Swedish Medical Benchmark, an evaluation framework for LLMs in the Swedish medical domain

Birger Moëll
KTH Royal Institute of Technology
bmoell@kth.se

Fabian Farestam
ETH Zürich
ffarestam@student.ethz.ch

Abstract

We present the Swedish Medical Benchmark (SMB), a comprehensive evaluation framework for assessing Large Language Models (LLMs) in the Swedish medical domain. The SMB addresses the lack of language-specific, clinically relevant benchmarks by incorporating four datasets: translated PubMedQA questions, Swedish Medical Exams, Emergency Medicine scenarios, and General Medicine cases. Our evaluation of state-of-the-art LLMs and Swedish open-source models reveals performance variations, with more capable models generally performing better. Claude-3.5 Sonnet. GPT-4 and GPT-4o emerge as the top performers, aligning with other evaluations of their capabilities. The SMB not only provides a rigorous evaluation tool but also offers insights into LLMs' capabilities and limitations in Swedish healthcare applications. By open-sourcing the SMB, we aim to democratize AI evaluation in healthcare, enabling researchers, clinicians, and developers worldwide to contribute to its refinement and expansion. This collaborative approach facilitates rapid iteration, cross-validation, and the development of culturally and linguistically adapted benchmarks. Our open-source initiative supports transparency in AI assessment, fosters innovation, and promotes the responsible development of AI-assisted medical practices. The SMB serves as a crucial step towards the ethical integration of AI in healthcare, emphasizing comprehensive, community-driven assessment before clinical deployment.

1 Introduction

The integration of language models (LMs) in medicine presents both significant opportunities and challenges. Large language models (LLMs), such as GPT-3.5 and GPT-4, have demonstrated promising capabilities in various healthcare applications, from answering medical questions to assisting in clinical decision-making (Omiye et al., 2023; Wang et al., 2023b). These models have the potential to enhance clinical decision-making, automate healthcare tasks, and improve patient outcomes (Li and Min, 2023; Wang et al., 2023b). However, the risks associated with their use are substantial; inaccurate medical advice generated by these models could lead to severe consequences, including misdiagnosis and inappropriate treatment, potentially compromising patient health (Ziaei and Schmidgall, 2023). Given these risks, the integration of LLMs in healthcare necessitates careful consideration and rigorous evaluation to maximize positive impacts while minimizing potential harm (Wang et al., 2023a; Barnard et al., 2023). A critical step in determining the suitability of language models for medical applications is to assess their performance using domain-specific benchmarks. While several benchmarks exist for evaluating language models in the medical domain, they primarily focus on English-language tasks and often include questions that do not fully represent real-world clinical scenarios (Omiye et al., 2023). To address these limitations, we introduce the Swedish Medical Benchmark (SMB), a comprehensive benchmarking suite consisting of Swedish-language questions in four distinct areas:

1. SWE-PUBMEDQA: Translated questions from established English-language benchmarks

2. SWE MEDICAL EXAMS: Questions from Swedish Medical Exams

3. SWE SPEC EM: Clinical specialist questions for the emergency medicine domain

4. SWE SPEC GM: Clinical specialist questions for the general practitioner domain

This benchmark aims to provide a more accurate assessment of LLMs' capabilities in the Swedish

medical context, ensuring that the models are evaluated against tasks that closely resemble clinical practice in Sweden (Wang et al., 2023b).

2 Method

The Swedish Medical Benchmark (SMB) was developed using a multi-faceted approach to create a comprehensive and robust evaluation tool for Large Language Models (LLMs) in the Swedish medical domain. Our methodology focused on three key strategies:

1. Translating existing English-language medical questions to Swedish using LLMs.

2. Creating high-quality clinical patient cases from medical information about various disorders, assisted by LLMs.

3. Incorporating standard medical tests used to assess medical students at Swedish Medical Universities.

2.1 Question Format and Evaluation

Our benchmarks are built using multiple-choice questions (MCQs) where the LLM needs to decide which answer is most plausible. The response format is a multiple-choice question with one correct answer. This approach aligns with established practices in medical education and assessment (Case and Swanson, 2002).

2.2 Open Source and Collaborative Approach

The Swedish Medical Benchmark is an open-source project. We actively encourage participants to contribute both with improvements to our benchmarking tool and by adding additional datasets. Benchmarks and evaluation code are available on github. [1]

2.3 Benchmarks

The Swedish Medical Benchmark comprises four distinct benchmarks, each designed to evaluate different aspects of medical knowledge and reasoning:

2.3.1 PubMedQA-Swedish-1000

PubMedQA-Swedish is a translated version of the PubMedQA dataset, including 1000 questions with yes / no / maybe answers.

- **Source:** Translated from the original English PubMedQA dataset (Jin et al., 2019).

---

[1]https://github.com/BirgerMoell/swedish-medical-benchmark/

Table 1: Swedish Medical Benchmark

Benchmark	Description
PubMedQA-Swedish	1000 yes/no/maybe questions; Translated from English PubMedQA (Jin et al., 2019); LLM translation with human review; Allows multilingual comparison
Medical Doctors Knowledge Test	535 multiple-choice questions (5 options); Adapted from Swedish clinical exams; Covers various medical specialties; Assesses broad medical knowledge (Norcini et al., 2011)
Emergency Medicine (SE-EM)	464 multiple-choice questions based on patient description with follow up questions(4 options); Focuses on time-critical scenarios; Tests ability to identify and prioritize severe medical issues
General Medicine (SE-GM)	666 multiple-choice questions based on patient description with follow up questions (4 options); Covers 200+ common disorders; Reflects >50% of patient interactions; Tests diagnosis and severity assessment in primary care

- **Translation Process:** Utilized state-of-the-art LLMs for translation, followed by human review to ensure accuracy and cultural appropriateness.

- **Evaluation:** When evaluating, each LLM is asked to answer only with "yes," "no," or "maybe" for each question.

- **Significance:** This dataset tests the model's ability to comprehend and reason about medical literature in Swedish, a crucial skill for evidence-based practice (Smith et al., 2018).

2.3.2 Swedish Medical Doctors Knowledge Test

This dataset consists of questions from the clinical exam for doctors in Sweden, adapted for LLM evaluation.

- **Content:** 535 multiple-choice questions, each with five different answer options.

- **Scope:** Covers a wide range of medical specialties and topics relevant to clinical practice in Sweden.

- **Adaptation:** Questions related to images have been omitted to focus on text-based reasoning.

- **Significance:** This dataset assesses the LLMs medical knowledge across various specialties, mimicking the breadth of knowledge required of practicing physicians (Norcini et al., 2011).

2.3.3 Emergency Medicine (SE-EM)

Specialist exam Emergency Service is a benchmark focused on time-critical healthcare issues that is in the domain of a specialist in emergency medicine.

- **Content:** 464 multiple choice question, each with four different answer options. Questions cover a range of emergency scenarios, focusing on realistic scenarios in emergency medicine.

- **Rationale:** Including these questions was crucial as it is vital for any model in healthcare to recognize when an issue needs immediate medical attention.

- **Significance:** Performance on these questions is a key indicator of a model's safety for general audience use, as it demonstrates the ability to identify and prioritize severe medical issues (Croskerry, 2013).

2.3.4 General Medicine (SE-GM)

Specialist Exam General Medicine focused on general medicine questions that is in the domain of a specialist in general medicine. General Medicine is the speciality that has the largest volume of patient interaction and the benchmark has questions on common medical issues.

- **Rationale:** Encompasses more than 50% of all patient interactions, reflecting the diverse nature of primary care.

- **Content:** 666 multiple choice questions, covering over 200 common disorders encountered in general medicine.

- **Significance:** This dataset tests the model's ability to accurately diagnose common disorders and assess their severity, mirroring the key skills required in general practice (Reilly, 2016).

2.4 Evaluation Metrics

We used accuracy as our evaluation metric for all our benchmarks. Furthermore a total SMB score was calculated with each benchmark weighted relative to the amount of questions included. The score was calculated with a total accuracy of 100 being a perfect score on all tests.

By combining these diverse datasets and rigorous evaluation methods, the Swedish Medical

Table 2: Composition of the Swedish Medical Benchmark

Benchmark	Number of Questions	Percentage
PQ-S-1000	1000	37.52%
SMDT	535	20.08%
EM	464	17.41%
GM	666	24.99%
Total	2665	100.00%

Benchmark aims to provide a comprehensive assessment of LLMs' capabilities in the Swedish medical domain, focusing on both broad medical knowledge and critical decision-making skills essential for safe and effective healthcare applications.

3 Result

We selected a range of state-of-the-art LLMs in healthcare including GPT-4o, GPT-4, Claude 3.5 and Llama-3.1, as well as Swedish open source LLMs such as Eir and evaluated their performance on our benchmark, available in Table 3. The results show that performance in the Swedish Medical domain matches performance in other domains, with generally more capable model also having better performance in the Swedish Medical domain. Notably, high performing model such as gpt-4, gpt-4o and Claude 3.5 are the best performing models, in line with overall ranking of models on general benchmarks such as Chatbot Arena (Chiang et al., 2024). Several models pass the SMDT medical test with a score above, 60 points (passing grade). The variation in scores between Llama3-70b and Llama-3.1-70b, an updated version of the same model is interesting. Llama-3.1-70b improves on SMDT and GM but has worse scores on EM. In a clinical setting, benchmarking before deployment of updated models could help improve patients outcomes by looking at, what areas does the model perform well in. These variations in score highlights why benchmarking language models is necessary in the medical domain. In a future where models will be deployed in the healthcare sector, benchmarking before updating a model will likely be a necessary safety precaution.

4 Discussion

The development of objective benchmarks, such as the Swedish Medical Benchmark (SMB) presented in this study, constitute a crucial step towards the re-

Table 3: Performance of LLMs on the Swedish Medical Benchmark

Model	PQ-S	SMDT	EM	GM	SMB
GPT-4o	-	**83.18**	90.51	88.88	
GPT-4	53.90	79.07	**93.10**	**93.09**	**75.57**
Claude-3.5	-	83.74	-	-	
Llama3-70b	**56.00**	69.91	74.35	67.57	64.88
Llama3-8b	50.50	41.68	-	-	
Llama3.1-70b	-	71.40	62.93	71.02	
Llama3.1-8b	-	6.36	-	-	
Gemma2-9b	-	61.31	-	-	
Gemma-7b	48.70	27.48	-	-	
EIR	46.50	-	-	-	
GPT-3.5	27.40	-	-	-	

Note: "-" indicates no evaluation. Accuracy in %. PQ-S: PubMedQA-Swedish-1000; SMDT: Swedish Medical Doctors Test; EM: Emergency Medicine; GM: General Medicine; SMB: Swedish Medical Benchmark.

sponsible utilization of LLMs in healthcare. These benchmarks serve multiple critical functions:

1. **Performance Assessment:** They provide a standardized measure of LLM performance in medical knowledge and reasoning tasks, allowing for comparative analysis across different models (Roberts et al., 2023).

2. **Safety Evaluation:** Benchmarks help identify potential gaps or inconsistencies in model knowledge, which is essential for patient safety (Challen et al., 2019).

3. **Transparency:** By offering clear metrics, benchmarks enhance transparency in AI capabilities, fostering trust among healthcare professionals and the public (Ghassemi et al., 2021).

4. **Guidance for Development:** They provide direction for further refinement and specialization of LLMs for medical applications (He et al., 2019).

It is important to note that while the public is already turning to LLMs for health-related information (Nadarzynski et al., 2019), the scientific community bears the responsibility of ensuring that this usage is informed by robust evidence. The SMB and similar benchmarks contribute to this goal by offering objective assessments of model performance.

4.1 Risks of medical LLMs

Deploying LLMs in healthcare poses risks, especially the potential overconfidence in model outputs, which could harm patient care (Challen et al.,

2019). While benchmarks like the SMB are useful for comparison, they may falsely suggest clinical readiness. High SMB scores indicate proficiency but do not ensure clinical safety or effectiveness. Therefore, subjective evaluations and real-world trials are crucial to complement benchmarks and ensure safe, effective deployment in healthcare (He et al., 2019).

4.2 Limitations

Our work in reviewing the questions in the benchmark is ongoing, and not all questions have been reviewed by a medical professional. As such, there might be inaccuracies in both questions and answers. We are actively working with medical professionals to review answers and hope to collaborate on an effort to medically validate the questions in the benchmark.

4.3 Open Source Medical AI

Open source is vital for developing robust medical AI solutions, enabling free collaboration and result sharing. Advancements made through open source can be utilized by all healthcare practitioners. Given health's universal importance, we welcome contributions from both medical and AI professionals to the ongoing development of the SMB.

5 Conclusion

The Swedish Medical Benchmark (SMB) offers a comprehensive, language-specific tool for evaluating LLMs in the medical domain. Our findings reveal both the potential and limitations of LLMs in healthcare applications. As these models evolve, frameworks like the SMB are crucial for ensuring their safe, effective, and ethical integration into healthcare systems. Understanding LLM capabilities is essential for their responsible healthcare integration. The SMB results demonstrate significant LLM proficiency in the Swedish medical domain, suggesting potential to enhance healthcare delivery. However, implementation requires caution. We advocate a measured approach, leveraging LLM capabilities while keeping human professionals central to decision-making. This human-in-the-loop model realizes AI benefits while mitigating risks. We urge healthcare stakeholders to develop frameworks for judicious LLM incorporation, balancing innovation with patient safety and ethical considerations.

References

Francois Barnard, Marlize Van Sittert, and Siri J. Rambhatla. 2023. Self-diagnosis and large language models: A new front for medical misinformation. *ArXiv*, abs/2307.04910.

Susan M Case and David B Swanson. 2002. Constructing written test questions for the basic and clinical sciences. *National Board of Medical Examiners Philadelphia, PA*.

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Pat Croskerry. 2013. From mindless to mindful practice—cognitive bias and clinical decision making. *New England Journal of Medicine*, 368(26):2445–2448.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.

Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Q. Li and X. Min. 2023. Unleashing the power of language models in clinical settings: A trailblazing evaluation unveiling novel test design.

Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study. *Digital Health*, 5:2055207619871808.

John Norcini, Brownell Anderson, Valdés Bollela, Vanessa Burch, Manuel João Costa, Robbert Duvivier, Robert Galbraith, Richard Hays, Alison Kent, Valerie Perrott, et al. 2011. Criteria for good assessment: Consensus statement and recommendations from the ottawa 2010 conference. *Medical teacher*, 33(3):206–214.

J. Omiye, Haiwen Gui, Shawheen J. Rezaei, James Zou, and Roxana Daneshjou. 2023. Large language models in medicine: the potentials and pitfalls. *ArXiv*, abs/2309.00087.

Brian M Reilly. 2016. The new landscape of the "general practice" specialty. *The New England journal of medicine*, 374(18):1701.

Michael Roberts, Philip Asare, Edward Choi, Abraham T Eshete, Judy Wawira Gichoya, Ophir Gozes, Rui Guo, Eddy Korot, Hoyeon Lee, Richard Lindsey, et al. 2023. Large language models in medicine. *Nature Medicine*, 29(7):1736–1744.

Valerie Smith, Declan Devane, Cecily M Begley, and Mike Clarke. 2018. Systematic reviews in medical education: A practical approach: Amee guide 94. *Medical teacher*, 40(2):117–132.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *ArXiv*, abs/2306.09968.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023b. Are large language models ready for healthcare? a comparative study on clinical language understanding. *ArXiv*, abs/2304.05368.

Rojin Ziaei and Samuel Schmidgall. 2023. Language models are susceptible to incorrect patient self-diagnosis in medical applications. *ArXiv*, abs/2309.09362.

A Dataset evaluation results

A.1 PubmedQA-Swedish-1000

The results of the evaluation can be seen in 4. Different LLM:s had different prompts, which are noted below. Note that "{question}" is where the actual question was inserted.

EIR had the following prompt:

> Var vänlig och överväg varje aspekt av medicinska frågan nedan noggrant. Ta en stund, andas djupt, och när du känner dig redo, vänligen svara med endast ett av de fördefinierade svaren: 'ja', 'nej', eller 'kanske'. Det är viktigt att du begränsar ditt svar till dessa alternativ för att säkerställa tydlighet i kommunikationen.
>
> Fråga: {question} svara bara 'ja', 'nej' eller 'kanske'

Gemma-7b-it, GPT-3.5-t, Llama3-8b had the following prompt:

> Du är en utmärkt läkare och skriver ett läkarprov. Var vänlig och överväg varje aspekt av medicinska frågan nedan noggrant. Ta en stund, andas djupt, och när

du känner dig redo, vänligen svara med endast ett av: 'ja', 'nej', eller 'kanske'. Det är viktigt att du begränsar ditt svar till dessa alternativ för att säkerställa tydlighet i kommunikationen. Du ska bara svara med 'kanske' ifall det kan vara fallet, annars svara 'ja' eller 'nej'. Svara med det svaret som du tror är mest korrekt.

Kom ihåg att svaret inte alltid är 'ja', så fundera ordentligt innan du svarar.

Fråga: {question}

Svara endast 'ja', 'nej' eller 'kanske'.

GPT-4-t had the following prompt:

Du är en utmärkt läkare och skriver ett läkarprov. Var vänlig och överväg varje aspekt av medicinska frågan nedan noggrant. Ta en stund, andas djupt, och när du känner dig redo, vänligen svara med endast ett av: 'ja', 'nej', eller 'kanske'. Det är viktigt att du begränsar ditt svar till dessa alternativ för att säkerställa tydlighet i kommunikationen.

Fråga: {question}

Svara endast 'ja', 'nej' eller 'kanske'.

Llama3-70b had the following prompt:

Du är en utmärkt läkare och skriver ett läkarprov. Var vänlig och överväg varje aspekt av medicinska frågan nedan noggrant. Ta en stund, andas djupt, och när du känner dig redo, vänligen svara med endast ett av: 'ja', 'nej', eller 'kanske'. Det är viktigt att du begränsar ditt svar till dessa alternativ för att säkerställa tydlighet i kommunikationen. Du ska bara svara med 'kanske' ifall det kan vara fallet, annars svara 'ja' eller 'nej'. Svara med det svaret som du tror är mest korrekt.

Kom ihåg att svaret inte alltid är 'ja', så fundera ordentligt innan du svarar.

Fråga: {question}

Svara endast 'ja', 'nej' eller 'kanske'.

Table 4: Performance metrics of models on PubMedQA-Swedish

Model	Acc (%)	Yes	No	Maybe	
EIR	46.50	56.35	44.72	6.36	
Gemma-7b-it	48.70	65.76	19.50	7.05	
GPT-4-t	53.90	67.56	36.23	14.07	Note:
GPT-3.5-t	27.40	41.36	14.59	17.58	
Llama3-8b	50.50	63.48	34.15	3.10	
Llama3-70b	56.00	67.77	41.83	1.72	

Acc = Accuracy. Yes, No, and Maybe columns show F1 scores. Sample sizes: Yes (n=552), No (n=338), Maybe (n=110).

A.2 Swedish medical doctors knowledge test

All LLM:s have been evaluated on the same prompt:

Du är en utmärkt läkare och skriver ett läkarprov. Var vänlig och överväg varje aspekt av medicinska frågan nedan noggrant. Ta en stund, andas djupt, och när du känner dig redo, vänligen svara med endast ett av alternativen. Svara med hela svarsalternativet. Utöver det är det viktigt att du inte inkluderar någon annan text i ditt svar.

The results for SMDT can be seen in Table 5.

Model	Accuracy (%)
gpt-4o-2024-08-06	83.18
gpt-4	79.07
llama3-70b	69.91
llama3-8b	41.68
llama3.1-70b-versatile	71.40
llama3.1-8b-instant	6.36
gemma2-9b-it	61.31
gemma-7b-it	27.48
claude-3-5-sonnet-20240620	83.74

Table 5: Performance metrics of evaluated models.

The low performance of llama3.1-8b-instant can be attributed to it not being willing to answer medical questions and asking for one to seek out a medical professional.

claude-3-5-sonnet-20240620 passed every exam, while gpt-4o-2024-08-06 and gpt-4 failed one exam out of the total 26 exams.

B Case Description

Emma, en 8-årig flicka, har haft buksmärtor i ett par dagar som förvärrats och nu åtföljs av illamående och kräkningar. Hon har också haft feber och hennes föräldrar märker att hon är ovanligt trött och orolig. Vid undersökning visar Emma ömhet i nedre högra kvadranten av buken och har svårt att röra sig utan att känna smärta.

Questions

Q1: Vilken av följande är den mest sannolika diagnosen för Emmas symptom?

    A) Obstipation
    B) Gastroenterit
    C) Appendicit
    D) Urinvägsinfektion

Correct Answer: C) Appendicit

Explanation: Emmas symptom på buksmärtor, illamående, kräkningar och ömhet i nedre högra kvadranten tyder på appendicit.

Q2: Vilken undersökning är förstahandsval vid misstanke om appendicit hos Emma?

    A) Lungröntgen
    B) Kolonröntgen
    C) Buköversikt (BÖS)
    D) Ultraljud

Correct Answer: D) Ultraljud

Explanation: Ultraljud är förstahandsval vid misstanke om appendicit hos barn eftersom det är en säker och effektiv metod för att visualisera en inflammerad appendix.

Q3: Vilken differentialdiagnos är viktig att överväga vid akut buksmärta hos barn som Emma?

    A) Pneumoni
    B) Otit
    C) Urinvägsinfektion
    D) Alla ovanstående

Correct Answer: D) Alla ovanstående

Explanation: Pneumoni, otit och urinvägsinfektion är alla viktiga differentialdiagnoser att överväga vid akut buksmärta hos barn, eftersom dessa tillstånd kan presentera sig med buksmärta.

Q4: Vad är den lämpligaste omedelbara åtgärden för Emma på akutmottagningen?

    A) Hemgång med poliklinisk uppföljning
    B) Observation och smärtlindring
    C) Akut operation
    D) Antibiotikabehandling och avvakta

Correct Answer: B) Observation och smärtlindring

Explanation: Initialt bör Emma observeras och få smärtlindring för att noggrant bedöma hennes tillstånd och behovet av eventuell kirurgi.

Recommended Treatment Plan

- **Misstänkt appendicit:** Ultraljud för att bekräfta diagnosen, följt av observation, smärtlindring och förberedelse för eventuell kirurgi.

- **Bekräftad appendicit:** Kirurgisk intervention (appendektomi) och perioperativ antibiotikabehandling.

- **Differentialdiagnostik vid buksmärta:** Uteslutning av andra orsaker som urinvägsinfektion, pneumoni och otit genom relevant anamnes, fysisk undersökning och laboratorieprover.

Example question general medicine

Jonas, en 42-årig man, upptäcker en mjuk, elastisk knöl på nacken som har vuxit långsamt under de senaste månaderna. Knölen är tydligt avgränsbar och förskjutbar mot underlaget. Den har nyligen börjat ömma och rodna.

Questions

Q1: Vilken diagnos är mest sannolik baserat på Jonas symtom och kliniska fynd?

    A) Lipom
    B) Furunkulos
    C) Aterom (talgkörtelcysta)
    D) Angiolipom

Correct Answer: C) Aterom (talgkörtelcysta)

Explanation: Jonas har en mjuk, elastisk, tydligt avgränsbar och förskjutbar knöl som nyligen börjat ömma och rodna, vilket är typiskt för ett aterom (talgkörtelcysta).

Q2: Vilken av följande är inte en riskfaktor för att utveckla aterom?

    A) Anabola steroider

    B) Hyperhidros

    C) Diabetes

    D) Akne

Correct Answer: C) Diabetes

Explanation: Riskfaktorer för att utveckla aterom inkluderar anabola steroider, hyperhidros och akne, men diabetes är inte en direkt riskfaktor.

Q3: Vilken behandling rekommenderas om Jonas aterom är inflammerat och smärtsamt?

    A) Antibiotikabehandling

    B) Incision och dränage

    C) Strålbehandling

    D) Excision

Correct Answer: B) Incision och dränage

Explanation: Vid inflammerat och smärtsamt aterom rekommenderas incision och dränage. Antibiotikabehandling är sällan nödvändigt.

Q4: Vilken differentialdiagnos bör övervägas om Jonas knöl inte är fritt förskjutbar och har en snabb tillväxt?

    A) Lipom

    B) Angiolipom

    C) Sarkom

    D) Fibrom

Correct Answer: C) Sarkom

Explanation: Om knölen inte är fritt förskjutbar och har snabb tillväxt, bör sarkom övervägas som differentialdiagnos.

Recommended Treatment Plan

- **Aterom utan inflammation:** Ingen behandling nödvändig om det inte är kosmetiskt störande för patienten.

- **Inflammerat aterom:** NSAID för symtomlindring. Vid mer uttalad inflammation: incision och dränage.

- **Återkommande inflammerade aterom eller kosmetiskt störande:** Excision i lugnt skede för att ta bort hela cystan inklusive kapseln.