Exploration of Declining Blood Pressure Control Through Large Language Models

Jiawen Chen, Hongqian Niu

Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

Hypertension, or high blood pressure (BP), is a major risk factor for cardiovascular disease, the leading cause of death worldwide, and exhibits concerning negative trends in BP control among US adults from 1999-2020 evidenced by the National Health and Nutrition Examination Survey (NHANES) results. While traditional methods such as generalized linear regression, random forest, and gradient boosting often perform well in classification, they are unable to offer contextual explanations of the relationships between cross-domain covariates. For example, understanding the effects of demographic variables, such as age, in conjunction with medication indicators for different hypertension drugs remains a challenge. Large Language Models (LLMs), such as GPT-4, have revolutionized the way large datasets can be processed and understood. The use of LLMs has even seen promising results in zero-shot classification, demonstrating the potential for leveraging the background information encoded from vast training datasets. Here, we use NHANES data to predict BP control and employ supervised LLM fine-tuning to explore key factors behind declining BP control in participants with hypertension[1].

[1]All code for data pre-processing, fine-tuning LLM, traditional method, and LLM prompting are available at `https://github.com/hong-niu/enar-datafest-2024`.

Introduction

Hypertension, commonly known as high blood pressure (BP), is a global public health concern, affecting a significant portion of the adult population. If left uncontrolled, it can lead to severe health complications such as heart disease, stroke, and kidney failure (Gaziano et al., 2006). Although interventions based on the aforementioned factors are often successful in reducing BP levels, recent studies suggest that the prevalence of BP control among adults with hypertension has worsened across varying time periods in the period between 2009 and 2020 (Muntner et al., 2022).

To investigate thiese trends, we analyze data from the National Health and Nutrition Examination Survey (NHANES), a comprehensive cross-domain dataset. Traditional methods to analyze these data include logistic regression, random forests, and XGBoost (Chen and Guestrin, 2016). Despite impressive classification performance, a common limitation to these traditional techniques is the inability to directly model covariate contexts such as names or categorical values, which are often reduced to numeric encodings. However, these names often carry critical domain-specific information, and can provide essential context that enhances the understanding of relationships in the data.

LLMs such as GPT-4, have revolutionized the way we approach complex tasks, showcasing their capabilities in a wide range of applications. These models, trained on vast corpora of text, excel at understanding and generating human language, making them powerful tools for diverse purposes (Roziere et al., 2023). Recognizing the potential benefits of LLMs in data analysis due to their inherent ability to comprehend the data labels themselves and incorporate previously trained contextual information (Hegselmann et al., 2023), we aim to evaluate LLMs for NHANES data reasoning. Our approach involves fine-tuning an LLM for a classification task to determine whether an individual with hypertension has uncontrolled BP based on cross-domain covariates such as age, race, and medication use. Following this, we query the fine-tuned LLM regarding factors that correlate with and contribute to predicting BP control. The final step involves a critical evaluation of the responses provided by the fine-tuned LLM, with a focus on assessing the accuracy of classification task and relevance of the information in relation to BP control.

Method

Data pre-processing. The NHANES data, encompassing survey responses from 59,799 individuals from 1999-2020, offers a rich information covering various health-related aspects. To specifically investigate the potential correlates and causal factors behind the observed worsening in BP control over the years among hypertension patients, we focused on a subset of the NHANES dataset. This subset includes 20,409 individuals who meet the criteria for hypertension as defined by the federal (JNC7) guidelines. This approach facilitates a more targeted examination of the trends, treatment effectiveness, and the impact of various health-related factors on hypertension management over a span of two decades. For detailed descriptive information about this data, please refer to Table 1.

Given the substantial amount of missing data (58%) in the cholesterol-related information, we have prepared two versions of input data for analysis: one including cholesterol-related variables and one excluding them. For traditional classification models, we utilize the restricted datasets without missing data to train the models. In contrast, for LLMs, we employ serialization techniques similar to those described in (Hegselmann et al., 2023) to reformat the data. This involves converting data into a format like "Variable 1: value1, ..., Variable m: value m" . Missing data is coded as "NA" in the text input. To indicate to the LLM that this is a classification task, we append an appropriate prompt: "Does this patient have uncontrolled blood pressure (SBP $\geq$ 140 mm Hg or DBP $\geq$ 90 mm Hg)". Our target variable for this study is uncontrolled blood pressure, defined as a systolic blood pressure (SBP) $\geq$ 140 mm Hg or diastolic blood pressure (DBP) $\geq$ 90 mm Hg. For standard classification models, the output is categorized as 0 and 1. In the case of LLMs, the outputs are designated as "Yes." and "No.".

LLM model and supervised fine-tuning. For this study, we chose QWEN (Bai et al., 2023) as our foundational model. QWEN is available in different sizes, with 14 billion, 7 billion, and 1.8 billion parameters, catering to various computational needs. These models have undergone extensive pre-training on a massive corpus, encompassing trillions of tokens. QWEN has proven to be a robust competitor among existing open-source models, exhibiting performance levels that rival, and in some cases, match those of proprietary models across comprehensive benchmarks and human evaluations. Its effectiveness is further highlighted by its selection as the base model by the winning team (Percent-BFD, 2023) in the A100-GPU track of the NeurIPS 2023 LLM Efficiency Challenge: 1LLM+1GPU+1Day. In this contest, QWEN demonstrated superior performance under the constraints of limited computing resources, making it an ideal choice for this datafest challenge. For this study, we used QWEN with 1.8 billion and 7 billion parameters.

For classification, we first perform supervised fine-tuning (SFT) on QWEN models to enhance the performance of pre-trained models. Initially, models like LLM are trained on vast, generalized datasets, which enables them to learn a broad range of patterns and relationships in the data. However, to tailor these models to specific tasks or domains, additional training, known as fine-tuning, is required. In supervised fine-tuning, the model is further trained on a smaller, task-specific dataset where the desired outputs are known. This "supervised" aspect helps adapt the model to the specific task. For the SFT of our models, we utilized code from the LLaMA-Factory repository (hiyouga, 2023). To enhance computational efficiency, we applied Low-Rank Adaptation (LoRA, Hu et al. (2021)) for the QWEN 1.8B model and a quantized version of LoRA (qLoRA, Dettmers et al. (2023)) for the QWEN 7B model. These adaptations ensure that the QWEN models are effectively tailored and optimized for the classification task, balancing performance with computational resource considerations.

Following the SFT, we employed the QWEN model for the classification task. In this task, we provided patient information to QWEN and asked it to determine whether the patient has uncontrolled BP. To further assess QWEN's understanding of the mechanisms underlying BP control, we utilized specific prompts aimed at eliciting its insights into this area. For instance, we posed the core question of the datafest challenge: "What are the potential causes or correlates of worsening BP control among US adults with hypertension over the past decade?". This approach is designed to gauge QWEN's capacity to identify and articulate factors influencing BP control. Additionally, to verify the accuracy and relevance of the variables identified by QWEN in its response, we conducted a comparison with the feature importance derived from traditional methods.

Results

Classification. We first compare classification accuracy between QWEN and traditional methods such as random forest (RF), gradient boosting (GB), and logistic regression (LR). We note that the traditional methods were trained on datasets excluding cholesterol information to accommodate a larger sample size. Our findings (Table 2) revealed that the accuracy of LLMs was inferior to that of these conventional techniques. Furthermore, we observed no significant difference in performance between the Qwen 1.8B and 7B models. Additionally, incorporating cholesterol information into the dataset did not result in any improvement in QWEN's performance. This outcome suggests that, in this specific context, traditional machine learning methods may hold an advantage over LLMs.

Feature Importance. Table 1 provides a comprehensive overview of the various factors influencing BP control in hypertensive patients. Furthermore, Figure 1 shows the top 15 most important features as determined by random forest, gradient boosting and logistic regression. Overall, there is a high level of agreement across the three models, where demographic covariates such as age, BMI, and race category appear in the two ensemble models, only age appears in the largest coefficients of the logistic regression model. The rest of the most important features all relate to different types of hypertension medication.

LLM Reasoning. Next, we evaluated QWEN's reasoning capabilities regarding BP control. As outlined in the classification section, after SFT, we provided detailed patient information to QWEN and inquired about BP control. In this scenario, QWEN's output is limited to either "Yes." or "No." (Figure 2, Prompt 1). To gain a deeper understanding of the mechanisms behind BP control, we asked a more general question to QWEN, focusing on potential causes or correlates of worsening BP control, which aligns with the core question of the datafest challenge (Figure 2, Prompt 2). QWEN's response highlighted various factors, including medication adherence, weight management, physical activity levels, smoking habits, alcohol consumption, and other lifestyle elements as potential contributors to changes in BP control. Interestingly, QWEN's response not only reflected the variables present in our dataset but also extended to include information beyond it. This demonstrates its capacity to utilize its extensive pre-training knowledge, provideing a more comprehensive understanding in complex health issues like BP control, where such extra information could guide the future data collection.

Furthermore, to delve deeper into the specific relationships within the NHANES dataset, we supplemented our queries to QWEN by providing the names of variables present in the data (Figure 2, Prompt 3). In this adjusted approach, QWEN's responses were more focused, including only those variables that were part of the dataset. This time, the majority of the factors identified by QWEN as influencing BP control were related to medication usage, disease history, and lifestyle choices.

Our analysis revealed a notable alignment between QWEN and traditional analytical methods, with both identifying medication as a top variable influencing BP control. To gain a more detailed understanding of the specific impact of different medications on BP control, we posed a further question to QWEN, inquiring about the most important medicine for managing BP (Figure 2, Prompt 4). Surprisingly, QWEN identified ACE (Acetolipidic coicosteroid) inhibitors as the most crucial medication for BP control. This response is in line with the findings from traditional methods, where ACE inhibitors consistently appeared as a top medicine in terms of coefficient or feature importance.

This congruence between QWEN's responses and traditional analytical methods underscores the effectiveness of LLMs in identifying key factors in healthcare scenarios. Both approaches highlight medication and disease history as shared important variables in the management of BP. The consistency in QWEN's responses with established analytical findings demonstrates its capability to understand and process complex medical data, offering insights that align with conventional medical knowledge and practices. This alignment further validates the potential of LLMs like QWEN in complementing traditional methods for data analysis in healthcare.

Discussion

In this report, we utilized LLMs to analyze the NHANES dataset. While LLMs' classification performance lagged behind traditional methods, they excel in providing detailed explanations and accurate reasoning. This highlights the potential of LLMs in healthcare data analysis and opens doors to future applications of LLMs in medical studies.

References

Bai, J., S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Gaziano, T., K. S. Reddy, F. Paccaud, S. Horton, and V. Chaturvedi (2006). Cardiovascular disease. *Disease Control Priorities in Developing Countries. 2nd edition*.

Hegselmann, S., A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag (2023). Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR.

hiyouga (2023). Llama factory. `https://github.com/hiyouga/LLaMA-Factory`.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Muntner, P., M. A. Miles, B. C. Jaeger, L. Hannon Iii, S. T. Hardy, Y. Ostchega, G. Wozniak, and J. E. Schwartz (2022). Blood pressure control among us adults, 2009 to 2012 through 2017 to 2020. *Hypertension 79*(9), 1971–1980.

Percent-BFD (2023). Neurips large language model efficiency challenge: 1 llm + 1 gpu + 1 day submission. `https://github.com/Percent-BFD/neurips_submission`.

Roziere, B., J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, et al. (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Table 1: Descriptive information of people with hypertension in the NHANES data

	Uncontrolled blood pressure	
	No (n=9327)	Yes (n=11082)
Age category, years (%)		
18 to 44	917 (9.8)	1602 (14.5)
45 to 64	4043 (43.3)	4117 (37.2)
65 to 74	2472 (26.5)	2561 (23.1)
75+	1895 (20.3)	2802 (25.3)
Race/ethnicity (%)		
Non-Hispanic White	4364 (46.8)	4692 (42.3)
Non-Hispanic Black	2525 (27.1)	3112 (28.1)
Non-Hispanic Asian	402 (4.3)	444 (4.0)
Hispanic	1732 (18.6)	2461 (22.2)
Other	304 (3.3)	373 (3.4)
Gender = Women (%)	4943 (53.0)	5492 (49.6)
Self-reported antihypertensive medi use = Yes (%)	9327 (100.0)	5363 (48.6)
Number of antihypertensive medi classes (%)		
None	246 (2.7)	5108 (46.5)
One	3345 (36.3)	2388 (21.8)
Two	3237 (35.1)	1939 (17.7)
Three	1667 (18.1)	1051 (9.6)
Four or more	718 (7.8)	491 (4.5)
Combination therapy = Yes (%)	2029 (22.0)	1071 (9.8)
Time since cholesterol measured (%)		
In the past year	3137 (78.2)	2564 (58.3)
1 to 5 years ago	636 (15.9)	981 (22.3)
>5 years ago (possibly never)	238 (5.9)	851 (19.4)
Total cholesterol, mg/dL (mean (SD))	188.01 (42.35)	201.25 (44.55)
Taking a cholesterol-lowering medication = Yes (%)	1935 (47.3)	1257 (27.5)
Very high ASCVD risk = Yes (%)	770 (18.8)	653 (14.3)
Smoking status (%)		
Never	4623 (49.6)	5672 (51.5)
Former	3185 (34.2)	3351 (30.4)
Current	1508 (16.2)	1986 (18.0)
Body mass index, kg/m2 (%)		
<25	1400 (15.6)	2645 (24.6)
25 to <30	2874 (32.0)	3651 (34.0)
30 to <35	2361 (26.3)	2375 (22.1)
35+	2339 (26.1)	2072 (19.3)
Prevalent diabetes = Yes (%)	2926 (31.4)	2450 (22.1)
Prevalent chronic kidney disease = Yes (%)	2816 (30.2)	3726 (33.6)
History of myocardial infarction = Yes (%)	979 (10.5)	769 (7.0)
History of coronary heart disease = Yes (%)	1448 (15.6)	1140 (10.4)
History of stroke = Yes (%)	843 (9.1)	823 (7.5)
History of ASCVD = Yes (%)	2014 (21.6)	1725 (15.6)
History of heart failure = Yes (%)	838 (9.0)	612 (5.6)
History of CVD = Yes (%)	2292 (24.6)	1949 (17.6)

Table 2: Performance Metrics of random forest, gradient boosting, logistic regression models and LLMs.

Metric	RF	GB	LR	Q1.8	Q1.8-C	Q7B	Q7B-C
Accuracy	0.7533	0.7530	0.7459	0.6046	0.6024	0.6001	0.6024
ROC	0.7656	0.7634	0.7549	-	-	-	-
F1-Score	0.7238	0.7313	0.7287	0.6090	0.6042	0.6057	0.6061

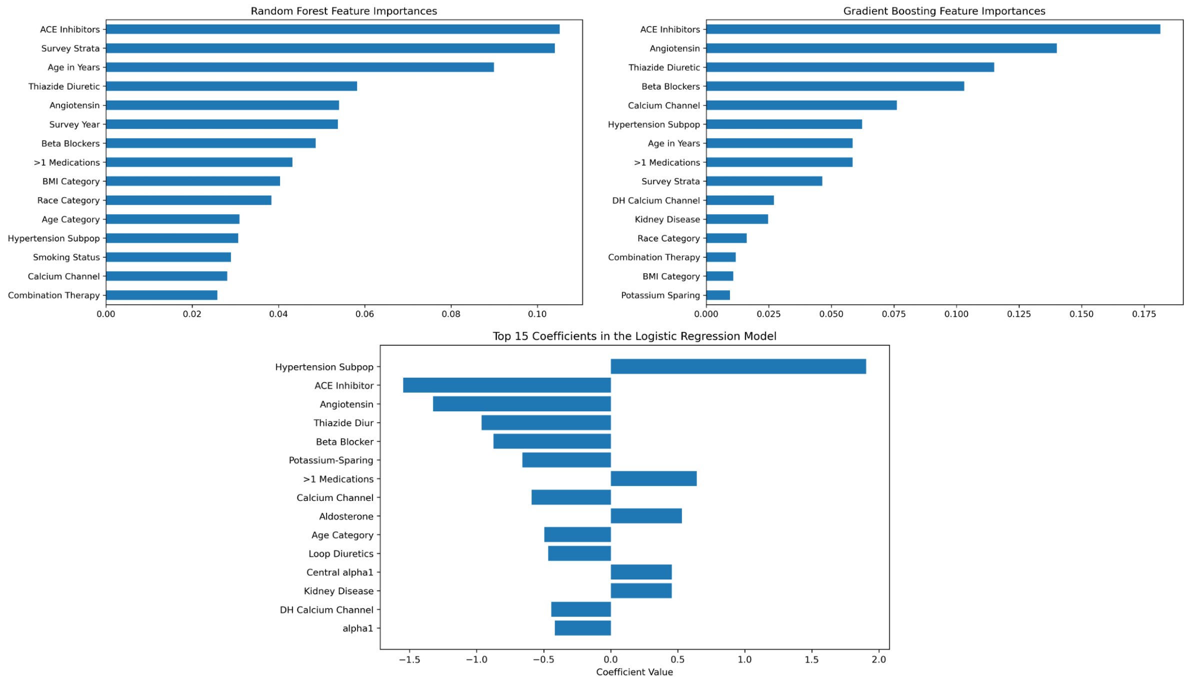

Figure 1: Feature importance for best fit random forest classifier (top left), best fit gradient boosting classifier (top right), and largest coefficients for logistic regression model (bottom).

Prompt 1:

- Input: Mobile examination center weights: 7991.127474, [other variable names and values used], History of CVD: No. Does this patient have uncontrolled blood pressure (SBP $\geq$ 140 mm Hg or DBP $\geq$ 90 mm Hg)?
- Output: Yes.

Prompt 2:

- Input: What are the potential causes or correlates of worsening BP control among US adults with hypertension over the past decade?
- Output: The potential causes include: 1. Reduced BP control measures: A reduction in BP control measures, such as taking medication as prescribed, monitoring blood pressure regularly, and increasing physical activity, can contribute to worsening BP control. 2. Increasing weight: Excess weight can put additional strain on the body and lead to increased blood pressure, as well as other health problems. 3. Reduced physical activity: Reduced physical activity can also lead to increased blood pressure and a decline in overall health. 4. Unhealthy lifestyle choices: Unhealthy lifestyle choices such as smoking and excessive alcohol consumption can increase blood pressure and other health problems. 5. Lifestyle factors: Other lifestyle factors, such as diet, sleep, and stress, can also contribute to worsening BP control. 6. Other health conditions: Other health conditions, such as kidney disease, diabetes, and heart disease, can also impact BP control.

Prompt 3:

- Input: Suppose I give you the information including [variable names used in the study] I want you to tell me whether this person has uncontrolled blood pressure. What are the variables that affect your decision?
- Output: The causal variables for worsen BP control are: 1. The number of antihypertensive medication classes, number of antihypertensive medication pills, combination therapy, taking two or more antihypertensive medication pills, ACE inhibitors, Aldosterone antagonists, (not showing some variable names due to the length of the response), Direct renin inhibitors, Direct vasodilators, Smoking status, Body mass index, Prevalent diabetes, Prevalent chronic kidney disease, History of myocardial infarction, History of coronary heart disease, History of stroke, History of ASCVD, History of heart failure, History of CVD.

Prompt 4:

- Inputs: Which medicine is most important for blood pressure control in hypertension patient?
- Answer: The most important medicine for blood pressure control in hypertension patient is ACE inhibitors (Acetolipidic coicosteroid). Explanation: Acetolipidic coicosteroids, like ACE inhibitors, are drugs that block the enzymes that convert the hormone angiotensin I into angiotensin II. This means that the body is less likely to produce angiotensin II, and thus prevents the "coughing" that occurs when the body cannot effectively control blood pressure. ACE inhibitors also help to lower blood pressure by relaxing the blood vessels and reducing the amount of protein in the blood, which is a sign of inflammation. They are used in hypertension patients to control blood pressure and to prevent complications like heart disease and stroke.

Figure 2: Prompts for LLM reasoning.