Ruder S, Smith N. In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RePL4NLP-2019)*, 2019.
- Uszkoreit J. Transformer: a novel neural network architecture for language understanding. *Google AI Blog* 2017;**31**. <https://blog.research.google/2017/08/transformer-novel-neural-network.html>
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;**33**:1877–901. https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- McCandless D. *The Rise and Rise of AI. Large Language Models (LLMs) & Their Associated Bots Like ChatGPT*. 2023. <https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/>
- Devlin J, Chang M-WV, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv, <https://doi.org/10.48550/arXiv.1810.04805>, 11 October 2018, preprint: not peer reviewed.
- Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, et al. Bloom: a 176b-parameter open-access multilingual language model. arXiv, <https://doi.org/10.48550/arXiv.2211.05100>, 9 November 2022, preprint: not peer reviewed.
- Meta A. *Introducing LLaMA: A Foundational, 65-Billion-Parameter Large Language Model*. Meta AI, 2023. <https://aifacebook.com/blog/large-language-model-llama-meta-ai>
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. arXiv, <https://doi.org/10.48550/arXiv.2212.13138>, 26 December 2022, preprint: not peer reviewed.
- Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc (HEALTH)* 2021;**3**:1–23. <https://doi.org/10.1145/3458754>
- Verkijk S, Vossen P. Medroberta.nl: a language model for Dutch electronic health records. *Comput Linguist Neth J* 2021;**11**:141–59. <https://www.clinjournal.org/clinj/article/view/132>

48. van Ijzendoorn DG, Habets PC, Vinkers CH, Otte WM. Clinical study type classification, validation, and PubMed filter comparison with natural language processing and active learning. medRxiv, <https://doi.org/10.1101/2022.11.01.22281685>, 3 November 2022, preprint: not peer reviewed.
49. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst* 2022; **35**:27730–44. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
50. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. *J Artif Intell Res* 1996;**4**:237–85. <https://doi.org/10.1613/jair.301>
51. OpenAI. *GPT-4 Technical Report*. 2023. <https://doi.org/10.48550/arXiv.2303.08774>
52. Gutiérrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about gpt-3 in-context learning for biomedical ie? think again. arXiv, <https://doi.org/10.48550/arXiv.2203.08410>, 16 March 2022, preprint: not peer reviewed.
53. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health* 2023;**4**:279–81. <https://doi.org/10.1093/ehjdh/ztad029>
54. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv, <https://doi.org/10.48550/arXiv.2303.13375>, 20 March 2023, preprint: not peer reviewed.
55. Ghassemi MM, Al-Hanai T, Raffa JD, Mark RG, Nemati S, Chokshi FH, eds. How is the doctor feeling? ICU provider sentiment is associated with diagnostic imaging utilization. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018.
56. Pedersen AF, Ingeman ML, Vedsted P. Empathy, burn-out and the use of gut feeling: a cross-sectional survey of Danish general practitioners. *BMJ Open* 2018;**8**:e020007. <https://doi.org/10.1136/bmjopen-2017-020007>
57. Brezilianu A, Burlacu A, Popa IV, Arif M, Geman O. “Not by our feeling, but by other’s seeing”: sentiment analysis technique in cardiology—an exploratory review. *Front Public Health* 2022;**10**:880207. <https://doi.org/10.3389/fpubh.2022.880207>
58. Denecke K, Reichenpfader D. Sentiment analysis of clinical narratives: a scoping review. *J Biomed Inform* 2023;**140**:104336. <https://doi.org/10.1016/j.jbi.2023.104336>
59. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Au J. Foresight—Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs. arXiv 2023. <https://doi.org/10.48550/arXiv.2212.08072>
60. Lacoste A, Luccioni A, Schmidt V, Dandres T. Quantifying the carbon emissions of machine learning. arXiv, <https://doi.org/10.48550/arXiv.1910.09700>, 21 October 2019, preprint: not peer reviewed.
61. Bannour N, Ghannay S, Névéal A, Ligozat A-L, eds. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In: *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 2021.
62. Narayanan D, Shoeybi M, Casper J, LeGresley P, Patwary M, Korthikanti V, et al., eds. Efficient large-scale language model training on gpu clusters using megatron-lm. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021.
63. Hisamoto S, Post M, Duh K. Membership inference attacks on sequence-to-sequence models: is my data in your machine translation system? *Trans Assoc Comput Linguist* 2020;**8**:49–63. https://doi.org/10.1162/tacl_a_00299
64. Fredrikson M, Jha S, Ristenpart T, eds. Model inversion attacks that exploit confidence information and basic countermeasures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.
65. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al., eds. Extracting training data from large language models. In: *USENIX Security Symposium*, 2021.
66. ENISA. *ENISA Threat Landscape: Health Sector*. 2023. <https://www.enisa.europa.eu/publications/health-threat-landscape>
67. HIPAA. *Healthcare Data Breach Statistics*. 2023. <https://www.hipaajournal.com/healthcare-data-breach-statistics/>
68. Mansfield-Devine S. *IBM: Cost of a Data Breach*. MA Business London, 2022.
69. Huang Y, Gupta S, Zhong Z, Li K, Chen D. Privacy implications of retrieval-based language models. arXiv, <https://doi.org/10.48550/arXiv.2305.14888>, 24 May 2023, preprint: not peer reviewed.
70. Lukas N, Salem A, Sim R, Tople S, Wutschitz L, Zanella-Béguelin S. Analyzing leakage of personally identifiable information in language models. arXiv, <https://doi.org/10.48550/arXiv.2302.00539>, 1 February 2023, preprint: not peer reviewed.
71. Tan S. *Web Service for Deduce, a Python Library for De-identification of Dutch Medical Text*. 2023 [updated 31-01-2023]. <https://github.com/umcu/deduce-service>
72. Honnibal M, Montani I. Spacy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2020;**7**:411–20. <https://doi.org/10.5281/zenodo.1212303>
73. Verkijk S, Vossen P, eds. Efficiently and thoroughly anonymizing a transformer language model for Dutch electronic health records: a two-step method. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022.
74. Dinh MH, Fioretto F. Context-aware differential privacy for language modeling. arXiv, <https://doi.org/10.48550/arXiv.2301.12288>, 28 January 2023, preprint: not peer reviewed.
75. NHS. *Secure Data Environment Service*. 2023 [updated 28-07-2023]. <https://digital.nhs.uk/services/secure-data-environment-service>
76. anDREa. *Trusted Research Environment Unburdening & Empowering Organizations*. 2023. <https://andrea-cloud.eu/>
77. Commission E. *Proposal for a Regulation—The European Health Data Space*. 2022. https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en
78. Mauran C. *Whoops, Samsung Workers Accidentally Leaked Trade Secrets via ChatGPT*. 2023. <https://mashable.com/article/samsung-chatgpt-leak-details>
79. Matefi R. The artificial intelligence impact on the rights to equality and non-discrimination in the light of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *Rev Universul Juridic* 2021;130–6.
80. Bill C. 27. *An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to Make Consequential and Related Amendments to Other Acts*. House of Commons of Canada. Minister of Innovation, Science and Industry, 2022.
81. *Chapter 2—Federal Trade Commission; Promotion of Export Trade and Prevention of Unfair Methods of Competition*. 2022. <https://uscode.house.gov/view.xhtml?path=/prelim@title15/chapter2&edition=prelim>
82. Rohit C. *Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems*. 2022. <https://www.eeoc.gov/joint-statement-enforcement-efforts-against-discrimination-and-bias-automated-systems>
83. Health and Human Services UDO. *HIPAA for Professionals*. 2013. <https://www.hhs.gov/hipaa/for-professionals/index.html>
84. White HT. *Applying the Blueprint for an AI Bill of Rights*. 2023. <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>
85. Ahmad T, Desai NR, Yamamoto Y, Biswas A, Ghazi L, Martin M, et al. Alerting clinicians to 1-year mortality risk in patients hospitalized with heart failure: the REVEAL-HF randomized clinical trial. *JAMA Cardiol* 2022;**7**:905–12. <https://doi.org/10.1001/jamacardio.2022.2496>
86. Ghazi L, Yamamoto Y, Riello RJ, Coronel-Moreno C, Martin M, O’Connor KD, et al. Electronic alerts to improve heart failure therapy in outpatient practice: a cluster randomized trial. *J Am College of Cardiol* 2022;**79**:2203–13. <https://doi.org/10.1016/j.jacc.2022.03.338>
87. Edward M. Clinical investigation of Wysa. In: Plymouth Uo, ed. clinicaltrials.gov/2023.
88. Nazanin A. Comparing clinical decision-making of AI technology to a multi-professional care team in eCBT for depression. In: University Qs, ed. clinicaltrials.gov/2022.
89. Sarah K. CBTpro: scaling up CBT for psychosis using simulated patients and spoken language technologies (CBTpro). In: Washington Uo, ed. clinicaltrials.gov/2023.
90. University WMCoC. Chatbot to maximize hereditary cancer genetic risk assessment. In: University WMCoC, ed. clinicaltrials.gov/2023.
91. Medicine ACCaP. ChatBot and activity monitoring in patients undergoing chemoradiotherapy. In: Medicine ACCaP, ed. clinicaltrials.gov/2023.
92. University Y. Artificial intelligent clinical decision support system simulation center study for technology acceptance. In: University Y, ed. clinicaltrials.gov/2023.
93. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. *Nature* 2023;**619**:357–62. <https://doi.org/10.1038/s41586-023-06160-y>
94. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;**183**:589–96. <https://doi.org/10.1001/jamainternmed.2023.1838>
95. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. Future-ai: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv, <https://doi.org/10.48550/arXiv.2109.09658>, 20 September 2021, preprint: not peer reviewed.
96. Magdziarczyk M, ed. Right to be forgotten in light of regulation (eu) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. In: *6th International Multidisciplinary Scientific Conference on Social Sciences and Art Sgem* 2019, 2019.
97. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept “black box” medicine? *Ann Intern Med* 2020;**172**:59–60. <https://doi.org/10.7326/M19-2548>
98. Cuttillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med* 2020;**3**:47. <https://doi.org/10.1038/s41746-020-0254-2>
99. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A, eds. What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Machine Learning for Healthcare Conference*. PMLR, 2019.
100. Vig J, ed. Bertviz: a tool for visualizing multihead self-attention in the BERT model. In: *ICLR Workshop: Debugging Machine Learning Models*, 2019.

101. Bastings J, Filippova K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? arXiv, <https://doi.org/10.48550/arXiv.2010.05607>, 12 October 2020, preprint: not peer reviewed.
102. Alammargithub J. *Interfaces for Explaining Transformer Language Models*. Jay Alammargithub, 2020. <https://jalammargithub.io/explaining-transformers>
103. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;**3**:e745–e50. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
104. Morgan A. 2023. [cited 2023]. <https://www.comet.com/site/blog/explainable-ai-for-transformers/>
105. Cinà G, Fernandez-Llaneza D, Mishra N, Röber TE, Pezzelle S, Calixto I, et al. Fixing confirmation bias in feature attribution methods via semantic match. arXiv, <https://doi.org/10.48550/arXiv.2307.00897>, 3 July 2023, preprint: not peer reviewed.
106. Cinà G, Röber TE, Goedhart R, Birbil ŞI. Semantic match: Debugging feature attribution methods in XAI for healthcare. arXiv, <https://doi.org/10.48550/arXiv.2301.02080>, 5 January 2023, preprint: not peer reviewed.
107. Behnamoh. *Is it Just Me or GPT-4's Quality Has Significantly Deteriorated Lately?* 10-08-2023 ed. Hacker News, 2023. <https://news.ycombinator.com/item?id=36134249>
108. Sam-998. *Has Anyone Noticed a Decline in GPT-4's Performance Recently?* 10-08-2023 ed. Reddit, 2023. https://www.reddit.com/r/ChatGPT/comments/13oonr9/has_anyone_noticed_a_decline_in_gpt4s_performance/
109. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv, <https://doi.org/10.48550/arXiv.2307.09009>, 18 July 2023, preprint: not peer reviewed.
110. Gomez D, Quijano N, Giraldo LF. Learning transferable concepts in deep reinforcement learning. arXiv, <https://doi.org/10.48550/arXiv.2005.07870>, 16 May 2020, preprint: not peer reviewed.
111. Shanahan M, Mitchell M. Abstraction for deep reinforcement learning. arXiv, <https://doi.org/10.48550/arXiv.2202.05839>, 10 February 2022, preprint: not peer reviewed.
112. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. arXiv, <https://doi.org/10.48550/arXiv.2009.03300>, 7 September 2020, preprint: not peer reviewed.
113. Shanahan M, Crosby M, Beyret B, Cheke L. Artificial intelligence and the common sense of animals. *Trends Cogn Sci* 2020;**24**:862–72. <https://doi.org/10.1016/j.tics.2020.09.002>
114. Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv, <https://doi.org/10.48550/arXiv.2206.04615>, 9 June 2022, preprint: not peer reviewed.
115. HuggingFace. *Transformers Agent*. 2023. https://huggingface.co/docs/transformers/transformers_agents
116. Murphy RM, Dongelmans DA, Yasrebi-de Kom I, Calixto I, Abu-Hanna A, Jager KJ, et al. Drug-related causes attributed to acute kidney injury and their documentation in intensive care patients. *J Crit Care* 2023;**75**:154292. <https://doi.org/10.1016/j.jcrc.2023.154292>
117. Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: a scoping review of supervised learning methods. *PLoS One* 2023;**18**:e0279842. <https://doi.org/10.1371/journal.pone.0279842>
118. Mellia JA, Basta MN, Toyoda Y, Othman S, Elfanagely O, Morris MP, et al. Natural language processing in surgery: a systematic review and meta-analysis. *Ann Surg* 2021;**273**:900–8. <https://doi.org/10.1097/SLA.0000000000004419>
119. Subramanian S, Baldini I, Ravichandran S, Katz-Rogozhnikov DA, Ramamurthy KN, Sattigeri P, et al., eds. A natural language processing system for extracting evidence of drug repurposing from scientific publications. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
120. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al., eds. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, 2019.
121. Gattepaille LM, Hedfors Vidlin S, Bergvall T, Pierce CE, Ellenius J. Prospective evaluation of adverse event recognition systems in Twitter: results from the Web-RADR project. *Drug Saf* 2020;**43**:797–808. <https://doi.org/10.1007/s40264-020-00942-3>
122. Yu D, Vidyiswaran VV. An assessment of mentions of adverse drug events on social media with natural language processing: model development and analysis. *JMIR Med Inform* 2022;**10**:e38140. <https://doi.org/10.2196/38140>
123. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;**23**:bbac409. <https://doi.org/10.1093/bib/bbac409>
124. van Mens HJ, van Eysden MM, Nienhuis R, van Delden JJ, de Keizer NF, Cornet R. Evaluation of lexical clarification by patients reading their clinical notes: a quasi-experimental interview study. *BMC Med Inform Decis Mak* 2020;**20**:1–8. <https://doi.org/10.1186/s12911-020-01286-9>
125. Schubbe D, Scalia P, Yen RW, Saunders CH, Cohen S, Elwyn G, et al. Using pictures to convey health information: a systematic review and meta-analysis of the effects on patient and consumer health behaviors and outcomes. *Patient Educ Couns* 2020;**103**:1935–60. <https://doi.org/10.1016/j.pec.2020.04.010>
126. Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. *Sci Rep* 2023;**13**:4569. <https://doi.org/10.1038/s41598-023-31341-0>
127. Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, Alyafei Z, et al. Multitask prompted training enables zero-shot task generalization. arXiv, <https://doi.org/10.48550/arXiv.2110.08207>, 15 October 2021, preprint: not peer reviewed.
128. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 2023;**55**:1–35. <https://doi.org/10.1145/3566815>
129. Zuccon G, Koopman B. Dr ChatGPT, tell me what I want to hear: how prompt knowledge impacts health answer correctness. arXiv, <https://doi.org/10.48550/arXiv.2302.13793>, 23 February 2023, preprint: not peer reviewed.
130. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;**388**:1233–9. <https://doi.org/10.1056/NEJMr2214184>
131. Van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023;**614**:224–6. <https://doi.org/10.1038/d41586-023-00288-7>
132. Abdelnabi S, Fritz M, eds. Adversarial watermarking transformer: towards tracing text provenance with data hiding. In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021.
133. Schinkel M, Paranjape K, Nanayakkara P. Written by humans or artificial intelligence? That is the question. *Ann Intern Med* 2023;**176**:572–3. <https://doi.org/10.7326/M23-0154>
134. Bills S, Cammarata N, Mossing D, Tillman H, Gao L, Goh G, et al. *Language Models Can Explain Neurons in Language Models*. OpenAI, 2023. <https://openai.com/research/language-models-can-explain-neurons-in-language-models>
135. Searle T, Ibrahim Z, Teo J, Dobson R. Summarisation of electronic health records with clinical concept guidance. arXiv, <https://doi.org/10.1016/j.jbi.2023.104358>, 14 November 2022, preprint: not peer reviewed.
136. Searle T, Ibrahim Z, Teo J, Dobson RJ. Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. *J Biomed Inform* 2023;**141**:104358. <https://doi.org/10.1016/j.jbi.2023.104358>
137. Farajidavar N, O'Gallagher K, Bean D, Nabeebaccus A, Zakeri R, Bromage D, et al. Diagnostic signature for heart failure with preserved ejection fraction (HFpEF): a machine learning approach using multi-modality electronic health record data. *BMC Cardiovasc Disord* 2022;**22**:1–13. <https://doi.org/10.1186/s12872-022-03005-w>
138. Sammani A, Bagheri A, van der Heijden PG, Te Riele AS, Baas AF, Oosters C, et al. Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks. *NPJ Digit Med* 2021;**4**:37. <https://doi.org/10.1038/s41746-021-00404-9>
139. Idnay B, Dreisbach C, Weng C, Schnall R. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *J Am Medical Inform Assoc* 2022;**29**:197–206. <https://doi.org/10.1093/jamia/ocab228>
140. Weis JM, Levy PC. Copy, paste, and cloned notes in electronic health records. *Chest* 2014;**145**:632–8. <https://doi.org/10.1378/chest.13-0886>
141. Thornton JD, Schold JD, Venkateshaiah L, Lander B. Prevalence of copied information by attendings and residents in critical care progress notes. *Crit Care Med* 2013;**41**:382–8. <https://doi.org/10.1097/CCM.0b013e3182711a1c>
142. Rule A, Bedrick S, Chiang MF, Hribar MR. Length and redundancy of outpatient progress notes across a decade at an academic medical center. *JAMA Netw Open* 2021;**4**:e2115334. <https://doi.org/10.1001/jamanetworkopen.2021.15334>
143. Gabriel RA, Kuo T-T, McAuley J, Hsu C-N. Identifying and characterizing highly similar notes in big clinical note datasets. *J Biomed Inform* 2018;**82**:63–9. <https://doi.org/10.1016/j.jbi.2018.04.009>
144. Gantzer HE, Block BL, Hobgood LC, Tufte J. Restoring the story and creating a valuable clinical note. *Ann Intern Med* 2020;**173**:380–2. <https://doi.org/10.7326/M20-0934>
145. Cheng C-G, Wu D-C, Lu J-C, Yu C-P, Lin H-L, Wang M-C, et al. Restricted use of copy and paste in electronic health records potentially improves healthcare quality. *Medicine (Baltimore)* 2022;**101**:e28644. <https://doi.org/10.1097/MD.00000000000028644>
146. Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. *Appl Clin Inform* 2017;**26**:12–34. <https://doi.org/10.4338/ACI-2016-09-R-0150>
147. Searle T, Ibrahim Z, Teo J, Dobson R. Estimating redundancy in clinical text. *J Biomed Inform* 2021;**124**:103938. <https://doi.org/10.1016/j.jbi.2021.103938>
148. Liu J, Capurro D, Nguyen A, Verspoor K. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *J Biomed Inform* 2022;**133**:104149. <https://doi.org/10.1016/j.jbi.2022.104149>



Patients on ENTRESTO experience fewer hospitalisations, reduced risk of CV death and improved QoL versus ACEi (enalapril)^{*3-7}

QoL based on post hoc analysis^{6,7}

Current, expert-led ESC guidelines recommend ENTRESTO as a first-line treatment option for eligible patients with symptomatic chronic HFrEF in combination with a BB, SGLT2i and MRA⁸

Explore ENTRESTO further



~2% of the NHS budget is spent on HF^{†9,10}



~70% of the cost of HF to the NHS is due to hospitalisation⁹

Versus ACEi (enalapril), at a median follow-up of 27 months, ENTRESTO significantly reduced the risk of:^{‡3}

Composite of death from CV causes or first hospitalisation for worsening HF
20% RRR (ARR=4.7%; p<0.001)

Death from CV causes
20% RRR (ARR=3.1%; p<0.001)

First hospitalisation for worsening HF
21% RRR (ARR=2.8%; p<0.001)

Starting ENTRESTO first-line could add 1 to 2 years to patients' lives vs ACEi⁴

Based on actuarial estimates from the PARADIGM-HF trial, and assuming that protective effects of ENTRESTO remain consistent with long-term use; extrapolated from available short-term follow-up data. Results were found in patients who were 45-75 years of age.⁴

The most commonly reported adverse reactions with ENTRESTO were hypotension (17.6%), hyperkalaemia (11.6%) and renal impairment (10.1%); angioedema was reported in patients treated with ENTRESTO (0.5%; uncommon).^{1,2}

For further safety information, please refer to the Summary of Product Characteristics^{1,2}

ACEi, angiotensin converting enzyme inhibitor; ARR, absolute risk reduction; BB, beta blocker; CV, cardiovascular; DHSC, Department of Health and Social Care; EF, ejection fraction; ESC, European Society of Cardiology; HF, heart failure; HFrEF, heart failure with reduced ejection fraction; MRA, mineralocorticoid receptor agonist; QoL, quality of life; RCT, randomised controlled trial; RR, risk reduction; SGLT2i, sodium-glucose cotransporter 2.

*PARADIGM HF (N=8,442) was a double-blind RCT of patients with class II, III or IV HF and an EF of ≤40% randomised to receive either ENTRESTO (200 mg twice daily) or enalapril (10 mg twice daily) in addition to recommended therapy. Primary outcome was a composite of death from CV causes or hospitalisation for HF¹; †NHS budget 2020-2021 based on DHSC departmental expenditure limit of £130.38 billion¹⁰; ‡N=8,399.

References: 1. ENTRESTO GB Summary of Product Characteristics. Available at: <https://www.medicines.org.uk/emc/product/7751/smpc> [Accessed September 2023]; 2. ENTRESTO NI Summary of Product Characteristics. Available at: <https://www.emcmedicines.com/en-gb/northernireland/medicine?id=a1393009-6872-4eb0-9aaf-e19f456e9dff&type=smpc> [Accessed September 2023]; 3. McMurray JJV, et al. *N Engl J Med* 2014;371(11):993-1004; 4. Claggett B, et al. *N Engl J Med* 2015;373(23):2289-2290; 5. Solomon SD, et al. *JACC Heart Fail* 2016;4(10):816-822; 6. Lewis EF, et al. *Circ Heart Fail* 2017;10(8):e003430; 7. Chandra A, et al. *JAMA Cardiol* 2018;3(6):498-505; 8. McDonagh TE, et al. *Eur Heart J* 2021;42(36):3599-3726; 9. National Institute for Health and Care Excellence. NG106: Chronic heart failure in adults: diagnosis and management. Available at: <https://www.nice.org.uk/guidance/ng106> [Accessed September 2023]; 10. Department of Health and Social Care. DHSC annual report and accounts: 2018 to 2019. Available at: <https://www.gov.uk/government/publications/dhsc-annual-report-and-accounts-2018-to-2019> [Accessed September 2023].

Adverse events should be reported. Reporting forms and information can be found at www.mhra.gov.uk/yellowcard. Adverse events should also be reported to Novartis via uk.patientsafety@novartis.com or online through the pharmacovigilance intake (PVI) tool at www.novartis.com/report. If you have a question about the product, please contact Medical Information on 01276 698370 or by email at medinfo.uk@novartis.com