When Are LLMs for Clinical Documentation Considered Medical Devices Under the European Medical Device Regulation (MDR)?

Andrea Blatti*

ABSTRACT: This article examines how AI systems that assist but do not provide a final diagnosis are classified under the EU Medical Device Regulation (MDR). With the rise of AI in healthcare, especially in managing EHRs and clinical documentation, the paper reconsiders regulatory interpretations in light of new developments. These include the shift from general-purpose LLMs to specialised medical models, like MedGemma, and the 2025 MDCG guidance, which updates classification rules for software. Addressing a gap in current literature, the paper analyses the regulatory status of AI tools participating in the diagnostic process that do not provide the final diagnosis.

KEYWORDS: medical device; MDCG; MDR AI; clinical documentation; AI scribe

1. Introduction

Integrating artificial intelligence (AI) into healthcare is rapidly transforming clinical workflows, particularly in the management of electronic health records (EHR) and clinical documentation.[1] The subjects of this paper are AI systems that participate in the diagnostic process without providing the final diagnosis. In other words, this analysis will focus on the regulatory classification of AI systems under the European Union's Regulation 2017/745 on medical devices[2] (MDR), which form the knowledge basis for physicians to autonomously diagnose the presence or absence of a pathology.

* *P.h.D student at Sant'Anna School of Advanced studies. Mail: andrea.blatti@santannapisa.it.* *The article was subject to a double-blind peer review process.*

[1] A. BRACKEN, et al., *Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review*, in *Journal of medical systems*, 49(28)/2025; A.R. BONGURALA, et al., *Transforming health care with artificial intelligence: redefining medical documentation*, in *Mayo clinic proceedings: digital health*, 2(3)/2024.

[2] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EE.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

Establishing whether a software falls within the MDR's scope is a notoriously challenging task. The determination is critical because it dictates the entire regulatory pathway. While many have highlighted the friction between the EU Regulation 2024/1689 on artificial intelligence[3] (AI Act) and the MDR,[4] the antecedent question must be answered first: does the MDR even apply to these assistive AI tools?

AI-based clinical documentation tools do not suggest a diagnosis or a given therapy, but rather curate, structure, and present information, thereby acting as sophisticated intermediaries between raw clinical data and the clinician's final judgment.

This paper will use two of the most prominent and marketed types of AI systems for EHR curation as representative examples: speech recognition (SR) tools and summarisation systems.

AI speech recognition tools aim to accelerate documentation activities, which have been proven to consume precious time that could be dedicated to patient care, sometimes even leading to physician burnout[5]. By seamlessly converting spoken language into structured text within the EHR, they streamline critical but burdensome administrative tasks.[6]

AI-powered summarisation systems are particularly useful to extract the large volumes of unstructured clinical data, such as narrative notes, reports, and discharge letters, that constitute patients' clinical documentation.[7] These systems can retrieve key health information and generate concise, coherent summaries, enabling clinicians to quickly grasp a patient's history and status without sinking into infinite documentation.[8]

While both types of systems serve the overarching goal of optimising the physician's documentation activities, they pose a common and complex regulatory question regarding their status as medical devices.

This research is motivated by several converging developments in technology, regulation, and clinical practice that render previous analyses incomplete and necessitate a renewed examination of the topic. There are six primary motivations for this paper.

First, the technological landscape of the AI models underpinning these systems has evolved significantly: recent discussions on the regulatory status of AI in medicine used to focus on general-purpose Large Language Models (LLMs), such as early versions of GPT.[9] Under EU law, software developed for general purposes without specific medical purposes typically does not qualify as a medical device.[10] However, the

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828.

[4] Ex multis, see D. ONITIU, S. WACHTER, B. MITTELSTADT, *How AI challenges the medical device regulation: patient safety, benefits, and intended uses*, in *Journal of Law and the Biosciences*, 2024.

[5] J.J.W. NG, et al., *Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review*, in *BMC Medical Informatics and Decision Making*, 25(236)/2025, 2.

[6] A.R. BONGURALA, et al., *op. cit*., 343-344.

[7] See J.D. OLIVEIRA, et al., *Development and evaluation of a clinical note summarization system using large language models*, in *Communications medicine*, 5(376)/2025.

[8] M. ALKHALAF, et al., *Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records*, in *Journal of biomedical informatics*, 156/2024, 2, 4, 7.

[9] Ex multis, see K. LUDVIGSEN, S. NAGARAJA, A. DALY, *When Is Software a Medical Device? Understanding and Determining the "Intention" and Requirements for Software as a Medical Device in European Union Law*, in *European journal of risk regulation*, 13(1)/2021, 83.

[10] D. WORKUM, et al., *Is my LLM application considered a medical device under the MDR?*, in Ssrn, 2025, 10. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5202088 (last access: September 8, 2025); T. MINSSEN, E.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

current reality is far more nuanced. We are witnessing a rapid proliferation of LLMs that are no longer solely general-purpose: the field has advanced to include models that are fine-tuned for specific medical objectives and, more recently, foundational models built from the ground up for medical purposes.[11] Google's Med-PaLM2 was an early example of a model trained with curated medical data. The recent development of models like MedGemma represents a further leap: as its creators noted, "while general-purpose (non-medically tuned) LMMs demonstrate impressively broad abilities, generic models can lack nuanced medical understanding and the ability to interpret and reason about medical data in a robust way… Recognising this gap, we created MedGemma, a new suite of open, medically-tuned, vision-language foundation models the MedGemma models are designed to interpret and reason about medical images and text while retaining the strong general-purpose capabilities present in Gemma 3".[12]

This shift from general-purpose to context-specific medical AI fundamentally alters the regulatory results achieved in the past and requires a new evaluation of how these systems should be classified.

Second, the 2019 Medical Device Coordination Group (MDCG) guidance on software classification was severely criticised in relation to software as medical devices classification rules; for this reason, the just released 2025 guide deserves great attention.

Third, due to the approval of the European Health Data Space regulation in Europe, electronic health records are likely to spread throughout the Union, increasing documentation duties and the time required for them.[13] Such an effect has already been studied in the USA, where the whole academic community has witnessed the relationship between documentation time and burnout.[14] This paper takes inspiration from that context, trying to foresee the possible consequences in the European context.

Fourth, there is a significant gap in the academic and legal literature concerning AI systems that do not perform a final diagnosis. The majority of existing research focuses on AI that provides diagnostic or prognostic outputs;[15] thus, this paper will seek to fill this void by focusing exclusively on this underexamined software category. Moreover, when scholars approached AI systems that do not provide the final diagnosis, they were not inclined to consider them medical devices;[16] In this research, the opposite will be argued.

VAYENA, G. COHEN, *The challenges for regulating medical use of chatgpt and other large language models*, in *Jama*, 330(4)/2023, 2.

[11] For an historical excursus see H.R. SAEIDNIA, M. NILASHI, *From MYCIN to MedGemma: A Historical and Comparative Analysis of Healthcare AI Evolution*, in *InfoScience trends*, 2/2025.

[12] A. SELLERGREN, et al., *Medgemma technical report*, in *arXiv*, 2025, 2.

[13] D. FÅHRAEUS, J. REICHEL, S. SLOKENBERGA, *The European health data space: challenges and opportunities*, in *Sieps*, 2024, 11.

[14] See E. GESNER, P. GAZARIAN, P. DYKES, *The Burden and Burnout in Documenting Patient Care: An Integrative Literature Review*, in L. OHNO-MACHADO, B. SÈROUSSI, *MEDINFO 2019: Health and Wellbeing e-Networks for All*, Proceedings of the 17th World Congress on Medical and Health Informatics, 2019.

[15] For example, see J.D. WORKUM, et al., *op. cit*., 9.

[16] See R.C. SCHOONBEEK, et al., *Completeness, Correctness and Conciseness of Physician-written versus Large Language Model Generated Patient Summaries Integrated* in *Electronic Health Records, in Preprints with The Lancet*, 2024, 6. Equally, in U.S. they are still not considered medical devices, see K.E. GOODMAN, P.H. YI, D.J. MORGAN, *AI-Generated Clinical summaries require more than accuracy*, in *Jama*, 331(8)/2024, 637; S.A. MESS, A.J. MACKEY, D.E. YAROWSKY, *Artificial Intelligence Scribe and Large Language Model Technology in Healthcare Documentation: Advantages, Limitations, and Recommendations*, in *Plastic and reconstructive surgery*, in *Global open*, 13(1)/2025, 3.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

Fifth, the practical urgency of this issue has intensified. These AI systems are no longer theoretical constructs; they are being actively deployed in clinical settings, including within publicly accessible healthcare facilities, particularly in the United States.[17] Their increasing availability and use within national health system facilities make clarifying their regulatory status not merely an academic exercise but a practical necessity to ensure patient safety and clarify legal responsibilities.[18]

Finally, the classification of these AI systems carries profound consequences for the applicable legislative framework. The AI systems under analysis surely fall within the scope of the EU's AI Act, given their reliance on technologies like LLMs that fit the AI Act's broad definition of an AI system.[19] However, their potential collocation under the MDR is far less clear.

Under the AI Act, high-risk AI systems are regulated by burdensome provisions. However, it is always the provider assessing the risk class of her AI systems. As noted, it is foreseeable that providers will classify their systems as non-high risk systems so to avoid the most demanding discipline.[20] AI systems capable of working in healthcare settings can be classified as high-risk systems only if they are medical devices according to MDR or IVDR, or if they are used for the activities outlined by annex III AI Act. For all the activities not covered by annex III, it will be crucial to assess if the system in question is a medical device or not (both under MDR and IVDR).

The paper is divided as follows: section 1 provides an overview of the landscape of AI scribe systems, highlighting their utility and risks. Section 2 describes the salient passages of the new guidance on the classification of software as medical devices, outlining the specific procedure to be followed in order to establish whether software must be classified as a medical device. The following sections will discuss each step, superseding on the first two.

Section 3 delves into the third passage, focused on the action that the software pursues on data. To discuss how LLMs work on data, a clarification of their historical development and their modern functioning is provided, exemplifying how SR and summarisation AI systems produce their output.

Section 4 focuses on whether these systems provide an individual benefit to patients. Before the conclusions, section 5 will instead clarify whether these AI systems serve medical purposes in the meaning of art. 2 MDR.

---

[17] G. Burke, H. Schellmann, *Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said,* in *APnews*, 2024.

[18] J. Ge, M.B. Delk, J.C. Lai, *A Comparison of a Large Language Model vs Manual Chart Review for the Extraction of Data Elements From the Electronic Health Record,* in *Gastroenterology*, 166(4)/2024, 708.

[19] European Law Institute, *The concept of 'AI system' under the new AI Act: Arguing for a Three-Factor Approach*, 2024, 15; for a critical view, see A. Rossi, et al, *The AI system definition under the AI Act, a new nomen rosae?, forthcoming,* in D. Dall'Anna, G. Gezici, G. Rossetti (edited by), *Proceedings of HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence* (HHAI), Pisa, 9–13 June 2025.

[20] B. Solaiman, A. Malik, *Regulating algorithmic care in the European Union: evolving doctor–patient models through the Artificial Intelligence Act (AI-Act) and the liability directives,* in *Medical law review*, 33/2025, 11.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

2. AI for Clinical Documentation

2.1. Clinical Documentation Problems

Current EHRs, while allowing for efficient data storage, retrieval, and transfer, suffer from usability challenges that undermine the primary expectation for the EHR, which is to help clinicians find the necessary information to deliver care.[21]

Generally, data stored in EHRs can be subgrouped into structured, unstructured, and semi-structured data. Structured data refers to information organised and presented in a standardised format, permitting easy storage, retrieval, and analysis.[22] The structure follows predefined schemes, ensuring consistency and interoperability between EHR systems.[23]

Unstructured data, conversely, are not arranged according to predetermined schemes, free-text narratives, such as clinical notes, scanned documents, images, etc..[24] Semi-structured data, as the name suggests, is stored in hybrid sections combining predefined data fields with varying degrees of flexibility in allowing the insertion of additional information.[25]

Since the beginning, early EHRs' adoption was accompanied by great enthusiasm; the improvement of patient care combined with reduced healthcare costs seemed to be within reach; however, the opposite has occurred[26]. The widespread adoption of EHRs in USA has expanded clinical documentation to unprecedented levels, resulting in a significantly challenging burden for the entire healthcare system.[27] Indeed, modern EHRs require physicians to keep record of basically every patient interaction: whether it be compiling diagnostic reports, writing progress notes or synthesising a patient's treatment history across different specialists[28]. Despite the promise of having a better knowledge basis for clinical decision-making, this workload significantly increased clinicians' burden.[29] As a result, healthcare systems are under

[21] M.W. FRIEDBERG, et al., *Factors Affecting Physician Professional Satisfaction and Their Implications for Patient Care, Health Systems, and Health Policy* in *Rand health quarterly*, 3(4)/2014; see also Y. A. KUMAH-CRYATAL, *Electronic Health Record Interactions through Voice: A Review*, in *Applied clinical informatics*, 9/2018, 542.

[22] See S.I. SHYLA, V.R.B. BLESSIE, *An introduction to electronic health records,* in P.M. FATHIMAL, et al. (edited by), *Advances of Machine Learning for Knowledge Mining,* in *Electronic Health Records*, 2025, 4.

[23] S.I. SHYLA, V.R.B. BLESSIE, *op. cit*., 5.

[24] Unstructured data can be medical images such as X-rays scan, medical notes, evaluation reports, social media posts, speech therapy sessions in audio format, faxed versions of structured data etc., see S.I. SHYLA, V.R.B. BLESSIE, *op. cit*., 5.

[25] S.I. SHYLA, V.R.B. BLESSIE, *op. cit*., 7.

[26] C. ROSE, J.H. CHEN, *Learning from the EHR to implement AI in healthcare* in *npj Digital Medicine*, 330/2024, 1.

[27] A.A. TIERNEY, et al., *Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation,* in Nejm catalyst Innovations in Care Delivery, 5(3)/2024, 2.
EHRs documentation duties must be considered in addition to others: see E. JOUKES, et al., Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record, in Applied clinical informatics, 9(45)/2018.

[28] D. VAN VEEN et al., op. cit., 1134.

[29] F. MORAMARCO et al., Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, 2022, 5739.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

increasing pressure because of growing patient' data volumes.[30] According to scholars, clinical documentation has become the most time-consuming aspect of EHR use.[31]

One of the major issues is the unfriendly design and complexity of existing EHR systems.[32] As extensive literature shows, the time spent on EHRs' documentation is unsustainable for healthcare facilities. The amount of time clinicians spend on documentation is disputed, as it can vary depending on the practice or hospital being considered.[33] Research has shown that it can vary from two[34] hours per day, to 35%[35] or even 50%[36] of the working time (rounded figures).[37] This massive workload led to two important phenomena: clinicians' burnout[38] and flow in patient care.[39] Focusing on the latter consequence, the impact is twofold. Firstly, physicians have less time to focus on patient care,[40] and secondly, the poor design of the EHRs complicates data entry activities, leading to low-quality information[41]. Errors in EHRs are common, and part of the reason is the overburden that clinicians have to carry.[42] It is worth noting that for these reasons some clinicians resist the digitalisation of data, preferring traditional paper documentation

[30] D. Anderson, et al., *Paging Dr. GPT: Extracting Information from Clinical Notes to Enhance Patient Predictions*, in *arXiv*, 2025,1.

[31] Deficiencies of EHRs were described in M.W. Friedberg, et al., *op. cit*.

[32] K.E. Goodman, P.H. Yi, D.J. Morgan, *op. cit*., 637; D. Karaferis, D. Balaska, Y. Pollalis, *Design and Development of Data-Driven AI to Reduce the Discrepancies in Healthcare EHR Utilization*, in *American journal of clinical and medica research*, 5(1)/2025, 3; A.A. Tierney, et al., *op. cit*., 3.

[33] L.S. Rotenstein, et al., *System-Level Factors and Time Spent on Electronic Health Records by Primary Care Physicians*, in *Jama Network*, 6(11)/2023, 2.

[34] A. Gaffney, et al., *Medical Documentation Burden Among US Office-Based Physicians in 2019: A National Study*, in *JAMA Internal Medicine*, 182(5)/2022, 565.

[35] M.D. Tipping, *Where did the day go?--a time-motion study of hospitalists*, in *Journal of hospital medicine*, 5(6)/2010, 325; E. Joukes, et al., *op. cit*., 47; J.P. Avendano, et al., *Interfacing With the Electronic Health Record (EHR): A Comparative Review of Modes of Documentation*, in *Cureus*, 14(6)/2022, 1.

[36] M. Tai-Seale, et al., *Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine*, in *Health affairs*, 36/2017, 7; see also C. Sinsky, et al., *Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties*, in *Annals of internal medicine*, 165(11)/2016, 757.

[37] See S. Ghatnekar, A. Faletsky, V.E. Nambudiri, *Digital scribe utility and barriers to implementation in clinical practice: a scoping review*, in *Health and technology*, 11/2021; B.G. Arndt, et al., *Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations*, in *Annals of Family Medicine*, 15(5)/2017.

[38] C. Rose, J.H. Chen, *op. cit*., 1.

See also E. Gesner, P. Gazarian, P. Dykes, *The burden and burnout in documenting patient care: an integrative literature review*, in *Studies in health technology and informatics*, 2019; J.M. Ehrenfeld, J.P. Wanderer, *Technology as friend or foe? Do electronic health records increase burnout?*, in *Current opinion in anesthesiology*, 31(3)/2018; N. Khamisa, K. Peltzer, B. Oldenburg, *Burnout in relation to specific contributing factors and health outcomes among nurses: a systematic review*, in *International journal of environmental research and public health*, 10(6)/2013.

[39] These two phenomena are not disconnected: see K.E. Goodman, P.H. Yi, D.J. Morgan, *op. cit*., 637; A.A. Tierney, et al., *op. cit*., 2.

[40] A. Bracken, et al., *op. cit*., 2; V.S. Kumari et al., *op. cit*., 147; D. Van Veen et al., *op. cit*., 1135.

[41] C. Rose, J.H. Chen, *op. cit*., 1; A. Bracken, et al., *op. cit*., 2.

[42] S.K. Bell, et al., *Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes*, in *Jama Network*, 3(6)/2020, 2.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

methods.[43] Moreover, healthcare facilities started hiring medical scribes to let physicians focus only on patients.[44]

2.2. Advantages and Risks of AI for Clinical Documentation

Thanks to NLP techniques, it is possible today to convert unstructured data into structures, such as human speech into text (for example, patient-clinician interactions), medical information and codes from free-text notes, radiology reports, etc., so to provide ready and structured information to physicians.[45]
When evaluated, these systems highlighted undeniable advantages in terms of enhanced visit fluency and reduced documentation time, allowing clinicians to focus solely on patients.[46] Unlike AI tools designed to support clinical decision-making, which directly impact patient treatment and diagnosis, the AI systems in question are primarily aimed at enhancing the efficiency of data handling, reducing administrative burdens, and improving the overall workflow.[47] As mentioned, speech recognition and summarisation AI systems will be analysed as examples of the wide range of AI documentation-related systems.
For speech recognition tools, it could be considered Microsoft Dragon Copilot, DeepScribe, or AWS HealthScribe, which are specifically designed for the healthcare sector.
Speech recognition AI systems are tools designed to convert physicians' and patients' spoken words into text. This might happen, for example, during the first encounter for anamnesis' purposes: once transcribed, the text is inserted within the EHR as a record of the encounter.[48]
To provide a use case, consider the telemedicine guidelines issued by the Italian Ministry of Health, which explicitly refer to the chance to apply AI systems to convert the dialogues conducted through videocalls between medical doctors and patients into text.[49] In cases like this, the system is only intended to convert the spoken words into written words.[50]
In most primary healthcare practices, the record of a clinician-patient encounter is structured using the SOAP (Subjective, Objective, Assessment, Plan) model, designed to capture the patient's history, the clinician's observations, diagnosis, and management plan[51]. As said before, the traditional manual insertion of the relevant data into the EHR takes a lot of time, and this has led to outsourcing solutions. One is that

[43] V. S. KUMARI, et al., *op. cit.*, 144.
[44] Y. LIN, T.D. SHANAFELT, S.M. ASCH, *Reimagining Clinical Documentation With Artificial Intelligence*, in *Mayo Clinic Proceedings*, 93(5)/2018, 563.
[45] M. KAVITHA, K. AKILA, *op. cit.*, 56.
[46] See A. BRACKEN, et al., *op. cit.*; K. NAWAB, *Artificial intelligence scribe: a new era in medical documentation*, in *Artificial intelligence in health*, 2024; S.A. MESS, A.J. MACKEY, D.E. YAROWSKY, *op. cit.*; A.A. TIERNEY, et al., *op. cit.*; M. SASSEVILLE, et al., *The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review, in healthcare*, 13/2025, 9.
[47] S. DORN, *AI Summaries Are About To Spread Across Healthcare*, in *Forbes*, 2025. Available at https://www.forbes.com/sites/spencerdorn/2025/02/13/ai-summaries-are-about-to-spread-across-healthcare (last access: September 8, 2025).
[48] See the example of the system called Suki described in V.S. KUMARI, et al., *op. cit.*; J.P. AVENDANO et al., *op. cit.*, 1.
[49] B. STEWART, *Front-end speech*, in *Klas research*, 2015. Available at https://klasresearch.com/report/front-end-speech-what-are-the-value-adds/1027 (last access: September 8, 2025).
[50] For Dragon Copilot see B. WANG, *The application and challenges of artificial intelligence in speech recognition*, in *Proceedings of the 5th International Conference on Computing and Data Science*, 2023, 38.
[51] See P.F. PEARCE, et al., *The essential SOAP note in an EHR age*, in *The nurse practitioner*, 41(2)/2016, 30 ff.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

of professional medical transcriptionists, who manually transcribe the document based on the clinician's audio record and return it to the physician for review.[52]

The other is the use of speech recognition (SR) systems. The inefficiencies and risks of professional medical transcriptionists led to increasing interest in AI SR systems.[53] With the help of these systems, the conversation can automatically be transcribed without requiring the clinician (or a delegated actor) to write down everything she heard from the patient. Another application of SR AI systems is in procedures called self-anamnesis, where patients answer questions about their personal medical history without interacting directly with a doctor or medical assistant.[54]

While SR technology presents clear potential for cost savings and improved productivity,[55] it also introduces significant concerns.[56] The risk of medical errors and patient harm due to miscommunication or inaccuracies introduced into the permanent medical record raises critical questions about the broader implementation of such technologies, especially concerning patient safety.[57] Although speech-to-text AIs can generate written transcriptions far more rapidly than human transcribers, they are not without flaws. Uncorrected transcription errors, even as small as a misinterpreted word, may result in unclear documentation, embarrassing mistakes, or even serious patient safety risks.[58]

A particularly troubling phenomenon associated with AI transcription tools is hallucinated transcriptions, which involve software generating content that was never actually spoken.[59] These hallucinations are also non-deterministic, meaning they can yield different erroneous outputs each time the same audio is processed.[60]

A considerable body of research has documented the limitations and error rates of SR technology across medical fields. In radiology, in 2008, Quint et al. found that nearly 22% of SR-generated reports contained significant errors. Basma et al. in 2011 showed that breast imaging reports generated via SR were as likely

---

[52] F. R. Goss, L. Zhou, S. G. Weiner, *Incidence of speech recognition errors in the emergency department*, in *International journal of medical informatics*, 93/2016, 2; B.M. Fogleman, et al., *Charting tomorrow's healthcare: a traditional literature review for an artificial intelligence driven future*, in *Cureus*, 16(4)/2024, 3.

[53] For a prominent case of erroneous human transcription that led to a patient's death see *Quick Safety, Speech recognition technology translates to patient risk*, 2022. Available at https://www.jointcommission.org/en/knowledge-library/newsletters/quick-safety/issue-12 (last access: September 8, 2025). See also *the problem of sound-alike medicines in World Health Organisation, Medication safety for look-alike, sound-alike medicines*, 2023.

[54] K. Denecke, et al., *Self-anamnesis with a conversational user interface: concept and usability study,* in *Methods of information in medicine*, 57/2018.

[55] A. Koenecke, et al., *op., cit*., 1672; G.K. Patra, et al., *Voice classification in AI: harnessing machine learning for enhanced speech recognition*, in *Global Research and Development Journals*, 8(12)/2023, 4. According to the Canadian Medical Association, physicians spend around ten hours per week on administrative tasks. Based on this, in 2024, the Ottawa Hospital began using Microsoft Dragon Copilot. Available at https://www.ottawahospital.on.ca/en/newsroom/the-ottawa-hospital-uses-microsoft-ai-to-help-increase-access-to-care-and-reduce-physician-burnout/ (last access: September 8, 2025).

[56] For an overview see B. Kiran, et al., *Automatic speech recognition through artificial intelligence*, in *International journal for multidisciplinary research*, 5(6)/2023, 3.

[57] F. R. Goss, L. Zhou, S. G. Weiner, *Incidence of speech recognition errors in the emergency department*, in *International journal of medical informatics*, 93/2016, 5.

[58] Y. A. Kumah-Cryatal, *op. cit*., 543.

[59] A. Koenecke, et al., *op. cit*., 1674.

[60] A. Koenecke, et al., *op. cit*., 1674.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

as reports generated with conventional dictation transcription to contain major errors.[61] Goss et al., in 2016, identified that nearly 71% of their notes contained errors, with an average of 1.3 errors per emergency department note, and 15% of those were deemed critically significant.[62] More recently, Miner et al, in 2020, showed that SR transcripts had a higher word error rate of 25% compared to human-generated transcripts and a semantic distance of 1.2.[63] Even the modern SR AI systems, as described by Barnwal and Gupta in 2022 and Koenecke et al in 2024, are prone to significant wrong transcriptions.[64]

Among the various factors impacting transcript correctness, environmental factors (such as background noise during dictation) were also shown to negatively affect SRs' performance, further increasing error rates. At the same time, conventional transcription appears to be more resilient to such conditions.[65]

The most common error type in SR reports was the addition of unspoken words, as the software relies on recognising and 'guessing' strings of phonemes and words rather than isolated words. Additionally, word substitution, omission, and punctuation errors are common across both SR and conventional transcription systems.[66]

Furthermore, as described by Szymanski et al., "the structure of spontaneous human conversations is diametrically different from the prescriptive written language used to train language models. These models can use the grammatical structure present in the training corpora, such as part-of-speech sequences, dependency trees, and dialog acts. On the other hand, spontaneous conversations lack sentence structure. They contain repetitions, back-channelling, phatic expressions, and other artifacts of turn-taking".[67]

Still, alarmingly low transcription accuracy was highlighted for minorities: Zolnoori et al shed light on the discrimination operated by SR AI systems on black patients, whose words were systematically, wrongly transcribed.[68]

According to Topaz et al., malpractices in which SR systems played a role increased over time, and the trend is set to continue (especially given the wider application of SR systems in EHRs). To conclude, it may be worth noting that phonic issues are not only related to computer systems but are also well-known human problems.[69]

In one documented case, an AI tool misinterpreted a physician's reassurance about a lack of allergy into a false medical record indicating the presence of an allergy, an error with the potential for direct patient harm. Alondra Nelson, former head of the US White House Office of Science and Technology Policy,

---

[61] S. BASMA, et al., *Error Rates in Breast Imaging Reports: Comparison of Automatic Speech Recognition and Dictation Transcription*, in *Health care policy and quality*, 2011, 925.

[62] F. R. GOSS, L. ZHOU, S. G. WEINER, *op. cit.*, 4.

[63] A.S. MINER, et al., *Assessing the accuracy of automatic speech recognition for Psychotherapy*, in *npj Digital medicine*, 3(82)/2020, 5.

[64] S.K. BARNWAL, *Evaluation of AI system's voice recognition performance in social conversation*, in *5th International Conference on Contemporary Computing and Informatics*, 2022, 806; A. KOENECKE, et al., op. cit., 1674.

[65] S. BASMA, et al., *op. cit.*, 926.

[66] S. BASMA, et al., *op. cit.*, 926.

[67] P. SZYMANSKI, et al., *Why aren't we NER yet? Artifacts of ASR errors in named entity recognition in spontaneous speech transcripts*, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, 2023, 1746.

[68] M. ZOLNOORI, et al., *Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare*, in *JAMIA open*, 7(4)/2024, 10; see also G.K. PATRA et al., *op. cit.*, 2.

[69] World Health Organisation, Medication safety for look-alike, sound-alike medicines, 2023, 4.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

expressed concern about the use of AI transcription tools in clinical settings: "stakes are high and the consequences may be grave".

The literature about SR systems lacks incidents (fortunately), likely due to their scarce prolonged application in real clinical settings. However, recently, especially in the US, hospitals started using SR systems such as Whisper by OpenAI, and scholars began to focus on them. Research revealed Whispers' critical limitations, particularly its tendency to generate hallucinations.[70]

An investigation by the Associated Press (AP) in 2024 showed that nearly 40% of Whisper's hallucinations are not just benign inaccuracies but potentially harmful or concerning, raising serious issues about its reliability and safety.[71] As to AP, software engineers and researchers indicate that hallucinations occur in many transcripts, even in well-recorded, short audio samples.[72]

Other US hospitals, such as Mankato Clinic in Minnesota and Children's Hospital Los Angeles, have started using a Whisper-based tool built by Nabla. The problem was that the original audio was deleted, making it impossible to verify or correct erroneous transcripts.[73]

In sum, while SR technologies like Whisper offer promise, especially when fine-tuned for specific domains such as medicine, the persistence of hallucinations demands legal inquiries about their best regulation.

To sum up the challenges faced by SR AI systems it could be helpful to refer to the Kumar's review, according to which they are, mainly: medical terminology and jargon, accents and dialects, sensitive data and privacy concerns, variability in speaking styles, limited context understanding, integration with electronic health records (EHR), data bias, handling errors and misinterpretations, real-time processing, ambient noise and acoustics and variable quality in human writing.[74]

Examples of summarisation AI tools are MedGemma, Falcon by AWS, Healthcare natural language API by Google, and Arkangel AI.[75]

These systems are designed to remove the unnecessary information in long documents or EHRs, providing a ready and clear overview of patients' conditions.[76] These systems are significantly helpful for analysing unstructured clinical data, allowing for a comprehensive view of the patients' health status.[77] EHRs, if well implemented, are meant to contain extensive patient histories, requiring physicians to navigate through a vast amount of medical data.[78] Reichert et al found that EHRs' summarisation is not easy; in fact, discovering and identifying clinical information resulted in cognitively complex and time-consuming tasks.[79] Moreover, as obvious, this process introduces the possibility for errors.[80] Thanks to AI-driven summaries, it is possible to generate concise summaries that highlight only the important elements (prior diagnoses,

[70] A. KOENECKE, et al., *op. cit*., 1672.
[71] A. KOENECKE, et al., *op. cit*., 1672.
[72] G. BURKE, H. SCHELLMANN, *op. cit*.
[73] G. BURKE, H. SCHELLMANN, *op. cit*.
[74] Y. KUMAR, *A Comprehensive Analysis of Speech Recognition Systems in Healthcare: Current Research Challenges and Future Prospects*, in *SN computer science*, 5(137)/2024, 13.
[75] See https://arkangel.ai/ (last access: September 8, 2025).
[76] See J.D. OLIVEIRA, et al., *op. cit*.
[77] J.D. OLIVEIRA, et al., *op. cit*.
[78] D. KARAFERIS, D. BALASKA, Y. POLLALIS, *op. cit*., 6.
[79] D. REICHERT, D. KAUFMAN, B. BLOXHAM, et al. *Cognitive analysis of the summarization of longitudinal patient records*, in *Proceedings of the Annual American Medical Informatics Association Fall Symposium (AMIA)*, 2010, 667–671.
[80] D. VAN VEEN et al., *op. cit*., 1134.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

treatment histories, and current medication regimens, etc.), which is particularly convenient during patient consultations.[81]

Thus, applying AI systems for summarisation purposes could become an important advantage for clinicians.[82]

For instance, the study described in Van Veen et al. showed that LLM-generated summaries were often preferred over those created by medical experts due to higher completeness, correctness, and conciseness scores[83]. These promising results were also visible in Shemtob et al., where AI-generated summaries were comparable with physician-generated ones. In other cases, human-generated notes contained fewer omissions and hallucinations than those produced by the LLM.[84]

Despite the optimism surrounding these systems, several limitations of LLMs have been identified that make them unreliable for widespread clinical use. Transforming unstructured notes in EHRs into structured, meaningful summaries remains a significant technical challenge. This is due to inconsistencies in format, length variations, typographical errors, and the use of highly specialized medical language.[85]

Pal et al. witnessed models' tendency towards hallucination and high sensitivity to the prompts used.[86] In Xu et al, the AI systems omitted critical patient history-related information (such as present illness).[87]

Asgari et al. found that out of 12,999 sentences in 450 clinical notes generated by LLMs, 191 (1.47%) contained hallucinations, with 44% of these being "major", meaning that they could adversely affect patient diagnosis or therapy.[88] Moreover, of the 49,590 sentences contained in the analysed transcripts, 1,712 sentences were omitted (3.45%) in the LLM-generated notes, of which 286 (16.7%) were classified as major and 1,426 (83.3%) as minor.[89]

Hallucinations, omissions and similar mistakes were also documented concerning the most well-known AI systems.[90] For instance, concerning ChatGPT, studies have provided concrete examples of hallucinations.[91] in one case, it inserted fabricated, overly empathetic sentences into clinical summaries, such as "please know that we are here to support you every step of the way".[92] In another case, it fabricated

[81] D. KARAFERIS, D. BALASKA, Y. POLLALIS, *op. cit*., 6.
[82] K.E. GOODMAN, P.H. YI, D.J. MORGAN, *op. cit*., 637.
[83] VAN VEEN et al., *op. cit*., 1138.
[84] F. MORAMARCO et al., *op. cit*., 5744; R. SHANKAR, A. BUNDELE, A. MUKHOPADHYAY, *Natural language processing of electronic health records for early detection of cognitive decline: a systematic review*, in *npj Digital Medicine*, 8(133)/2025, 6.
[85] M. ALKHALAF, et al., *op. cit*., 2.
[86] "As the prompts are changed from ambiguous to more specific and direct, the accuracy of the tasks improved", A. PAL, L.K. UMAPATHI, M. SANKARASUBBU, Med-HALT: Medical Domain Hallucination Test for Large Language Models, in arXiv, 2023, 8.
[87] X. DU, Generative Large Language Models in Electronic Health Records for Patient Care Since 2023: A Systematic Review, in *medRxiv*, 9.
[88] E. ASGARI, et al., *A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation*, in *medRxiv*, 5.
[89] E. ASGARI, et al., *op. cit*., 5.
[90] See A. BRACKEN, et al., *op. cit*., 2; A. AARON, et al., *Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation*, in *New England Journal of Medicine Catalyst*, 5(3)/2024; S.R. ALI, et al., *Whitaker IS. Using Chat-GPT to write patient clinic letters*, in *Lancet Digit Health*, 5(4)/2023.
[91] X. DU, *op. cit*., 10.
[92] A.C. STONEHAM, et al., *Can artificial intelligence make elective hand clinic letters easier for patients to understand?* in *The Journal of hand surgery, European volume*, 49(10)/2024.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

patient names in structured thyroid ultrasound reports,[93] or misstated clinical findings, such as denying a lateral malleolus fracture that was present or inventing a ligament tear that was not.[94]
These findings reinforce the view that hallucinations and omissions may be intrinsic properties of current LLMs.[95]
Despite their high confidence, these models may output unactual or unfaithful text, which is particularly dangerous in high-stakes environments like healthcare.[96] To address this, researchers are working on frameworks to quantify and minimize the clinical impact of such errors.[97] Once identified and measured, hallucinations and omissions can be mitigated through improved prompt design, workflow integration, and engineering refinements.[98]
In conclusion, while LLMs are transforming the landscape of clinical documentation and medical summarisation, significant limitations persist, such as hallucinations and omissions, which pose real threats to patient safety and diagnostic accuracy.[99]

3. Classification Rules: the Medical Device Regulation and the MDCG

This section focuses on the MDR provisions relevant to medical devices' classification, and on the interpretative guidelines issued by the Medical Device Coordination Group in 2025. The combined reading of these two sources will require an interdisciplinary analysis, which will be conducted in sections three, four and five, focusing on both the technical features of AI and medical-legal concepts such as "diagnosis".
Turning to medical devices' classification, it must be premised that the software provider always determines whether a device is a medical device.[100]
What is a medical device is established by art. 2 (1) MDR, to which the IVDR explicitly refers in art. 2 (1), and it is meant as "means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: — diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease, — diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability, — investigation, replacement or modification of the anatomy or of a physiological or pathological process or state, — providing information by means of *in vitro* examination of specimens derived from the human body, including organ, blood and tissue donations, and which does

[93] H. JIANG et al., *Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography*, in *European Journal of Radiology*, 175/2024.
[94] J.J. BUTLER, *From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model*, in *Foot and ankle surgery*, 30(4)/2024.
[95] Z. XU, S. JAIN, M. KANKANHALLI, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, in *arXiv*, 2024.
[96] B. MESKÓ, E.J. TOPOL, *The imperative for regulatory oversight of large language models (or generative AI) in healthcare*, in *npj digital medicine*, 2023, 3.
[97] See S.M.T.I. TONMOY, et al., *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*, in *arXiv*, 2024, 2.
[98] S.M.T.I. TONMOY, et al., *op. cit*., 2024, 2.
[99] Y. A. KUMAH-CRYATAL, *op. cit*., 543.
[100] More precisely "the decision of whether a software product qualifies as a medical device is made by the developer or, using the terminology of the MDR, the manufacturer (Article 2[30] MDR)", L. KEUTZER, U. S.H. SIMMONSSON, *Medical device apps: an introduction to regulatory affairs for developers*, in *JMIR Mhealth Uhealth*, 8(6)/2022, 2-3.

Downloaded from https://teseo.unitn.it/biolaw
ISSN 2284-4503

not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means".

Moreover, software is considered "active devices" by at. 2 (4) MDR, meaning "any device, the operation of which depends on a source of energy other than that generated by the human body for that purpose, or by gravity, and which acts by changing the density of or converting that energy. Devices intended to transmit energy, substances or other elements between an active device and the patient, without any significant change, shall not be deemed active devices. Software shall also be deemed to be an active device".

Instead, according to art. 2 (2) IVDR, in vitro diagnostic medical device is "any medical device which is a reagent, reagent product, calibrator, control material, kit, instrument, apparatus, piece of equipment, software or system, whether used alone or in combination, intended by the manufacturer to be used *in vitro* for the examination of specimens, including blood and tissue donations, derived from the human body, solely or principally for the purpose of providing information on one or more of the following: (a) concerning a physiological or pathological process or state; (b) concerning congenital physical or mental impairments; (c) concerning the predisposition to a medical condition or a disease; (d) to determine the safety and compatibility with potential recipients; (e) to predict treatment response or reactions; (f) to define or monitoring therapeutic measures. Specimen receptacles shall also be deemed to be *in vitro* diagnostic medical devices". The latter definition does not apply to the systems previously described; therefore, the analysis will proceed only in relation to the MDR.

The manufacturer establishes if the software is a medical device based on the "intended purpose",[101] defined in the same way by art. 2 (12) MDR and art. 2 (12) IVDR as "the use for which a device is intended according to the data supplied by the manufacturer on the label, in the instructions for use or in promotional or sales materials or statements and as specified by the manufacturer in the clinical evaluation". The intended purpose description must include the patient's clinical benefit.[102] As clearly visible, these classification rules do not assign any role or weight to the actual use of the software, and this approach allows manufacturers choose the most suitable regulatory framework according to their interests.[103] The same was under the former directives, as stated by the European Court of Justice (C-219/11).[104]

[101] Mdcg 2019-11 Rev.1, Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR, 2025, 8; S. STOPPACHER, P. S. MULLNER, *Software as medical device In Europe*, in C. BAUMGARTNER, J. HARER, J. SCHROTTNER (edited by*), Medical device and in vitro diagnostics*, 2023, 190.

[102] L. KEUTZER, U. S.H. SIMMONSSON, *op. cit*., 3.

[103] L. SVEMPE, *Exploring impediments imposed by the medical device regulation EU 2017/745 on software as a medical device*, in *Jmir medical informatics*, 12/2024, 6.

[104] "As regards software, the legislature thus made unequivocally clear that in order for it to fall within the scope of Directive 93/42 it is not sufficient that it be used in a medical context, but that it is also necessary that the intended purpose, defined by the manufacturer, is specifically medical", Brain Products GmbH v BioSemi VOF and Others (C-219/11).

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

In 2019[105] and 2025, [106] the Medical Device Coordination Group (MDCG) issued guidelines on software classification under MDR and IVDR. The MDCG explained what is and what is not a software medical device.

Focusing on the latest guidelines, they specify that not all the software used in healthcare settings is to be considered medical devices[107], which is perfectly in line with recital 19 MDR stating that "…software for general purposes, even when used in a healthcare setting, or software intended for life-style and well-being purposes is not a medical device. The qualification of software, either as a device or an accessory, is independent of the software's location or the type of interconnection between the software and a device". For instance, simple information retrieval systems are not classified as medical devices as long as they do not serve medical purposes, even if conducted using NLP. [108]

It emerges that what really makes the difference is the presence of a medical purpose according to art. 2 MDR. Indeed, the MDCG states that "software which is intended to process, analyse, interpret calculate, create or modify medical information may be qualified as a MDSW if the creation or modification of that information is governed by a medical intended purpose".[109] Therefore, it must be assumed that software intended for activities such as processing, analysing, creating or modifying medical information for purposes other than medical ones is not a medical device.

Other examples of software not being medical device are invoicing systems, staff planning, e-mailing, web or voice messaging, data parsing, word processing, and back-up, wellness or fitness apps.[110] On the opposite, there are software searching image for findings that uphold clinical hypotheses as to the diagnosis or evolution of therapy, [111] as well as software that locally amplifies the contrast of the finding on an image display so that it serves as a decision support or suggests an action to be taken by the user, [112] or AI systems designed to influence other medical devices' functioning: it is the case of an AI software working

[105] Mdcg 2019-11, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR,* 2019.

[106] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR,* 2025.

[107] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR,* 2025, 8.

[108] "However, software would not be considered as conducting 'Simple search' if it contributes to achieving a medical purpose. E.g. User Interface (UI) search feature or a software that performs search using Natural Language Processing (NLP) where those actions contribute to achieving a medical purpose", Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR, 2025, 8.*

[109] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR,* 2025, 8.

In respect to the guidelines issued in 2019, the new one adds the reference to calculation and interpretation, 6.

[110] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 9.

[111] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR,* 2025, 8.

[112] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 8.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

in combination with a computed tomography scanner, and performing auto-contouring for the delineation between cancer lesions and healthy tissue. [113]

The guide suggests establishing whether a software is a medical device through specific steps: if it is a software, it has to be checked whether it is an MDR annex XVI device,[114] an accessory to a medical device,[115] or a software that drives or influences the use of a medical device. If so, it is a medical device. If not, it is to be determined whether the software performs an action on data beyond storage, archival, communication, simple search, or lossless compression (i.e. using a compression procedure that allows the exact reconstruction of the original data).[116] If it does, the last step is to ascertain whether this action is intended for the benefit of individual patients. Software that only aggregates population data, provides generic diagnostic or treatment pathways not directed to individual patients, or is used for scientific or epidemiological purposes is not considered to benefit individual patients and is therefore excluded,[117]

The last step involves establishing whether the software matches the definition of medical device software (MDSW) provided in the guidance. [118] It defines MDSW as a software that is intended to be used, alone or in combination, for a purpose as specified in the definition of a "medical device" in the medical devices regulation or *in vitro* diagnostic medical devices regulation".[119]

Furthermore, the guide highlights that "the risk of harm to patients, users of the software, or any other person, related to the use of the software within healthcare, including a possible malfunction is not a criterion on whether the software qualifies as a medical device".[120]

To conclude, the MDCG provides examples of software that should not be considered medical devices. It enlists electronic health records,[121] patient data management systems,[122] email systems, mobile telecommunication systems, video communication systems, paging, speech-to-text systems, etc. to transfer

[113] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 10.

[114] Annex XVI medical devices are exceptional, because they have not intended medical purposes.

[115] Accessories are defined by art. 2(2) MDR: "accessory for a medical device' means an article which, whilst not being itself a medical device, is intended by its manufacturer to be used together with one or several particular medical device(s) to specifically enable the medical device(s) to be used in accordance with its/their intended purpose(s) or to specifically and directly assist the medical functionality of the medical device(s) in terms of its/their intended purpose(s)".

[116] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025,

[117] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 12.

[118] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 12.

[119] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 7.

[120] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 9.

[121] "Electronic Health Record (EHR) systems are primarily intended to store and transfer electronic patient records, serving as repositories for various documents and data related to individual patients. These systems, when used solely to replace traditional paper-based patient files, do not meet the definition of a medical device…", Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 25.

[122] Mdcg 2019-11 Rev.1*, Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 26.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

electronic information. The communication systems mentioned are not intended medical devices because, generally, they are intended for general purposes and used for transferring both medical and non-medical information.[123]
Following, the aforementioned steps will be discussed in relation to the AI systems under analysis.
Less attention will be paid to the first two steps: whether the product at issue can be considered software or an accessory to a medical device or whether it is enlisted under annex VXI MDR (considering medical devices without medical purposes).[124] The AI systems under analysis are indeed software, and this research concerns their classification only when considered by themselves (and not in combination with other medical devices); lastly, none of the systems focused on in this paper is covered in annex XVI MDR.

4. Step 3: Actions on Data Performed by Large Language Models

The third step indicated by the guide to establish whether a software is a medical device focuses on whether the software performs an action on data beyond storage, archival, communication, simple search, or lossless compression (i.e. using a compression procedure that allows the exact reconstruction of the original data).[125] To this end, this paragraph will summarise the functioning of language models, from their beginnings to today's large language models, so as to account for their actual actions on input data.

4.1. History

The history[126] of language models began between the 1980s and 1990s with the rise of Statistical Language Models (SLMs).[127] These models approached language from a probabilistic perspective, aiming to calculate the probability of a sequence of words occurring.[128] The most prominent SLM was the n-gram

[123] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 26.
[124] "1. Contact lenses or other items intended to be introduced into or onto the eye. 2. Products intended to be totally or partially introduced into the human body through surgically invasive means for the purpose of modifying the anatomy or fixation of body parts with the exception of tattooing products and piercings. 3. Substances, combinations of substances, or items intended to be used for facial or other dermal or mucous membrane filling by subcutaneous, submucous or intradermal injection or other introduction, excluding those for tattooing. 4. Equipment intended to be used to reduce, remove or destroy adipose tissue, such as equipment for liposuction, lipolysis or lipoplasty. 5. High intensity electromagnetic radiation (e.g. infra-red, visible light and ultra-violet) emitting equipment intended for use on the human body, including coherent and non-coherent sources, monochromatic and broad spectrum, such as lasers and intense pulsed light equipment, for skin resurfacing, tattoo or hair removal or other skin treatment. 6. Equipment intended for brain stimulation that apply electrical currents or magnetic or electromagnetic fields that penetrate the cranium to modify neuronal activity in the brain", Annex XVI, points 1-6.
[125] Mdcg 2019-11 Rev.1*, Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025.
[126] For an historical analysis see D. KHURANA, et al., *Natural language processing: state of the art, current trends and challenges*, in *Multimedia tools and applications*, 82/2023.
[127] E.D. LIDDY, *Natural language processing*, in M.A. DRAKE (edited by), *Encyclopedia of Library and Information Science*, 2001, 7; J. HIRSCHBERG, C.D. MANNING, *Advances in natural language processing*, in *Science*, 349/2015, 261.
[128] Z. WANG, et al., *History, development, and principles of large language models: an introductory survey*, in *AI and Ethics*, 5/2025, 1957; M.A.K. RAIAAN et al., *A review on large language models: architectures, applications, taxonomies, open issues and challenges*, in *IEEE Access*, 12/2024, 26843.

Downloaded from https://teseo.unitn.it/biolaw
ISSN 2284-4503

model, which predicts the next word in a sequence based on the preceding *n-1* words.[129] By analysing massive text corpora, n-gram models calculated the frequency of word sequences to estimate these probabilities.[130] However, SLMs faced significant limitations, for instance, they still had limitations in predicting the semantic relationship between concepts and the context of the language with long-range dependencies[131]. This led to data sparsity issues, where many valid sequences would not appear in the training data, requiring smoothing techniques to compensate.[132] More fundamentally, their reliance on a fixed, short context window meant they could not capture long-range dependencies or the nuanced semantic relationships inherent in human language.[133]

The transition to Neural Language Models (NLMs) marked a pivotal shift: instead of relying on word frequencies, NLMs utilised neural networks to learn language patterns.[134] A key breakthrough was the development of distributed word representations, or word embeddings[135]: models like Word2Vec and GloVe learned to represent words as dense, low-dimensional vectors in a continuous space.[136] This approach was revolutionary because it captured semantic relationships, so words with similar meanings were closer together in the vector space, allowing models to grasp concepts like analogy.[137]

Early NLMs often employed Recurrent Neural Networks (RNNs) and their more advanced variant, Long Short-Term Memory (LSTMs), to process sequential data.[138] These models processed text one word at a time, maintaining a hidden state that served as a memory of previous words.[139] While an improvement over n-grams, RNNs and LSTMs were inherently sequential, which prevented parallelisation during training and made them inefficient.[140] They also struggled with the vanishing gradient problem, limiting their ability to capture long-range dependencies.[141]

LLMs are essentially trained massive neural networks.[142] They use deep neural networks to grasp the statistical patterns and semantic nuances of language, improving the understanding and generation of

[129] M.A.K. Raiaan et al., *op. cit*., 26843; R.K. Singh, D. Rana, *Advancements in natural language processing: an in-depth review of language transformer models, in International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 2024, 1721.

[130] Y. Annepaka, P. Pakray, *Large language models: a survey of their development, capabilities, and applications, in Knowledge and Information Systems*, 67/2025, 2976.

[131] Y. Annepaka, P. Pakray, *op. cit*., 2976.

[132] Y. Annepaka, P. Pakray, *op. cit*., 2976.

[133] Y. Annepaka, P. Pakray, *op. cit*., 2976.

[134] Z. Wang, et al., *op. cit*., 1959; M.A.K. Raiaan et al., *op. cit*., 26843.

[135] See M. Mars, *From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough*, in *applied sciences*, 12/2022, 2.

[136] M.A.K. Raiaan et al., *op. cit*., 26843; Z. Wang, et al., *op. cit*., 1959.

[137] Z. Wang, et al., *op. cit*., 1959.

[138] B.N. Patro, V.S. Agneeswaran, *Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, Applications, and Challenges*, in *Engineering applications of artificial intelligence*, 159/2025, 1; R.K. Singh, D. Rana, *op. cit*., 1721.

[139] B.N. Patro, V.S. Agneeswaran, *op. cit*., 1.

[140] B.N. Patro, V.S. Agneeswaran, *op. cit*., 1.

[141] A.R. Sajun, I. Zualkernan, D. Sankalpa, *A historical survey of advances in transformer architectures*, in *applied sciences*, 14/2024, 2; B.N. Patro, V.S. Agneeswaran, *op. cit*., 1.

[142] P. Kaur, et al., *From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility*, in *arXiv*, 1.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

language.[143] The term "large" highlights the vast scale of parameters, often billions, contributing to the model's capacity to comprehend and generate human-like text.[144] During the training process, through backpropagation and gradient descent, the model adjusts the parameters seeking to minimise the difference between the model's predictions and the actual target sequences.[145]

Most LLMS today are based on transformers. The landscape of NLP was fundamentally reshaped in 2017 with the introduction of transformers,[146] which departed from traditional sequence approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which processed sequences step-by-step or via sliding windows,[147]

The systems discussed in section 1 are based on LLM, which are often built on transformers.[148] Many important foundational models are LLMs, and they are available in open-source;[149] this allowed other developers and companies to base their systems on such models. Nowadays, the healthcare sector has been flooded by general-purpose and healthcare-specific LLMs. For this reason, the following section will briefly describe how transformers work on data to answer the *quaesitum* posed by the MDCG (whether the software performs an action on data beyond storage, archival, communication, simple search, or lossless compression).

4.2. Transformers

Transformers rely on a powerful mechanism known as self-attention, enabling them to model relationships between all parts of a sequence simultaneously.[150]

The original transformers, followed by many subsequent models, adhere to an encoder-decoder structure.[151] The encoder's primary function is to transform an input sequence of symbolic representations, such as tokenised words, into a sequence of continuous representations, essentially interpreting the input's meaning.[152] On the other hand, the decoder takes the representations generated by the encoder

[143] Z. WANG, et al., *op. cit*., 1961.
[144] P. KAUR, et al., *op. cit*., 3.
[145] P. KAUR, et al., *op. cit*., 3.
[146] Introduced in the paper A. VASWANI, et al., *Attention is all you need*, in *arXiv*, 2017.
[147] M.A.K. RAIAAN et al., *op. cit*., 26844.
[148] S. LATIF, et al., *Transformers in speech processing: a survey*, in *Computer science review*, 58/2025, 1; S. SINGH, M. SINGH, V. KADYAN, *Speech recognition transformers: topological-lingualism perspective*, in *arXiv*, 2024, 2; A. LOUBSER, P. DE VILLIERS, A. DE FREITAS, *End-to-end automated speech recognition using a character based small scale transforer architecture*, in *Expert systems with applications*, 252/2024, 2; D. AL-FRAIHAT, et al., *Speech Recognition Utilizing Deep Learning: A Systematic Review of the Latest Developments*, in *Human-centric Computing and Information Sciences*, 2024, 12. Concerning the most famous LLMs: "all GPT (Generative Pre-trained Transformer) models, including the most recent GPT-3 model, are built based on the core technology of Transformers", A. KOUBAA, et al., *Exploring chatgpt capabilities and limitations: a survey*, in *IEEE Access*, 11/2023, 118702; "BERT's model architecture is a multilayer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. (2017)", J. DEVLIN, et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in *arXiv*, 2019, 3.
[149] Y. ANNEPAKA, P. PAKRAY, *op. cit*., 2970.
[150] A. VASWANI, et al., *op. cit*., 3.
[151] A. VASWANI, et al., *op. cit*.
[152] M. GIUNTI, *Tutto quello che avreste voluto sapere su ChatGPT ma non avete mai osato chiedere*. *Note sui Transformer decoder-only*, in *ResearchGate*, 2025**,** 4. DOI: 10.13140/RG.2.2.31444.62084/3 (last access: September 30, 2022).

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

and produces an output sequence of symbols, one element at a time.[153] This generation process is auto-regressive, meaning each new token is generated by incorporating the previously generated symbols as additional input.[154] The decoder's role is to selectively focus on the most relevant information from the input to produce the desired output.[155]

The most pronounced distinction lies in the architectural mechanism for processing sequential data. Previous deep learning models, particularly RNNs and LSTMs, relied on recurrence to handle sequence modelling.[156] RNNs processed sentences sequentially, often struggling with vanishing gradients and failing to capture relationships among words beyond a certain length due to inherent limitations in short-term memory.[157]

The transformer architecture, introduced by Vaswani et al.,[158] discarded recurrence and convolution entirely,[159] Its core component, the self-attention mechanism (or multi-head attention), enabled parallelization and efficient handling of long-range dependencies.[160] Self-attention enables the model to weigh the importance of different words relative to each other and capture dependencies across all positions simultaneously, irrespective of their distance in the sequence.[161] Crucially, this mechanism overcomes the computational inefficiency of sequential methods, as it enables massive parallel processing of tokens, drastically improving training times, especially when dealing with large datasets.[162]

Transformers excel where previous LMs struggled: handling long-range dependencies. In contrast, the attention mechanism inherently discounts the distance between tokens, allowing Transformer-based models to capture dependencies over long spans effectively.[163]

Furthermore, early neural models, such as the initial Generative Pre-trained Transformer (GPT-1), utilized a left-to-right (unidirectional/autoregressive) approach, meaning each token in the self-attention layer could only attend to previous tokens.[164] However, the versatility of the transformer allowed for the rapid development of models like BERT (Bidirectional Encoder Representations from Transformers), which employed an encoder-only architecture trained using a masked language model objective.[165] This design enabled a deep bidirectional understanding where the representation is jointly conditioned on both the left and right context in all layers, a feature typically absent in pre-Transformer models and early unidirectional transformer variants.[166]

---

[153] M. GIUNTI, *op. cit*., 4.
[154] M. GIUNTI, *op. cit*., 5.
[155] J. JONS, *Text summarization using a transformer architecture*, in *Diva*, 2024, 18. Available at https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1896208&dswid=-4519 (last access: September 30, 2025).
[156] B. HELEEN M. ABUBAKAR, S. DODDA, *Exploring transformer models in natural language understanding*, in *Ssrn*, 2025, 2. Available at https://ssrn.com/abstract=5198588 (last access: September 30, 2025).
[157] M.A.K. RAIAAN et al., *op. cit*., 26844.
[158] A. VASWANI, et al., *op. cit*.
[159] A.R. SAJUN, I. ZUALKERNAN, D. SANKALPA, *op. cit*., 2.
[160] M.A.K. RAIAAN et al., *op. cit*., 26840.
[161] A. VASWANI, et al., *op. cit*., 3.
[162] T. WU, Y. WANG, N. QUACH, *Advancements in natural language processing: exploring transformer-based architectures for text understanding*, in *arXiv*, 2022, 2.
[163] A.R. SAJUN, I. ZUALKERNAN, D. SANKALPA, *op. cit*., 2.
[164] M.A.K. RAIAAN et al., *op. cit*., 26847.
[165] J. DEVLIN, et al., *op. cit*.
[166] R.K. SINGH, D. RANA, *op. cit*., 1723.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

The scalable nature of the transformer architecture fundamentally supported the massive increase in model size, leading directly to the classification of modern systems as Large Language Models (LLMs).[167] The new generation, starting with models like GPT-3, showcased a significant discrepancy in scale, moving from parameters in the range of millions (e.g., the largest GPT-2 model with 1.5 billion parameters trained on 40 GB of text data) to hundreds of billions (e.g., GPT-3 with 175 billion parameters trained on 570 GB of text data).[168]

This immense scaling, coupled with training on vast, diverse corpora, resulted in the emergence of advanced capabilities not seen in smaller, previous models, such as sophisticated few-shot or zero-shot learning and the ability to leverage contextual information during inference (e.g., GPT-3 possessing this ability while GPT-2 lacked it).[169] In summary, the transformer models surpassed their predecessors by introducing the efficiency of parallelisation, superior handling of long-range dependencies, and the architectural robustness necessary for the unprecedented scaling that defines modern LLMs.

Despite this being a mere simplification of AI speech recognition functioning, it still makes clear that these systems also go beyond storage, archival, communication, simple search, or lossless compression. The phases in which data are modified or transformed are several: the transformation of sound vibration into visual data (spectrogram or mel-spectrogram); techniques to enhance the sound's quality (e.g. noise reduction); segmentation into short temporal windows; the translation of the phonetic units in numeric vectors; finally, the generation of the output text based on guessing the most probable next token.

More generally, from the way LLMs impacted society, scholars pointed to their transformative power.[170] One of the reasons is the way they work on data: for instance, in the copyright domain, the fair use doctrine highlighted the way protected artworks are transformed through LLMs,[171] the same was held to compare LLMs to techniques aimed at data transformation.[172]

Regardless of the name given to the type of action LLMs perform on data, it is possible to claim that these models go well beyond storage, archival, communication, simple search, or lossless compression.

These models don't *understand* text in a human sense; instead, they function by statistics,[173] predicting the most likely next token in a sequence.[174] This generative process is described as inherently stochastic, meaning it involves randomness.[175] Because LLMs operate on a statistical basis, their outputs will always

[167] Z. Wang, et al., *op. cit.*, 1957.
[168] Z. Wang, et al., *op. cit.*, 1958; T. Wu, Y. Wang, N. Quach, *op. cit.*, 2.
[169] M.A.K. Raiaan et al., *op. cit.*, 26845.
[170] A. Srivastava, et al., *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*, in *arXiv*, 2023.
[171] R. Bommasani, et al., *On the Opportunities and Risks of Foundation Models, Center for Research on Foundation Models (CRFM)*, in *arXiv*, 2022, 146.
[172] S. Ghazzai, et al., *Harnessing GPT for Data Transformation tasks*, in IEEE *International Conference on Web Services (ICWS)*, 2024.
[173] For the statistics behind LLMs see W. Ji, et al., *An Overview of Large Language Models for Statisticians*, in *arXiv*, 2025.
[174] M. Giunti, *op. cit.*, 6, 23; W. Ansar, S. Goswami, A. Chakrabarti, *A survey on transformers in NLP with focus on efficiency*, in *arXiv*, 2024, 4.
[175] W. Su, *Do large language models (really) need statistical foundations?,* in *arXiv*, 2025, 3.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

involve variability and uncertainty.[176] This explains why even the most powerful models still make mistakes.[177]

Of particular interest is the English NHS Guidance on the use of AI-enabled ambient scribing products in health and care settings,[178] according to which speech recognition systems are not medical devices due to their simple functionality, but they become medical devices if generative AI is used to summarise the transcript.[179] The English guide distinguishes between AI systems with low and high functionality; however, it does not clarify how to establish such a scale.[180] The US Federal Drug Administration (FDA), instead, still did not take a position on the matter,[181] however, according to US scholars, these systems could not be considered regulated medical devices under the US Federal Food, Drug, and Cosmetic Act.[182]

5. Step 4: Individual Benefit

The next step is to ascertain whether the system benefits individuals. Software that only aggregates population data, provides generic diagnostic or treatment pathways not directed to individual patients, or is used for scientific or epidemiological purposes, is not considered to benefit individual patients and is therefore excluded.

The guide uses the term "benefit" to refer to the clinical benefit as defined in the art. 2(53) MDR as "the positive impact of a device on the health of an individual, expressed in terms of a meaningful, measurable, patient-relevant clinical outcome(s), including outcome(s) related to diagnosis, or a positive impact on patient management or public health".

In the MDR framework, the beneficial effect of a medical device must be demonstrated through clinical evaluation and overcome the risks associated with it.[183] As clarified by the previous 2016 MEDDEV guide, the benefit must concern the individuals' health. With this, the framework refers to positive impacts on clinical outcomes (such as reduced probability of adverse outcomes, e.g. mortality, morbidity; or improvement of impaired body function); patients' quality of life (like simplifying care or improving the clinical management of patients); outcomes related to diagnoses; public health impact, etc.[184]

For the purposes of this research, it is essential to note that the clinical benefit also encompasses the enhanced clinical management *of* the patient. The MDR, the MEDDEV guide and the MDCG guide seem

[176] W. SU, *op. cit*., 3.

[177] I am aware that these models do not make mistakes, rather, they simply do not respect our expectations.

[178] NHS England, *Guidance on the use of AI-enabled ambient scribing products in health and care settings*, 2025, 2.

[179] NHS England, *Guidance on the use of AI-enabled ambient scribing products in health and care settings*, 2005, 7.

[180] The guide poses questions related to functionality; however, it does not explain how to answer.

[181] "The FDA and other regulatory organizations have started to outline guidelines for AI in healthcare equipment, but they have not yet addressed documentation and administrative AI use in depth", S. AGARWAL, S.B. PETA, *From notes to billing: large language models in revolutionizing medical documentation and healthcare administration*, in *Scholars journal of applied medical sciences*, 13(8)/2025, 1561.

[182] See S. GERKE, D.A. SIMON, B.R. ROMAN, *Liability risks of ambient clinical workflows with artificial intelligence for clinicians, hospitals, and manufacturers*, in *JCO Oncology Practice*, 0/2025; see also C. CHEN, J.E. THORNTON, *AI-generated clinical summaries*, in *JAMA*, 331/2024.

[183] M. BRETTHAUER, et al., *The new European medical device regulation: balancing innovation and patient safety*, in *Medicine and public issues*, 2023, 845.

[184] MEDDEV 2.7/1, revision 4, Clinical evaluation: a guide for manufacturers and notified bodies under directives 93/42/EEC and 90/385/EEC, 2016, 43.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

to refer to patients' benefit; therefore, such benefit should be intended as patients being more able to manage themselves. On the other hand, it could also mean that software improving patients' management by others, such as physicians, should be regarded as a medical device (provided the other requirements of the guide are met). However, it would make poor sense to classify as medical devices only the software used by patients for their management, considering also that medical doctors use the vast majority of medical devices on patients. For these reasons, software used to improve physicians' management of patients should be considered as providing a clinical benefit for the individual patient in the meaning of art. 2(53) MDR.

For these reasons, AI systems devoted to speech recognition, summarisation, and other kinds of clinical documentation do pose clinical benefits for individual patients, given that such damage can also positively impact patients' management.

To conclude, it must be remembered that the MDR acknowledges the status of medical devices to products without clinical benefits.[185] Annex XVI serves the purpose to cover products without a medical purpose (and therefore clinical benefit) deemed to be similar to medical devices in terms of functioning and risk profile, as considered by recital 12 MDR.[186] The systems analysed in this paper are not listed in annex XVI; however, art. 1(5) MDR states that the Commission can amend this list, adding new groups of products.[187] Thus, the Commission should consider inserting these LLMs designed for healthcare settings in annex XVI to provide further clarity.

6. Step 5: Medical Purposes between Diagnosis and Anamnesis

The last step to classify a software as a medical device requires to establish whether the software matches the definition of medical device software (MDSW) provided in the guidance.[188] It defines MDSW as a "software that is intended to be used, alone or in combination, for a purpose as specified in the definition of a "medical device" in MDR or *in vitro* diagnostic medical devices regulation".[189] Such definition is in reality a non-definition, because it merely recalls the well-known definition provided by the MDR in art. 2 MDR, which requires devices to meet specific purposes to be classified as medical devices. The provision enlists different medical purposes, such as diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease; or providing information by means of *in vitro* examination of specimens derived from the human body, including organ, blood and tissue donations, etc.

---

[185] B. WILKINSON, R. VAN BOXTEL, *The medical device regulation of the european union intensifies focus on clinical benefits of devices,* in *Therapeutic Innovation & Regulatory Science*, 54(3)/2020, 614.

[186] Mdcg 2023-5, *Guidance on qualification and classification of Annex XVI products*, 2023, 3, 5.

[187] "Where justified on account of the similarity between a device with an intended medical purpose placed on the market and a product without an intended medical purpose in respect of their characteristics and risks, the Commission is empowered to adopt delegated acts in accordance with Article 115 to amend the list in Annex XVI, by adding new groups of products, in order to protect the health and safety of users or other persons or other aspects of public health".

[188] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 12.

[189] Mdcg 2019-11 Rev.1, *Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*, 2025, 7.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

Thus, do LLMs devoted to medical documentation serve medical purposes? Is it the summarisation of medical documents used for diagnosis or other purposes indicated in art. 2 MDR?

The MDR does not define concepts such as diagnosis or disability, though they are fundamental for classification purposes. Such concepts were, however, defined elsewhere.

According to the Encyclopaedia Britannica, diagnosis is "the process of determining the nature of a disease or disorder and distinguishing it from other possible conditions. The term comes from the Greek *gnosis*, meaning knowledge".[190] Similarly, the US National Cancer Institute defines it as "the process of identifying a disease, condition, or injury from its signs and symptoms. A health history, physical exam, and tests, such as blood tests, imaging tests, and biopsies, may be used to help make a diagnosis".[191] Both definitions underline that a diagnosis is a process. Other definitions, however, narrow the concept to the judgment given by the physician. For instance, the Cambridge dictionary defines diagnosis as "a judgment about what a particular illness or problem is, made after examining it".[192] The ambivalence of the concept was noted by scholars, who highlighted that a diagnosis is both the pre-defined set of categories to designate a specific pathological condition, and the whole process by which such categories are applied.[193] A simplification of the diagnostic process is described as follows: firstly, the patient is likely the first person to consider her symptoms and to engage with the healthcare system accordingly;[194] once the healthcare apparatus accepts the patient, an iterative process of information gathering follows. It entails information integration and interpretation. Defining the clinical history requires interviews, physical exams, diagnostic tests, and referring or consulting with other clinicians. The information-gathering techniques are continuous and can be employed at different phases: they involve continuous hypothesis generation and updating probabilities as more information is analysed.[195] In this line, the diagnostic process is described as "a complex, patient centered, collaborative activity that involves information gathering and clinical reasoning with the goal of determining a patient's health problem".[196] Accordingly, the diagnostic process comprehends four different activities aimed to information-gathering: the interview, physical exams, diagnostic testing, and referrals or consultations.[197] Thus, the clinical reasoning stage builds on different previous phases, where a more or less broad spectrum of potential diagnoses is narrowed

[190] Available at: https://www.britannica.com/science/diagnosis (last access: September 8, 2025).

[191] Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/diagnosis (last access: September 8, 2025).

[192] Available at: https://dictionary.cambridge.org/dictionary/english/diagnosis (last access: September 8, 2025).

[193] A. JUTEL, *Sociology of diagnosis: a preliminary review, in Sociology of Health & Illness*, 31(2)/2009, 278.

[194] E.P. BALOGH, B.T. MILLER, J.R. BALL, *Improving diagnosis in health care*, in *National academy press*, 2015. See also H. LLEWELYN, et al., *Oxford Handbook of Clinical Diagnosis*, 2014, 25; P. GØTZSCHE, P. GØTZSCHE, *Rational Diagnosis and Treatment: Evidence-Based Clinical Decision-Making*, 2008, 2.

[195] "Throughout the diagnostic process, there is an ongoing assessment of whether sufficient information has been collected. If the diagnostic team members are not satisfied that the necessary information has been collected to explain the patient's health problem or that the information available is not consistent with a diagnosis, then the process of information gathering, information integration and interpretation, and developing a working diagnosis continues", E.P. BALOGH, B.T. MILLER, J.R. BALL, *op. cit.*, 35-36.

[196] P. BALOGH, B.T. MILLER, J.R. BALL, *op. cit.*, 32.

[197] P. BALOGH, B.T. MILLER, J.R. BALL, *op. cit.*, 36. See also J.A. DE MORAIS RODRIGUES, et al., *The importance of anamnesis in the excellent clinical examination*, in *International Journal of Science and Research Archive*, 13(02), 2024, 3314.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

down into fewer likely options.[198] When, according to the physician, the remaining potential diagnoses mirror patients' symptoms, the more appropriate option is confirmed and conveyed to the patient as the final diagnosis.[199]

In conclusion, diagnosis can be the final judgement of the medical doctor, or the whole process that grounds it. Thus, also the exchange of information between physicians and patients is an integral part of the diagnostic process, as well as the final answer communicated to the patient. The very first part, where information is collected from the patient, is also called anamnesis.[200] This phase, as constituting the basis of the final diagnosis conveyed to the patients, can influence and determine the quality of the clinical reasoning.[201]

For this research purpose, what is relevant is: does art. 2 MDR refer to all the phases of the diagnostic process (anamnesis included), or only to the final clinical reasoning resulting in the answer submitted to the patient? The MDR does not answer directly. However, it must be remembered that it is a regulation on medical devices, not on clinicians' medical reasoning. Traditional medical devices do not operate the final diagnosis; they produce information used in the different stages of the diagnostic process.

Diagnoses can be performed by both patients and physicians using medical devices, which are usually prescribed by physicians for diagnostic purposes (including system type, usage instructions, and duration). When these devices are prescribed, the objective is to gather information to base the final reasoning on all the possible relevant data.

For example, the MDR regulates devices operating magnetic resonance imaging. These systems are intended to provide relevant information (in the form of images) on which to base the final diagnosis conveyed to the patient. Therefore, these devices are designed and commercialised to inform different stages of the diagnostic process.

Traditionally, the final diagnosis was not provided by medical devices; instead, it was determined only by physicians. To be more precise, this was the rule before the advent of data-driven technologies: today, medical devices can also provide a final diagnosis to physicians and to patients when powered by automation or AI techniques.[202] Therefore, when considering the medical devices regulated under the MDR,

---

[198] P. BALOGH, B.T. MILLER, J.R. BALL, *op. cit*., 34; see also J.P. KASSIRER, *Teaching Clinical Reasoning: Case-Based and Coached*, in *Academic medicine*, 85(7)/2010.

[199] P. BALOGH, B.T. MILLER, J.R. BALL, *op. cit*., 35-36.

[200] "An anamnesis or "taking a history" is a practice where a doctor inquires, not only into a patient's immediate complaints and symptoms, but also into their past and the history of their symptoms. This history may reveal particular clusters of symptoms or a recognizable development over time that again point to probable diagnoses and further investigations", K. TYBJERG, *Medical anamnesis. Collecting and recollecting the past in medicine*, in *Centaurus. Journal of the European Society for the History of Science*, 65(2)/2023, 236; see also J.A. DE MORAIS RODRIGUES, et al., *op. cit*., 3314; Z. JOKSIMOVIC, D. BASTAĆ, *Anamnesis – the skill and art of clinical medicine*, in *Timok medical gazette*, 47(4)/2022, 153.

[201] "Clearly, the most informative part of this broad diagnostic process will be the history…gold standards for final diagnoses are best based on the outcome of patients' symptoms combined with the result of histology, biochemistry, or some other measurements. So final diagnoses are often based on initial history-taking skills from which outcomes can be assessed", G. LIPSCHIK, et al., *Oxford American handbook of clinical diagnosis*, 2009, 57. See also H.C.B. HILLEN, *The IT anamnesis? Diagnosing software requirements*, 2012, 3; JOHRL, et al., *An evaluation framework for clinical use of large language models in patient interaction tasks*, in *Nature Medicine*, 31/ 2025, 77.

[202] The literature is infinite: ex multis, see J.A. NICHOLS, H.W.H. CHAN, M.A.B. BAKER, *Machine learning: applications of artificial intelligence to imaging and diagnosis*, in *Biophysical reviews*, 11/2019; P. SZOLOVITS, R.S. PATIL, W.B. SCHWARTZ,

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

we can include both those intended to provide information for diagnostic purposes and those aimed at delivering the diagnosis.
Speech recognition systems are deployed for anamnesis purposes;[203] they are beneficial for patient management because they allow physicians to focus on reasoning rather than wasting hours on documentation.
On the other hand, summarisation systems remove irrelevant information, thus providing physicians only with relevant information in order to determine the correct diagnosis. Including the wrong information can significantly impact the final diagnosis,[204] while focusing only on the important one improves the quality of the judgment.[205] Such considerations lead to the conclusion that AI summarisation tools also contribute to the diagnostic process.

7. Results and Conclusions

This section will first describe the results achieved through the analysis and then provide some final observations on the resulting picture.
The results will be analysed presenting the outcomes of each step required by the 2025 MDCG guidance. The final observation will concern the whole picture related to AI for documentation in the healthcare sector.
The steps described in the guide are the following: i) establishing whether the product is a software according to the definition indicated in the guide; ii) ascertaining whether the software is listed in MDR annex XVI device, is an accessory to a medical device, or a software that drives or influences the use of a medical device. If so, it is a medical device; iii) If not, is to be determined whether the software performs an action on data beyond storage, archival, communication, simple search, or lossless compression (i.e. using a compression procedure that allows the exact reconstruction of the original data); iv) If it does, the last step is to ascertain whether this action is intended for the benefit of individual patients; v) lastly, the software has to match the definition of medical device software provided in the guidance. Such a definition simply refers to art. 2 MDR, which requires products to serve the medical purposes listed therein.

7.1. Results

The previous sections described the steps required to establish whether software should be classified as a medical device.
No attention was paid to the first two steps, which required ascertaining whether the products at issue can be considered software or an accessory to a medical device and whether they are enlisted under

Artificial intelligence in medical diagnosis, in Annals of internal medicine, 108(1)/1988; N.G. NIA, E. KAPLANOGLU, A. NASAB, *Evaluation of artificial intelligence techniques in disease diagnosis and prediction*, in *Discover artificial intelligence*, 3(5)/2023.

[203] "Anamnesis, and electronic health records can be considered as critical data sources. In this context, Machine Learning (ML) and Deep Learning (DL) approaches have shown excellent capabilities in working with anamnesis and unstructured text data", S. LEMBO, et al., *AI4RDD: An Artificial Intelligence and Rare Disease Diagnosis Approach For Anamnesis Process, preprint*, 2025, 12.

[204] J.A. DE MORAIS RODRIGUES, et al., *op. cit.*, 3315.

[205] S. JOHRL, et al., *op. cit.*, 77.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503

annex VXI MDR (considering medical devices without medical purposes). The AI systems under analysis are indeed software, and this research concerned their classification only when considered by themselves (and not in combination with other medical devices); lastly, none of the systems focused on this paper is covered in annex XVI MDR.

The third step was deeply analysed. It requires to establish whether the software performs an action on data beyond storage, archival, communication, simple search, or lossless compression. The answer was preceded in the analysis by displaying how LLMs work, especially when based on transformer architecture. Such a description shows the transformative process conducted on data, explaining the reasons why multiple hallucinations and fabrications are experienced with the use of these systems, as outlined in section 1. It was therefore straightforward to conclude positively to the answer posed by the MDCG: yes, LLMs do perform actions on data that go beyond storage, archival, communication, simple search, or lossless compression.

On this basis, the paper proceeded to the fourth step, asking to deliberate whether the software's action is intended for the benefit of individual patients.

To answer, the analysis started from the concept of benefit, which refers to the clinical benefit defined in art. 2(53) MDR, according to which, a clinical benefit for the individual's health is also that improving patients' management. The systems under analysis, as speech recognition and summarisation AI tools, are sold advertising their positive impact on patients' management. In the case of summarisation tools, it should be noted that they are meant to remove the unnecessary information from extensive documentation. Such task is not only related to patients' management, but it is also relevant for patients' health; indeed, it brings the system's user to focus only on specific medical data, improving or decreasing the quality of care. On these bases, this paper suggests considering that such software's action is intended for the benefit of individual patients.

Proceeding to the last step, the guidance requires the software to conform to the definition of medical device software provided therein. Such a definition simply refers to art. 2 MDR, which requires products to serve the medical purposes listed therein.

The analysis shows that the concept of "diagnosis" is twofold: it relates to both the diagnosis as the entire process leading to the final decision and its final step, the diagnosis, when the decision is conveyed to the patient. The paper underlined that the MDR regulates systems, such as RX scans, that provide information to the medical doctor, which will serve as an informational ground for the physician to establish the presence or absence of a given pathology. Therefore, the traditional medical devices do not provide the final diagnosis, but they are nevertheless considered falling under the MDR scope because they participate in the diagnostic process. On this basis, the analysis concluded that the reference to "diagnosis" in art. 2 MDR concerns the entire diagnostic process. Such process starts from the anamnesis and ends with conveying the final decision to the patient. AI documentation systems participate in the process: for example, speech recognition systems help in the anamnesis phase, tracking everything said between doctors and patients; summarisation tools, instead, assist physicians in a subsequent phase, where the information is collected, and must be deprived of irrelevant factors. All this, as said, for the benefit of the individual patient.

Downloaded from https://teseo.unitn.it/biolaw
ISSN 2284-4503

For these reasons, this paper suggests classifying these systems as medical devices under the MDR. However, such analysis must be conducted on a case-by-case basis to determine whether each step of the MDCG guide is followed.

7.2. Conclusions

This matter, namely the classification of AI systems intended for medical documentation as medical devices, is significant.

Following the approval of the European Health Data Space regulation, electronic health records will spread across the Union, increasing documentation requirements and the time needed to complete them.[206] Such an effect has already been analysed in the USA, where documentation time and burnout have been widely observed.

The introduction of AI systems intended for medical documentation could, as described, alleviate physicians' burdens and burnout risks. However, the regulation of these systems is still in its infancy, and this paper aims to contribute to it.

The classification of these systems as medical devices would make a huge difference. Indeed, establishing that software is a medical device means including it under the umbrella of both the MDR and the AI Act, thereby increasing compliance duties, such as clinical evaluations and documentation.[207] Another relevant consequence would concern the principle of human oversight by design as established in art. 14 AIA. SR and summarisation systems based on AI will need to be provided in a way that allows users to effectively supervise the system. For instance, SR systems could be built so they can click on any word and listen to the specific segment of the patient's visit to check accuracy.[208]

The MDR has already been criticised for the large number of duties imposed on manufacturers. According to many scholars, this led medical companies to leave the European market, favouring the USA market.[209] Therefore, the output of this research (if followed by competent authorities) could severely impair the growth of this new market segment. However, the MDR is specifically designed to guarantee patients' safety, and classifying these systems as medical devices would both increase patients' trust and safety, provided that manufacturers furnish extensive evidence to notified bodies. That said, the MDCG should take a stance on the topic to counter the uncertainty caused by this regulatory gap.

---

[206] D. FÅHRAEUS, J. REICHEL, S. SLOKENBERGA, *The European health data space: challenges and opportunities*, in *Sieps*, 2024, 11.

[207] For a view on the combined duties see F. GENNARI, *O Complementarity, Where Art Thou? Wading through the Medical Device Regulation and the AI Act Compliance: The case of Software as a Medical Device. A Primer*, in *BioLaw*, 3/2024.

[208] This measure is implemented in the SR systems provided by Abridge, as described in https://support.abridge.com/hc/en-us/articles/30235128433811-Verify-a-Note-With-Linked-Evidence.

[209] "Manufacturers have adopted a "Europe last" approach and have started to carry out their clinical trials in the USA first, which is seen as a more manufacturing- and innovation-centred regulatory environment", O. MCDERMOTT, B. KEARNEY, *A review of the literature on the new European Medical Device Regulations requirements for increased clinical evaluation, in International Journal of Pharmaceutical and Healthcare Marketing*, 2023, 17.

Downloaded from https://teseo.unitn.it/biolaw ISSN 2284-4503