AutoGenHR: Automated Generation of Health Reports for Patients at Home*

Nicole Merkle[1,*,†]

[1]Karlsruhe Institute of Technology (KIT), Institute for Automation and Applied Informatics (IAI), Germany

Abstract

European countries have the highest proportion of people over the age of 50. This implies that age-related diseases, such as cardiovascular diseases, diabetes, cholesterol and dementia, are becoming increasingly common in our aging societies. Additionally, it can be observed that health systems are overburdened due to a lack of medical specialists, resources and capacities. To get an appointment with a medical specialist, patients have to wait up to several months. These circumstances imply that necessary health checks and detection of health risks cannot be carried out in a sufficient manner. Furthermore, the personalisation of medical treatments can hardly be guaranteed. To tackle these problems and enable preventive measures and interventions, a round-the-clock health monitoring, prediction of health risks, and personalised treatment would be necessary. This paper presents *Auto Generative Health Reports (AutoGenHR)*, a work in progress that allows the automated generation of health reports based on acquired context data from round-the-clock monitoring of patients in their domestic environment. In the proposed approach health reports are generated by means of dynamically created knowledge graphs representing patients and context information, and a language model fine-tuned for generating health reports. In this way an early detection of health risks and prevention of avoidable deaths is to be achieved, while patients can stay in their domestic environments without overstretching the capacities of the health system.

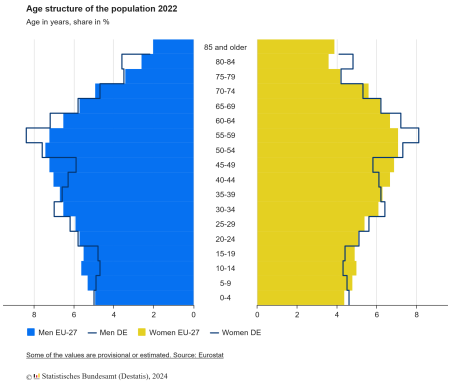
Keywords

Virtual Agents, Knowledge Graphs, Large Language Models, Transformer Neural Networks, Wearables, Digital Twins

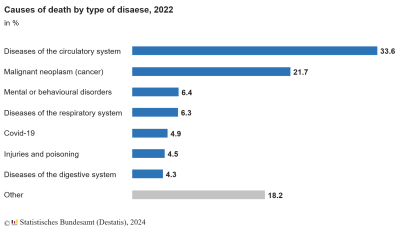
1. Introduction

On average, the population of European countries shows an ageing society, as the highest population density lies in the age range between 45 and 70 years [1] (see Fig. 1a). This leads to implications affecting the health systems, e.g. the increase of age-related diseases. Age cohorts from 45 up to 85 cause the highest costs for treating age-related illnesses [1] (see Fig. 1d). Another implication is that the healthcare systems of European countries are overburdened and inadequate due to lacking capacities, resources and health professionals. Thus, the quality of medical services is significantly decreased and even chronically ill patients do not receive the medical care and treatment they actually need. The statistics in Fig. 1b and Fig. 1c show that diseases of the *circulatory system* and *cancer* followed by *mental health* issues are the most common causes of high costs and deaths [1]. The approach presented aims at addressing the

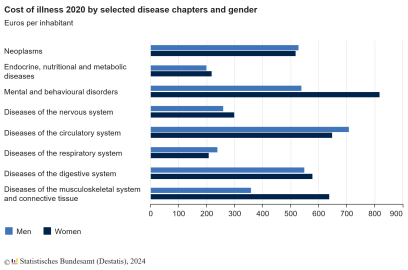
Semantics'24: 20th International Conference on Semantic Systems, September 17–19, 2024, Amsterdam, NL
*Nicole Merkle.
nicole.merkle@kit.edu (N. Merkle)
https://www.iai.kit.edu/2154_4249.php (N. Merkle)
0000-0001-6063-1689 (N. Merkle)



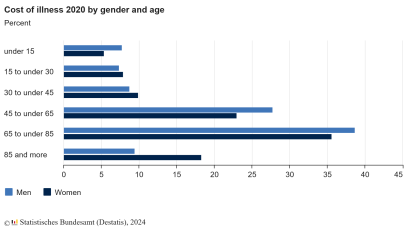
(a) Population distribution in European countries [1].



(b) Diseases as causes of death [1].



(c) Costs caused by diseases [1].



(d) Costs by age cohorts [1].

Figure 1: Costs caused for the health system distinguished by disease type and age for men and women [1].

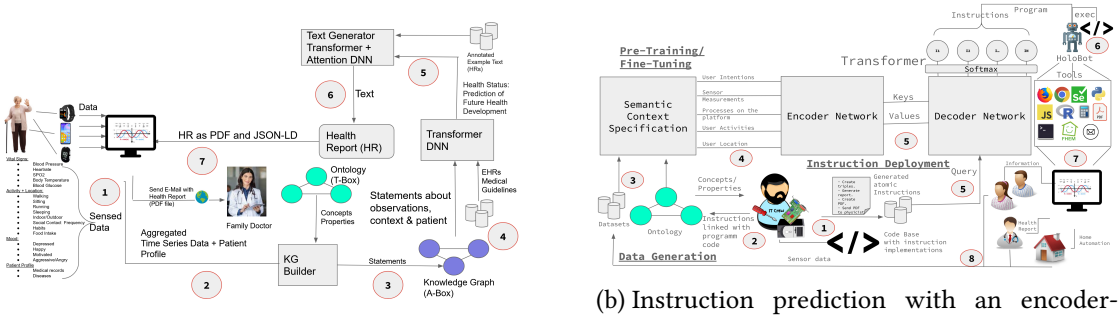
aforementioned problems through a dedicated AI platform that implements digital twins of patients represented by knowledge graphs (KGs) and enables the monitoring of them in their home environment. The purpose is to recognise health risks of the circulatory system and mental health issues at an early stage and inform the relevant contact points, e.g. emergency services, to intervene in risk situations. Furthermore, patients do not have to keep a health diary as the AI platform takes care of this. In the long term, the aim is to allow precision medicine and thus an improved treatment of patients while the costs caused by frequent diseases and usage of health resources is to be reduced.

2. Related Works

Some recent papers have proposed methods that share similarities with the proposed *AutoGenHR* platform. Wang et al. introduced *TWIN-GPT*, a LLM-based approach for *"creating personalised digital twins in clinical trials"* [2]. In contrast to AutoGenHR, TWIN-GPT focuses primarily on optimising clinical trials and not on continuous health monitoring and reporting at home. Fatemeh et al. propose a framework for personalised diabetes treatment by leveraging digital health twins and knowledge graphs [3]. However, their approach is limited to diabetes treatment

while AutoGenHR aims at addressing a wide range of diseases and treatments. Zhang et al. propose a *"digital twin framework for type 2 diabetes"* [4] that utilises KGs in order to describe multiomics data and interpret prediction results [4]. However, the authors state that a future improvement would be to provide a *"natural language interface employing large language models"* [4]. AutoGenHR attempts to close this gap by training and fine-tuning a medical language and health progression model. Gyrard et al. implement a *"Mental Health Knowledge Graph"* [5] to support digital twins with respect to the standardisation of *"communication protocols, data exchange and integration"* [5]. However, unlike AutoGenHR, they are limited to mental health aspects only. In future work, AutoGenHR intends to address some limitations of the considered related work by combining round-the-clock monitoring, dynamic knowledge graph generation, and LLM-driven reporting. In addition, the automated generation of health reports will address the need for shared communication and understanding between patients, physicians, developers and healthcare providers.

3. Approach



(a) Frequent generation and storage of HRs.

(b) Instruction prediction with an encoder-decoder transformer deep neural network.

Figure 2: AutoGenHR for automated generation of HRs and execution of context-related code instructions.

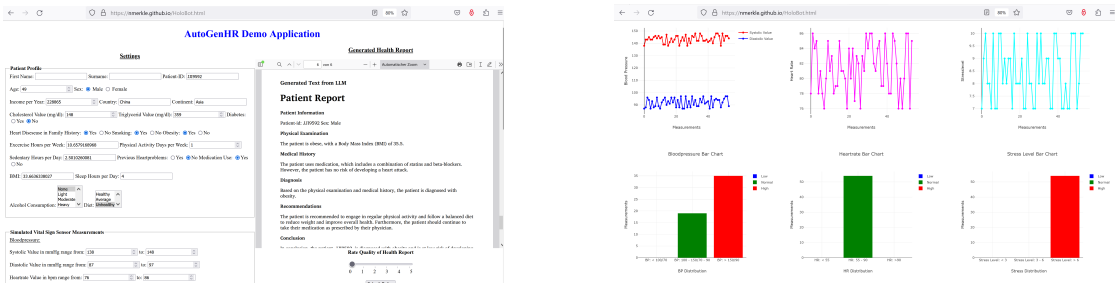
The AI platform proposed consists of different software modules, e.g. *KG Builder, 3 encoder-decoder transformer neural networks* for a) health progression prediction, b) text and c) instruction generation. A devised ontology serves as basis for the generation of dynamic KGs, that serve as patient and context representation and thus input to the *encoder-decoder transformers* [6]. Fig. 2a illustrates the process of health report (HR) generation. 1) Heterogeneous data is collected either in different time intervals or event driven. For this purpose, wearable edge devices, e.g. *earables* [7], smartwatches, are to be used for measuring different vital sign parameters and patient activities. 2) The collected data and the patient profile are leveraged in order to generate a corresponding KG representation by means of the *KG Builder*. 3) The KG Builder relies on the proposed ontology. 4) The generated KG serves together with medical datasets as input for the first *Transformer DNN* that predicts the patient's future health progression for a predefined time window. For this purpose, the KG elements are encoded as embedding vectors, while the proposed ontology represents the vocabulary for the encoder. 5) The output of the transformer

serves together with annotated medical data, as input for the *Text Generator* neural network, since medical knowledge in the language model is to be learned through medical guidelines. 6) The pre-trained language model of the *Text Generator* is obtained by training and evaluating it with the curated *FineWeb-EDU* [8] dataset because the language model should provide initially a general and reliable *world knowledge* adaptable to specific tasks. 7) The fine-tuned *Text Generator* generates the HR which is sent to a web platform that stores the HR in a knowledge base. The treating physician has access to this web platform and can use the HRs and the data provided to gain valuable insights that contribute to the personalised treatment of patients.
The platform also implements a virtual agent (VA) that orchestrates the services provided to the user. To do so the VA assembles instructions and performs them via available software tools (see Fig. 2b). 1) A developer deploys code and 2) creates a semantic specification of the code. 3) Context-related datasets are transformed into a KG that 4) serves as input to an encoder-decoder transformer. 5) The encoder encodes the context KG into *attention keys* and *values*, while the decoder takes the developer instructions and encodes them as *queries* in order to perform *cross-attention* [6] between context-related data and suitable instructions. 6) The VA executes the instructions based on the instructions' semantic representation and via installed tools and programs. 7) During the orchestration process, the VA must check whether required software is installed on the operating system and whether the executable instructions are authorised for execution. It is possible that some instructions are declared as critical or not permitted for the agent due to security reasons. Every change in the context may lead to new instructions that are executed sequentially in order to provide context-aware and personalised services to the patient. 8) Thus, the transformer model is regularly refined using newly acquired data.

The ontology published on Github[1], actually consists of 25 concepts and 54 properties while it will be extended in future. The concepts and properties comprise technical knowledge about a) the platform and its installed or required software tools, b) the executable codes, i.e. instructions and c) the patient profile and measurable vital sign parameters. Software developers and physicists will use the ontology to integrate their knowledge into the platform. While physicists will create patient profiles and provide medical records, developers of healthcare providers will create agent programs or code profiles for integrating their software. Moreover, the ontology has the purpose to personalise services provided by the VA to the patient. E.g., if a person is hard of hearing, text is displayed on an available display; if the person has a visual impairment, the text is read aloud via loudspeakers.

A web demo illustrates what the end result of the approach presented might look like. It should be mentioned that the demo[2] is not yet fully implemented, as the approach is still under development. The health reports received were generated and assigned separately for each data element in advance using the *Lama-3-8B-instruct* [3] model. This means that there is currently no life text generation in the background, but this will be adapted in future. The selected dataset[4] will be replaced by representative high quality datasets from medical studies.

[1] https://github.com/nmerkle/holomatik-ai/blob/main/Holomatik-Ontology.ttl
[2] https://nmerkle.github.io/HoloBot.html
[3] https://huggingface.co/nmerkle/Meta-Llama-3-8B-Instruct-ggml-model-Q4_K_M.gguf
[4] https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset



(a) (b)

Figure 3: In Fig. 3a the web demo creates a HR as a PDF file for randomly selected patient data. Fig. 3b shows how the simulated vital parameters develop over a certain period of time.

4. Conclusion

This paper presented AutoGenHR, a work in progress, that allows the round-the-clock health observation of patients and early prediction of health risks. Dynamically generated KGs represent the digital twin of the patient and interconnect this twin with the patient's context to serve as input to a fine-tuned language model. The main objectives of the approach are the personalisation of treatments and services and the reduction of the burden on the healthcare system by shifting diagnosis to the home environment. The local automated generation of HRs makes it no longer necessary that patients have to maintain themselves a health diary while it ensures the frequent documentation for physicists and thus better therapy opportunities.

Acknowledgements: *This contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health - Helmholtz Information and Data Science School for Health".*

References

[1] Destatis, Statistisches Jahrbuch 2024 für die Bundesrepublik Deutschland, Federal Statistical Office, Wiesbaden, 2024.
[2] Y. Wang, et al., Twin-gpt: Digital twins for clinical trials via large language model, 2024.
[3] F. Sarani Rad, et al., Personalized diabetes management with digital twins: A patient-centric knowledge graph approach, Journal of Personalized Medicine 14 (2024). doi:10.3390/jpm14040359.
[4] Y. Zhang, et al., A framework towards digital twins for type 2 diabetes, Frontiers in Digital Health 6 (2024). doi:10.3389/fdgth.2024.1336050.
[5] A. Gyrard, et al., Iot-based preventive mental health using knowledge graphs and standards for better well-being, 2024. arXiv:2406.13791.
[6] A. Vaswani, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, 2017.
[7] T. Röddiger, et al., Openearable: Open hardware earable sensing platform, Association for Computing Machinery, New York, NY, USA, 2023, p. 246–251.
[8] G. Penedo, et al., The fineweb datasets: Decanting the web for the finest text data at scale, 2024.