XMedQA: A Multilingual Benchmark for Region-Aware Medical Question Answering

Anonymous ACL submission

Abstract

Medicine combines global biomedical knowledge with country-specific healthcare systems, practices, and norms. However, existing Medical QA benchmarks rarely evaluate such regional specificity, often treating it as implicit in query language. We introduce **XMedQA**, a multilingual and region-aware Medical QA benchmark built from exam-style datasets in seven languages: Japanese, English, Chinese, French, Swedish, Spanish, and Russian. XMedQA preserves region-specific medical content, assigns regionality scores, and enables analysis of how regional specificity interacts with query language. Across 37 LLMs, we profile models along two axes: robustness to query language and adaptability to local medical context. The profiles reveal substantial variation, distinguishing models stable across both axes from those sensitive to language, regionality, or both. Accuracy declines as regionality increases, and this decline remains substantial even when each dataset is evaluated in its original source language. As a case study, we further find that Japanese adaptation can improve region-specific performance and transfer positively to other query languages. These findings highlight the need for region-aware alignment and evaluation of medical LLMs. The benchmark dataset will be publicly released.

1 Introduction

Medicine relies not only on globally shared biomedical knowledge but also on country-specific healthcare systems, clinical practices, cultural norms, and ethical expectations. Accordingly, the same clinical question may require different assumptions or answers across countries and regions. Recent evidence outside medicine suggests that LLM outputs can reflect country- and region-specific information environments through language-specific training data (Waight et al., 2026). Yet regional specificity is often treated as implicit in query language, assuming that questions in a given language reflect the corresponding medical context. However, recent work suggests that cross-lingual understanding and reasoning are not reducible to bilingual or parallel pretraining data (Shao et al., 2026), further motivating evaluation designs that disentangle query language from regional context.

Despite this, existing Medical QA benchmarks rarely make regional specificity explicit or support controlled multilingual comparison. Medical QA is widely used to evaluate the medical knowledge and reasoning capabilities of LLMs (Singhal et al., 2023; Liu et al., 2024; Wang et al., 2026), typically using multiple-choice or structured question-answer pairs from licensing examinations, educational resources, and benchmark collections (Jin et al., 2019, 2021). Because these questions are often expert-authored and clinically grounded, Medical QA serves as a proxy for measuring medical competence in LLMs. Without question-level regional annotation and aligned multilingual evaluation, however, existing benchmarks cannot analyze how regional specificity interacts with query language or how LLM performance varies across local medical contexts.

To address these issues, we construct **XMedQA**, a multilingual and region-aware Medical QA benchmark that explicitly captures regional specificity and its interaction with query language. XMedQA is built from diverse exam-style medical QA sources across seven languages, standardized into a unified multiple-choice format, and expanded into aligned multilingual versions. Each question is annotated with clinical categories and regionality scores (see **Fig. 1** for an overview). This design preserves country- and region-specific medical content while enabling controlled multilingual comparison. *XMedQA allows us to profile how LLMs handle region-specific medical knowledge, how this behavior changes across query languages, and how language-specific adaptation reshapes these profiles, including transfer to other languages.*

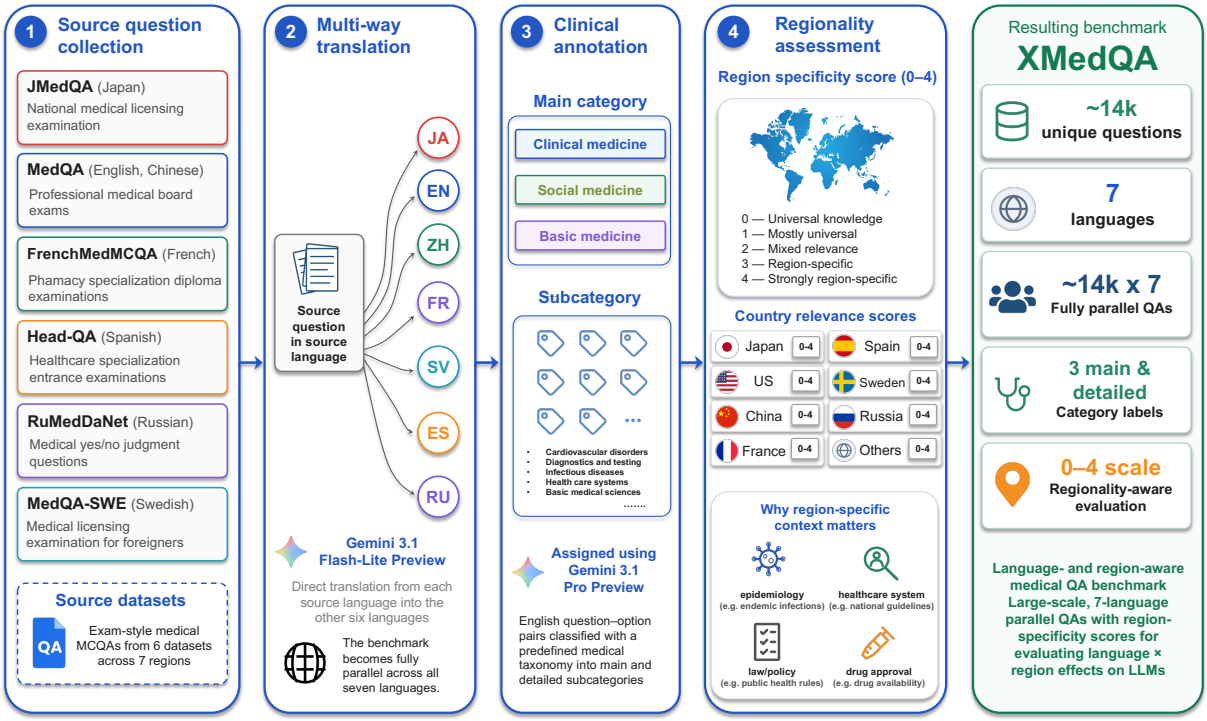

Figure 1: Overview of XMedQA benchmark construction. The pipeline standardizes source medical QA datasets into a unified format, creates fully parallel translations across seven languages, and adds clinical-category and regionality annotations for language- and region-aware evaluation.

In this work, we use XMedQA to evaluate 37 LLMs under controlled multilingual settings and analyze their behavior with respect to regionality and query language. We first profile models along two axes: robustness to query language and adaptability to local medical context. The resulting profiles reveal clear model-to-model variation, with some models robust to both axes and others fragile along one or both. We further show that accuracy declines as regionality increases, and that this decline remains substantial even when each dataset is evaluated in its original source language. Finally, we find that language-specific adaptation can improve region-specific performance and transfer positively to other languages, as observed with Japanese medical adaptation. These results show that medical LLM evaluation must account for both query language and local medical context, motivating region-aware alignment and evaluation.

Our main contributions are fourfold:

- We introduce **XMedQA**, a multilingual and region-aware Medical QA benchmark across seven languages with aligned questions, clinical-category annotations, and regionality scores.
- We provide a comprehensive profiling of 37 LLMs along two axes: robustness to query language and adaptability to local medical context.
- We show that regional specificity is a distinct source of difficulty beyond query language, with performance drops persisting under English-only and original-source-language evaluation.
- We provide a case study suggesting that Japanese adaptation can improve region-specific medical QA beyond Japanese-only evaluation, with positive transfer to other query languages depending on model family and adaptation stage.

2 Related Work

Medical Question Answering. Medical QA, especially in multiple-choice and exam-style formats, has become a standard testbed for evaluating biomedical knowledge, clinical reasoning, and medical comprehension in language models. Benchmarks such as BioASQ, PubMedQA, MedQA, and MedMCQA have established standardized Medical QA evaluation (Krithara et al., 2023; Jin et al., 2019, 2021; Pal et al., 2022).

Benchmark	#Questions	#Lang.	Parallel	Source	Regionality
MMedBench	∼54k	6		Multi-country	
MedExpQA	622	4	✓	Single-country	
WorldMedQA-V	568	5	✓	Multi-country	
Multi-OphthaLingua	1,184	7	✓	Authored	
MultiMed-X	350	7	✓	Multi-country	
XMedQA (ours)	**∼14k**	**7**	✓	**Multi-country**	✓

Table 1: Comparison of multilingual medical QA benchmarks. "#Questions" denotes the number of unique questions before multilingual expansion, where applicable. "Parallel" indicates whether the same questions are available across languages. "Regionality" indicates whether the benchmark explicitly annotates or analyzes country- or region-specific medical context; only XMedQA provides explicit question-level regionality annotations.

Subsequent studies have used these tasks to compare model architectures, domain adaptation methods, and prompting and inference strategies across general-purpose and medical-domain LLMs (Tu et al., 2024; Singhal et al., 2025; Wu et al., 2024; Zhang et al., 2023; Chen et al., 2024, 2023; Labrak et al., 2024; Nori et al., 2023).

Multilingual and Cross-Regional Medical Benchmarking. Recent benchmarks have extended medical LLM evaluation to multilingual and cross-regional settings. MedExpQA and MMedBench reported substantial performance variation across languages (Alonso et al., 2024; Qiu et al., 2024), but each language is based on different source questions, making it difficult to disentangle language effects from differences in question content and difficulty. WorldMedQA-V incorporated country-specific questions with English translations (Matos et al., 2025), while other resources such as MultiMed-X and Multi-OphthaLingua highlight multilingual disparities in medical reasoning (Gao et al., 2026; Restrepo et al., 2025).

Controlled multi-way parallel evaluation, where the same questions are evaluated across multiple query languages, can better isolate language effects, as demonstrated by general-domain aligned QA benchmarks (Lewis et al., 2020; Longpre et al., 2021). However, existing medical benchmarks do not provide such controlled parallel evaluation while preserving country- and region-specific medical content, nor do they annotate regional specificity at the question level.

Positioning of XMedQA. Beyond multilingual evaluation, XMedQA combines question-level regionality assessment with controlled multi-way parallel evaluation. This design distinguishes broadly shared medical knowledge from local medical context, including national guidelines, healthcare systems, drug availability, screening policies, epidemiology, and public health rules. It thereby enables analysis of how regional specificity and query language independently and jointly affect model performance. **Table 1** summarizes the comparison with existing benchmarks.

3 XMedQA Benchmark Construction

3.1 Overview

XMedQA is a multilingual multiple-choice medical QA benchmark built from exam-style medical QA datasets in seven languages: Japanese (ja), English (en), Chinese (zh), French (fr), Swedish (sv), Spanish (es), and Russian (ru). The construction pipeline standardizes heterogeneous source datasets into a unified format, creates aligned translated versions of each question, and adds clinical-category and regionality annotations. This design allows each source question to be evaluated both in its original language and in translated query languages. This section describes the source datasets, format standardization, translation and alignment procedure, and annotation process.

3.2 Source Datasets and Translation

XMedQA was constructed from multiple exam-style medical QA resources. The Japanese component, JMedQA, was independently constructed from the Japanese National Medical Licensing Examination (JNMLE), with details provided in **Appendix A**. The English and Chinese components were based on MedQA (Jin et al., 2021); the French component on FrenchMedMCQA (Labrak et al., 2022); the Swedish component on MedQA-SWE (Hertzberg and Lokrantz, 2024); the Spanish component on Head-QA (Vilares and Gómez-Rodríguez, 2019); and the Russian component on RuMedDaNet (Blinov et al., 2022). For the En-

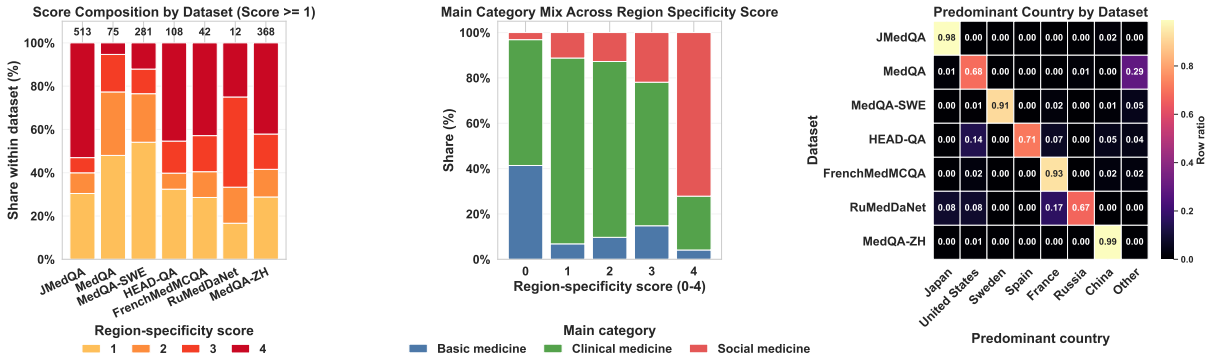

Figure 2: Regional characteristics of XMedQA. Left: Distribution of region-specificity scores among region-specific questions (score ≥ 1) by source dataset; numbers above the bars indicate the corresponding question counts. **Middle:** Main clinical-category composition by region-specificity score. **Right:** Predominant-country distribution by source dataset.

glish, Chinese, French, Spanish, and Russian components, we followed the dataset composition of the MMedBench test set (Qiu et al., 2024) and extracted the corresponding source questions from the datasets included in MMedBench. The source datasets were obtained from publicly available research resources under their stated licenses or access terms. XMedQA will be released for research and evaluation use under the applicable source-dataset terms.

Each question was translated into all non-source languages to construct aligned multilingual question sets. Translations were produced directly from each source language into each target language using Gemini 3.1 Flash-Lite Preview, with prompts designed to preserve medical meaning, answer-option order, and gold-answer consistency; details are provided in Appendix D.

3.3 Clinical and Regional Annotation

Clinical category annotation. To characterize the medical content of XMedQA beyond dataset identity and language, we annotated each question with clinical category labels. Each item was assigned to one of three main categories, defined as follows:

Category	Definition
Clinical medicine	Patient-oriented medical specialties and clinical practice, including diagnosis, treatment, and disease management.
Social medicine	Public health, healthcare systems, ethics, law, and epidemiology.
Basic medicine	Foundational biomedical sciences such as anatomy, physiology, pathology, and pharmacology.

Detailed subcategories were assigned for finer-grained analyses of topic composition across datasets and languages. Clinical categories were annotated with Gemini 3.1 Pro Preview using standardized prompts (Appendix D).

Regionality annotation. We annotated each question for regionality using an LLM-as-a-judge framework. This annotation produced two complementary scores.

First, each question was assigned an overall *region specificity score* on a 0–4 scale, measuring how strongly the question depends on country- or region-specific medical context rather than broadly shared medical knowledge. Scores range from 0 for no meaningful regional dependence to 4 for strong local-context dependence.

Second, each question received 0–4 *country relevance scores* for Japan, the United States, China, France, Spain, Sweden, Russia, and other countries to identify the most relevant national or regional medical context. In our analyses, the region specificity score quantified the degree of local dependence, whereas the country relevance scores identified the associated context.

A robustness analysis against GPT-5.5 showed substantial ordinal agreement for region-specificity scoring (quadratic weighted $\kappa = 0.775$; **Appendix B**). The prompts are provided in **Appendix D**.

3.4 Expert Review of Qualitative Examples

A physician with five years of clinical experience manually reviewed the qualitative examples shown in Appendix G. The review focused on the clinical plausibility of the displayed categories, region-specificity scores, and rationales. No major incon-

sistencies were identified in these examples. This review was intended to verify the face validity of the illustrative cases, rather than to provide a systematic reannotation or quantitative assessment of annotation quality.

3.5 Characteristics of XMedQA

Overall composition. The final XMedQA benchmark contains 13,812 unique questions, yielding 96,684 question–language instances across seven languages. **Fig. C.1** summarizes key benchmark characteristics, including answer-option counts, English question length, clinical-area composition, and subclinical-area distribution. Dataset-wise subclinical-area distributions are provided in **Fig. C.2**.

Regional specificity. We further characterized the regional specificity of XMedQA using a structured 0–4 scoring scheme. Region-specific items appeared across all source datasets, but their frequency and severity varied substantially (**Fig. 2**, left): the Japan-derived component and MedQA-ZH contained the largest numbers of items with region specificity scores of 1 or higher, whereas RuMedDaNet contained relatively few. The clinical-category distribution also varied by score (**Fig. 2**, middle). Low-score items were dominated by basic and clinical medicine, whereas the highest-score items included a larger proportion of social medicine, reflecting stronger dependence on healthcare systems, laws, public-health policy, and local clinical practice.

Country relevance. We also examined country relevance scores for region-specific items (**Fig. 2**, right). For questions with region specificity scores of 1 or higher, we identified the country or region with the highest relevance score, including the "Other" category. This analysis showed that each source dataset largely retained its expected country or regional context. The relatively large "Other" component in MedQA should not necessarily be interpreted as dataset noise, because many such items explicitly describe a patient's country of origin, immigration background, or travel-related context.

4 Experiments

4.1 Models

We evaluated a broad set of general-purpose and medical-domain LLMs. The general-purpose models include the DeepSeek-V3 series (DeepSeek-AI et al., 2025), Qwen2.5 and Qwen3 series (Qwen et al., 2025; Yang et al., 2025), Llama 3 and Llama 4 series (Grattafiori et al., 2024; Meta AI, 2025), GPT-OSS models (OpenAI et al., 2025), and Gemma 3 models (Team et al., 2025). We also evaluated MedGemma models as medical-domain adaptations of the Gemma family (Sellergren et al., 2026).

Additionally, we evaluated Japanese-adapted models to examine whether Japanese adaptation transfers beyond Japanese-only settings. Qwen3-Swallow and GPT-OSS-Swallow are general-domain models adapted to Japanese through Japanese-oriented continued pretraining and post-training (Fujii et al., 2024; Ma et al., 2025). We also evaluated Preferred-MedRECT-32B, a Japanese medical-domain model developed for clinical text correction and reasoning (Iwase et al., 2025).

All models were evaluated using the inference configuration described in Appendix E.

4.2 Evaluation Protocol

Aggregation over query languages. We evaluated model accuracy under three settings that differ in how query languages are aggregated. The *all-language setting* aggregates predictions across all seven query languages for the main multilingual and regionality analyses. The *English-only setting* evaluates all questions in English, providing a common-language view of regionality effects. The *original-source-language setting* evaluates each dataset in its original source language, capturing performance in the language originally associated with that dataset.

Model profiling by language robustness and region-specificity adaptability. To profile how models handle regional medical context and how this behavior changes across query languages, we summarized each model along two axes that quantify the lack of robustness: *cross-lingual variability* and *region-specificity sensitivity*. Lower values on these axes correspond to greater language robustness and greater adaptability to region-specific medical context.

Let $\mathcal{L} = \{\mathrm{en}, \mathrm{es}, \mathrm{fr}, \mathrm{ru}, \mathrm{sv}, \mathrm{ja}, \mathrm{zh}\}$ denote the set of seven query languages, and let $a_{m,\ell}$ be the accuracy of model m in language ℓ. We first computed the mean accuracy across languages,

$$\bar{a}_m = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} a_{m,\ell}.$$

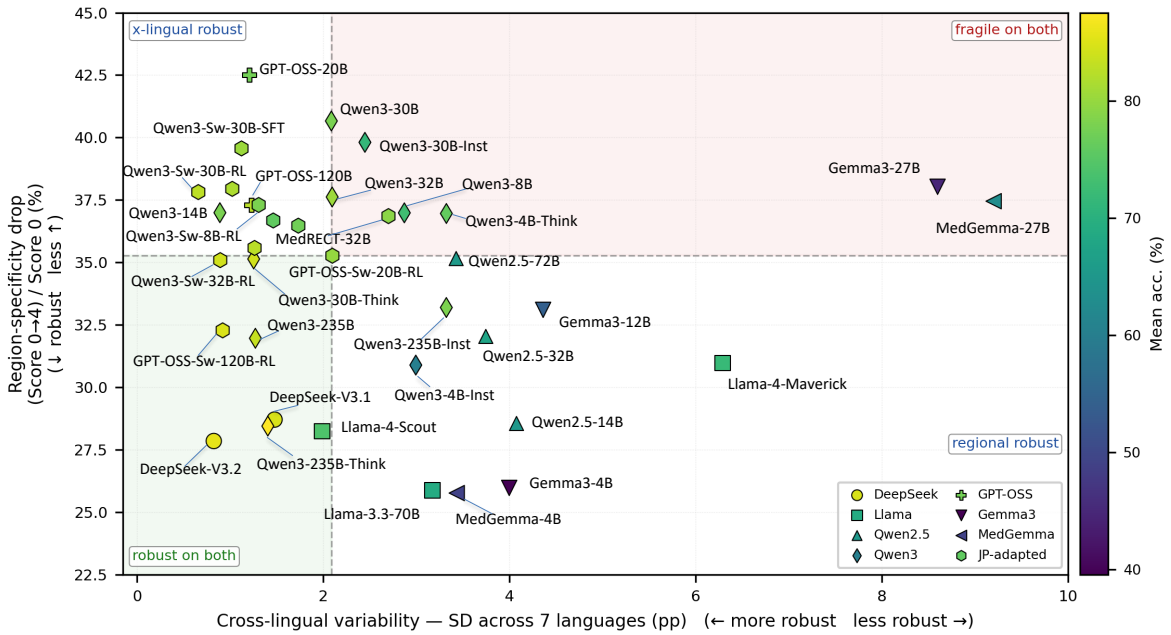

Figure 3: Model profiles by query-language robustness and region-specificity sensitivity. Each point represents one model. The horizontal axis shows cross-lingual variability across seven query languages, and the vertical axis shows the relative accuracy drop from region-specificity score 0 to 4. Lower values indicate greater robustness. Color indicates mean accuracy across all query languages.

We define cross-lingual variability as

$$X_m = \sqrt{\frac{1}{|\mathcal{L}|-1}\sum_{\ell\in\mathcal{L}}(a_{m,\ell}-\bar{a}_m)^2} \quad [\mathrm{pp}],$$

where lower X_m indicates greater language robustness, and pp denotes percentage points.

For region-specificity sensitivity, let $a_m^{(s)}$ be the all-language aggregated accuracy of model m for region-specificity score $s \in \{0,1,2,3,4\}$. We defined

$$Y_m = \frac{a_m^{(0)} - a_m^{(4)}}{a_m^{(0)}} \times 100 \quad [\%],$$

where lower Y_m indicates greater adaptability to increasing regional specificity.

We used (X_m, Y_m) to visualize model profiles, with color indicating the mean accuracy $\bar{a}_m$.

Regionality analysis. To examine how models handle region-specific medical knowledge, we stratified accuracy by region-specificity score under the all-language, English-only, and original-source-language settings. This analysis tests whether performance degradation with increasing regionality persists across different query-language settings. We additionally repeated the stratified analysis by clinical category to examine whether regionality effects differ across medical subdomains.

Japanese adaptation analysis. Finally, we analyzed Japanese-adapted models as a case study of whether language-specific or medical-domain adaptation reshapes these profiles and transfers beyond Japanese-only evaluation. We compared each adapted model with its corresponding base model and examined changes in the language-robustness and region-specificity profiles defined above. We also computed accuracy changes separately for high-region-specificity questions, defined as scores 2–4, and low-region-specificity questions, defined as scores 0–1. We further decomposed these changes into Japanese-language evaluation and the six non-Japanese query languages to assess whether Japanese adaptation improves only Japanese performance or also transfers to other languages.

5 Results

5.1 Model Profiles Vary Across Language Robustness and Regional Adaptability

We first profiled models along two axes: cross-lingual variability and region-specificity sensitivity. **Fig. 3** shows that models differ substantially not only in mean accuracy, but also in how robustly they handle query language and region-specific medical context.

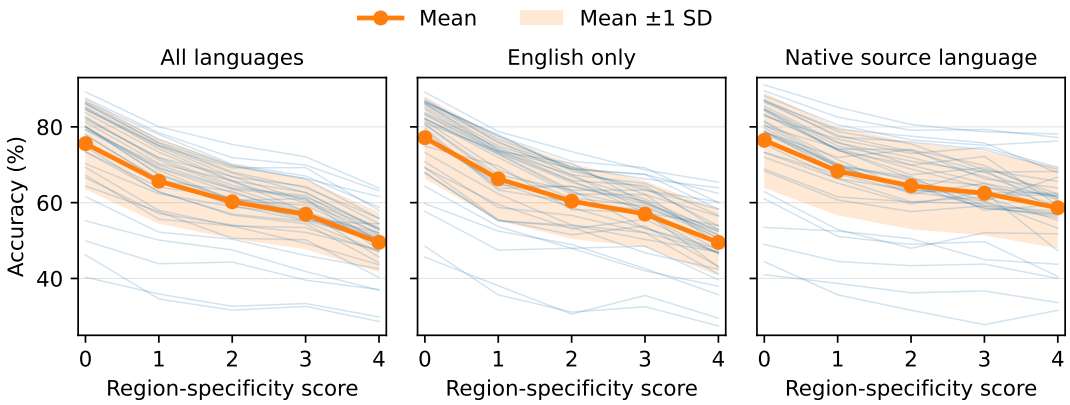

Figure 4: Accuracy by region-specificity score under three evaluation settings. Results are shown for all-language, English-only, and original-source-language evaluation. Accuracy decreases as region-specificity increases across all settings. Thin lines indicate individual models, thick lines indicate model averages, and shaded regions indicate one standard deviation.

Several high-performing models appear in the lower-left region, indicating both low cross-lingual variability and smaller accuracy drops on highly region-specific questions. In contrast, some models with high mean accuracy still show strong region-specificity sensitivity, suggesting that aggregate Medical QA performance does not necessarily imply adaptability to local medical context. Other models are positioned farther to the right, reflecting larger variation across query languages.

These profiles show that XMedQA captures distinct dimensions of medical LLM behavior that are obscured by aggregate accuracy alone. Models may be robust across languages but sensitive to regionality, or relatively stable across regionality but variable across query languages. This motivates the following analyses of regionality effects and language-specific adaptation.

Table F.1 summarizes the overall accuracy of representative models across the seven languages, with full model-wise results provided in **Table F.2**.

5.2 Regional Specificity Is an Independent Difficulty Factor

We next examined whether the performance drop associated with regional specificity can be explained by query-language effects. **Fig. 4** shows accuracy stratified by region-specificity score under all-language, English-only, and original-source-language evaluation.

Accuracy decreased as region-specificity increased in all three settings. Mean accuracy dropped from 75.6% to 49.5% in the all-language setting, from 77.1% to 49.5% in the English-only setting, and from 76.5% to 58.7% in the original-source-language setting. The comparable decline in English-only evaluation indicates that the difficulty cannot be attributed solely to non-English processing or translation artifacts, while the smaller decline in the original-source-language setting suggests that regional-language presentation partially mitigates, but does not eliminate, the difficulty. Full model-wise results are provided in **Table F.3**–**Table F.5**.

Controlling for clinical-category composition. Because high-region-specificity questions are more likely to involve social medicine, we further examined whether the trend simply reflects category composition. When stratified by main clinical category, accuracy still declined with increasing region-specificity across basic medicine, clinical medicine, and social medicine, including within clinical medicine. Thus, the effect is not explained solely by an overrepresentation of social-medicine or administrative questions. The decline was larger in clinical and social medicine than in basic medicine, and was partially mitigated by original-source-language evaluation. Full results are provided in **Fig. F.1**.

5.3 Cross-Lingual Gains from Japanese Adaptation in Region-Specific QA

Finally, we examined Japanese-adapted models as a case study of whether language-specific adaptation improves region-specific medical QA and transfers beyond Japanese. We focused on Japan-derived questions and compared performance on low-region-specificity questions (scores 0–1) and

Family	Size	Adaptation	Stage	Mean Δ		ja Δ		non-ja Δ		Improved non-ja	
				High	Low	High	Low	High	Low	High	Low
Qwen3	8B	Swallow	SFT	+2.44	+3.67	+4.20	+3.22	+2.15	+3.74	5/6	5/6
			RL	+3.80	+6.59	+5.05	+5.62	+3.59	+6.75	4/6	6/6
Qwen3	30B-A3B	Swallow	SFT	+2.92	+2.49	+4.21	+3.79	+2.71	+2.27	6/6	5/6
			RL	+4.92	+5.44	+3.93	+6.09	+5.09	+5.33	6/6	6/6
Qwen3	32B	Swallow	SFT	+3.72	+1.60	+4.49	+1.22	+3.59	+1.67	5/6	6/6
			RL	+7.68	+4.02	+11.21	+4.09	+7.09	+4.01	6/6	6/6
gpt-oss	20B	Swallow	SFT	+7.00	+0.11	+14.85	+2.00	+5.69	-0.21	5/6	2/6
			RL	+10.84	+2.11	+20.73	+5.98	+9.20	+1.46	6/6	5/6
gpt-oss	120B	Swallow	SFT	+4.01	+0.55	+6.72	+0.56	+3.55	+0.55	6/6	4/6
			RL	+8.93	+2.59	+14.28	+3.27	+8.04	+2.48	6/6	6/6
Qwen3	32B	MedRECT	SFT	+1.04	-0.98	+1.68	-2.36	+0.93	-0.74	4/6	3/6

Table 2: Japanese adaptation effects by region-specificity level. Evaluation is limited to Japan-derived questions in XMedQA, stratified by region-specificity level. Values indicate absolute percentage-point differences from the corresponding base model. High and Low denote region-specificity scores of 2–4 and 0–1, respectively; non-ja Δ denotes the mean change across the six non-Japanese query languages.

high-region-specificity questions (scores 2–4). **Table 2** summarizes the main results, with full language-wise results provided in **Table F.6**.

Japanese-adapted Swallow models generally improved performance on Japan-derived questions not only in Japanese but also in non-Japanese query languages. Most Swallow variants improved accuracy in five or six of the six non-Japanese languages, indicating that the gains were not limited to Japanese-language evaluation.

The gains were often larger for high-region-specificity questions than for low-region-specificity questions, especially in the GPT-OSS-Swallow models. For example, GPT-OSS-Swallow-120B-RL improved mean accuracy by +8.93 points on high-region-specificity questions and by +2.59 points on low-region-specificity questions. In contrast, Preferred-MedRECT-32B showed smaller and less consistent gains. Overall, this case study shows that Japanese adaptation can improve region-specific medical QA and transfer beyond Japanese, although the magnitude varies by model family, adaptation stage, and region-specificity level.

6 Conclusion

We introduced XMedQA, a multilingual and region-aware Medical QA benchmark for evaluating how LLMs handle query language and region-specific medical context. Across 37 LLMs, model profiles varied substantially in cross-lingual robustness and sensitivity to regional specificity, showing that aggregate Medical QA accuracy can obscure important differences in model behavior.

Our results show that regional specificity is a distinct source of difficulty. Accuracy consistently decreased as region-specificity increased, including under English-only and original-source-language evaluation, indicating that the effect cannot be explained solely by non-English processing or translation artifacts. This decline also persisted within clinical-category strata, including clinical medicine, suggesting that it was not merely driven by the overrepresentation of social-medicine questions among high-region-specificity items. Rather, clinically oriented questions may also depend on local epidemiology, screening practices, guideline-preferred management, or drug availability.

The Japanese-adaptation case study further showed that adaptation can improve performance on Japan-derived region-specific questions and transfer to non-Japanese query languages. This suggests that adaptation may reshape access to locally grounded medical knowledge, although such transfer varied by model family and adaptation stage.

These findings highlight the need for region-aware evaluation of medical LLMs. By preserving regional medical content rather than normalizing it away, XMedQA enables profiling across accuracy, language robustness, and sensitivity to locally grounded medical knowledge. More broadly, our results suggest that reliable medical LLM evaluation should distinguish globally shared biomedical knowledge from language-mediated access and country- or region-specific clinical context.

Limitations

This study has several limitations. First, XMedQA is an evaluation benchmark and does not provide training data for region-aware medical model development. Second, XMedQA covers seven languages and source regions, and therefore cannot represent the full diversity of healthcare systems, epidemiology, guidelines, and examination cultures worldwide. Third, translations were produced with a single machine translation model, Gemini 3.1 Flash-Lite Preview, without back-translation or professional medical-translator verification. Residual translation artifacts may therefore affect cross-lingual comparisons. Finally, clinical-category and regionality annotations were generated using a single LLM, Gemini 3.1 Pro Preview. Although a physician reviewed sampled items for plausibility, we did not conduct large-scale manual reannotation or measure inter-annotator agreement.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Pavel Blinov, Arina Reshetnikova, Aleksandr Nesterov, Galina Zubkova, and Vladimir Kokh. 2022. *RuMedBench: A Russian Medical Language Understanding Benchmark*, page 383–392. Springer International Publishing.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Fan Gao, Sherry T Tong, Jiwoong Sohn, Jiahao Huang, Junfeng Jiang, Ding Xia, Piyalitt Ittichaiwong, Kanyakorn Veerakanjana, Hyunjae Kim, Qingyu Chen, and 1 others. 2026. Med-coreasoner: Reducing language disparities in medical reasoning via language-informed co-reasoning. *arXiv preprint arXiv:2601.08267*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Niclas Hertzberg and Anna Lokrantz. 2024. Medqa-swe-a clinical question & answer dataset for swedish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186.

Naoto Iwase, Hiroki Okuyama, and Junichiro Iwasawa. 2025. Medrect: A medical reasoning benchmark for error correction in clinical texts. *Preprint*, arXiv:2511.00421.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific data*, 10(1):170.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard

Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. 2025. Building instruction-tuning datasets from human-written instructions with open-weight large language models. In *Second Conference on Language Modeling*.

João Matos, Shan Chen, Siena Kathleen V Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis Filipe Nakayama, José María Millet Pascual-Leone, and 1 others. 2025. Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7203–7216.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2026-05-20.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, Robyn Gayle Dychiao, Jose Carlo Artiaga, and 1 others. 2025. Multi-ophthalingua: a multilingual benchmark for assessing and debiasing llm ophthalmological qa in lmics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28321–28330.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Jiandong Shao, Raphael Tang, Crystina Zhang, Karin Sevegnani, Pontus Stenetorp, Jianfei Yang, and Yao Lu. 2026. The role of mixed-language documents for multilingual large language model pretraining. *Preprint*, arXiv:2601.00364.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev,

and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966.

Hannah Waight, Eddie Yang, Yin Yuan, Solomon Messing, Margaret E. Roberts, Brandon M. Stewart, and Joshua A. Tucker. 2026. State media control influences large language models. *Nature*.

Jinqiang Wang, Huansheng Ning, Yi Peng, Qikai Wei, Daniel Tesfai, Wenwei Mao, Tao Zhu, and Runhe Huang. 2026. A survey on large language models from general purpose to medical applications: Datasets, methodologies, and evaluations. *Neurocomputing*, page 132914.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 others. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885.

A Japanese-Derived Benchmark Construction

We constructed **JMedQA** as a recompiled Japanese medical QA benchmark based on the Japanese National Medical Licensing Examination, rather than directly using IgakuQA (Kasai et al., 2023). Specifically, we collected examination questions published between 2018 and 2025 from the official examination materials distributed as PDF files by the Ministry of Health, Labour and Welfare in Japan. Because these PDFs were not directly available as a structured benchmark, we used an OCR-assisted extraction pipeline with Gemini 3.0 Pro Preview to extract the question stem, answer options, answer key, year, and section information.

After extraction, we applied rule-based post-processing to clean OCR artifacts, normalize option labels, and standardize answer representations. The records were then converted into the unified XMedQA schema. Specifically, each item was assigned a source dataset identifier, source region, source language, normalized question text, normalized answer options, answer label, answer count, case-based question flag, linked-question flag, calculation flag, clinical-area label, year, section, and unique problem identifier. Additional non-standard fields were preserved as metadata to maintain traceability.

Because the original national examination includes questions that depend on figures, photographs, radiographs, pathological images, or other visual materials, we excluded image-dependent questions from the final JMedQA benchmark. This exclusion was applied to ensure that JMedQA evaluates text-based medical QA rather than multimodal image interpretation.

B Robustness of Region-Specificity Scoring

To assess the robustness of the LLM-based region-specificity annotations, we compared the Gemini-3.1-Pro-Preview scores used in the main analysis with scores independently assigned by GPT-5.5. Because the original score distribution was highly imbalanced, with most items assigned a score of 0, we used a stratified random subset rather than sampling from the natural distribution. Specifically, we sampled 150 items from each Gemini-assigned region-specificity score category, yielding 750 items in total. This design ensured adequate coverage across the full 0–4 ordinal scale.

The two models showed substantial ordinal agreement, with a quadratic weighted kappa of 0.775. The exact-match rate was 56.4%, while 89.3% of scores differed by no more than one point, and the mean absolute error was 0.58 on the 0–4 scale. Agreement was strongest at the lowest and highest score categories, with exact agreement of 92.7% for Gemini score 0 and 77.3% for Gemini score 4, whereas intermediate scores showed lower exact agreement. These results indicate that disagreements were concentrated in ambiguous intermediate categories rather than between clearly region-independent and strongly region-specific

items.

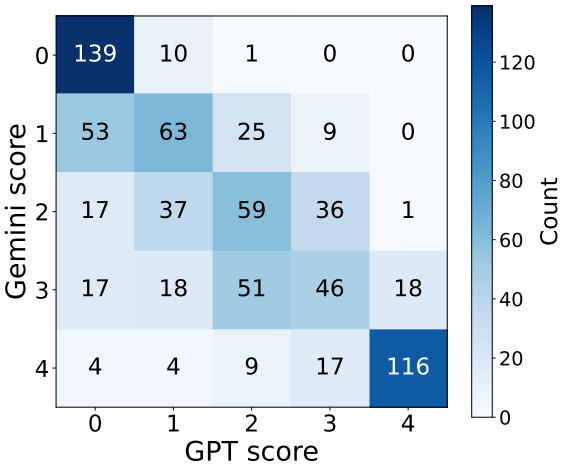

Figure B.1: Confusion matrix of region-specificity scores assigned by Gemini-3.1-Pro-Preview and GPT-5.5. Rows indicate Gemini-assigned scores used as the reference, and columns indicate GPT-5.5 scores. The subset was stratified by Gemini score, with 150 items sampled from each score category. Agreement was strongest at the extreme categories, whereas disagreements were concentrated among adjacent and intermediate scores.

C Detailed Characteristics of XMedQA

This appendix provides additional descriptive analyses of XMedQA beyond the summary presented in the main text. We first summarize the overall benchmark composition, including dataset size, answer-option counts, question length, and clinical-area coverage. We then provide finer-grained subclinical-area distributions and additional analyses of region-specificity types across clinical categories and source datasets.

Fig. C.1 provides an overview of the constructed benchmark. Because XMedQA integrates heterogeneous exam-style datasets, these statistics help characterize differences in question format, length, and clinical coverage before multilingual expansion.

Fig. C.2 further breaks down the clinical content of each source dataset. This analysis shows that the source datasets differ not only in language and country context, but also in their clinical topic composition.

Fig. C.3 characterizes how different types of regional dependence are distributed across clinical categories. This complements the main-text analysis by showing which kinds of regional factors are associated with basic medicine, clinical medicine, and social medicine.

Fig. C.4 shows how region-specificity types vary across source datasets. This analysis helps distinguish datasets that contain region-specific items mainly because of healthcare systems, local clinical practice, epidemiology, public-health policy, or other forms of local medical context.

D Annotation and Translation Prompts

This appendix summarizes the prompts used for multilingual translation, clinical-category annotation, and region-specificity scoring. For readability, we report the core instructions and output schemas rather than the full API wrapper.

D.1 Multilingual Translation Prompt

For each source item, we translated the question and answer options directly into each target language without using English as a pivot. The translation prompt is shown in **Fig. D.1**. It was designed to preserve the medical content and answer-option structure while returning a structured JSON output.

D.2 Clinical Category Classification Prompt

We assigned each item to a hierarchical clinical category using the English question and options as primary evidence. The clinical-category classification prompt is shown in **Fig. D.2**. The taxonomy consisted of three main categories: `basic_medicine`, `clinical_medicine`, and `social_medicine`, each with predefined subcategories.

D.3 Region-Specificity Scoring Prompt

We used an LLM-as-a-judge prompt to assess whether each question depended on country- or region-specific medical context. The region-specificity scoring prompt is shown in **Fig. D.3**. The primary score used in the main analysis was the `region_specificity_score`, while additional fields were collected to support interpretability and quality control.

E Inference Configuration

We used vLLM-based single-pass generation for all models. Unless otherwise noted, decoding parameters were fixed across models. For models that supported configurable thinking mode, thinking was enabled. Inference was run on a single compute node equipped with eight NVIDIA H200 GPUs.

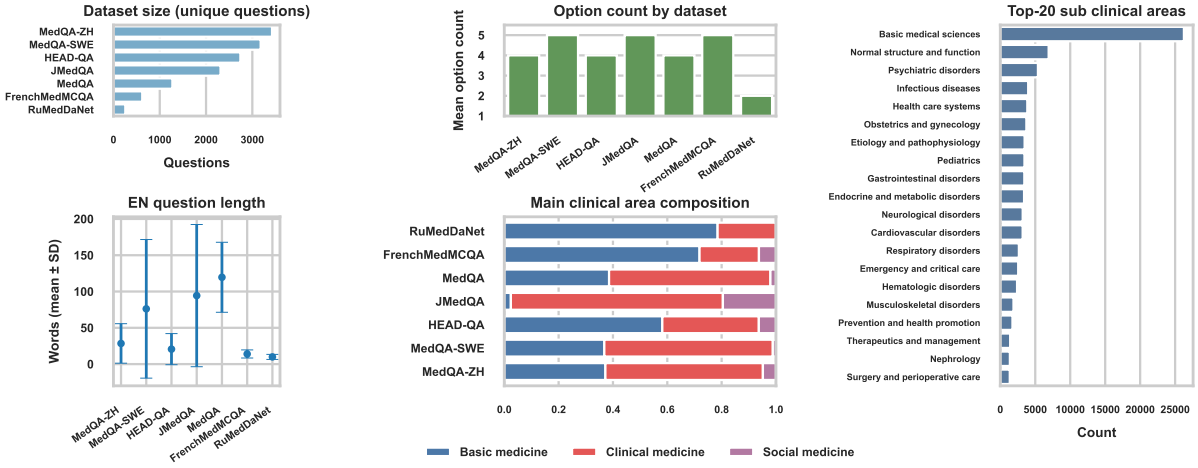

Figure C.1: Detailed characteristics of XMedQA. The figure summarizes dataset size, mean number of answer options, English question length, main clinical-area composition, and the distribution of the top 20 subclinical areas. Dataset size is shown as the number of unique questions before multilingual expansion. Question length is reported as mean ± standard deviation after English normalization. Clinical-area composition was calculated from standardized category mappings across source datasets.

Item	Value
Decoding temperature	0.2
Top-p	0.95
Top-k	-1
Max generation tokens	8192
Thinking mode	Enabled when configurable; otherwise default model behavior

Table E.1: Inference configuration used in the experiments.

F Additional Evaluation Results

This appendix provides additional model-wise results supporting the analyses in the main text. We first report overall accuracy by model and query language, followed by region-specificity-stratified results under the three evaluation settings. We then provide detailed language-wise results for the Japanese-adaptation case study.

F.1 Overall Accuracy by Query Language

Table F.1 reports the overall accuracy of representative models across the seven query languages. This table supports the model-profile analysis in **Section 5.1** by showing the language-wise performance patterns underlying the cross-lingual variability metric.

Table F.2 provides the full model-wise version of the same analysis for all evaluated models.

F.2 Accuracy by Region-Specificity Score

Table F.3 reports model-wise accuracy stratified by region-specificity score under the all-language setting. This table supports the regionality analysis in **Section 5.2**.

Table F.4 reports the corresponding results when all questions are evaluated in English. This setting controls for query language and is used to test whether region-specificity effects remain under a common evaluation language.

Table F.5 reports the corresponding results under original-source-language evaluation. This setting evaluates each source dataset in its original language.

Fig. F.1 further stratifies the region-specificity analysis by clinical category. This analysis supports the clinical-category control in **Section 5.2**, showing that the accuracy decline with increasing region-specificity is not explained solely by the overrepresentation of social-medicine or administrative questions.

F.3 Detailed Results for Japanese Adaptation

Table F.6 provides the full language-wise results for the Japanese-adaptation case study in **Section 5.3**. The table reports accuracy changes from the corresponding base model on Japan-derived XMedQA questions, separately for high- and low-region-specificity subsets and for each query language.

Model	en	es	fr	ru	sv	ja	zh	Mean	Range	SD
DeepSeek-V3.1	84.9	85.2	84.2	81.6	85.7	**86.2**	84.5	84.6	4.5	1.5
DeepSeek-V3.2	85.6	86.5	85.1	84.4	86.5	**86.6**	85.9	85.8	2.1	0.8
Llama-3.3-70B-Instruct	67.5	70.2	65.1	68.5	65.9	71.5	**74.0**	69.0	9.0	3.2
Llama-4-Scout-17B-16E-Instruct	71.3	74.7	75.1	71.8	**76.5**	75.9	73.9	74.2	5.2	2.0
Llama-4-Maverick-17B-128E-Instruct	66.0	74.7	66.1	65.9	70.8	79.6	**80.0**	71.9	14.1	6.3
Qwen3-4B-Thinking-2507	**78.4**	77.2	77.4	68.6	75.9	77.4	77.0	76.0	9.7	3.3
Qwen3-30B-A3B-Thinking-2507	84.7	84.0	83.6	82.2	84.1	**86.2**	83.3	84.0	4.1	1.2
Qwen3-235B-A22B-Thinking-2507	87.4	87.2	86.7	85.5	87.1	**89.4**	89.3	87.5	3.9	1.4
gpt-oss-20b	**79.1**	77.4	78.1	75.3	76.5	77.3	76.9	77.2	3.8	1.2
gpt-oss-120b	**84.3**	83.0	82.7	80.5	81.9	83.5	82.2	82.6	3.8	1.2
gemma-3-4b-it	**44.4**	36.3	40.2	32.4	42.0	41.2	40.5	39.6	12.0	4.0
gemma-3-27b-it	46.8	**55.9**	46.9	37.6	50.1	46.4	29.6	44.8	26.3	8.6
medgemma-4b-it	**56.2**	49.2	49.7	46.8	47.4	47.5	46.2	49.0	10.1	3.4
medgemma-27b-it	**73.0**	71.1	67.4	65.1	61.6	53.7	47.9	62.8	25.2	9.2

Table F.1: Overall accuracy of representative models by language. Mean, range, and standard deviation are computed across the seven languages. For each model, the highest language-specific accuracy is shown in bold, and the lowest is underlined.

G Qualitative Examples of Region-Specificity Scores

This appendix provides qualitative examples for each region-specificity score level (0–4), with five samples per score. For each sample, we show: (i) source dataset and language, (ii) region-specificity score, (iii) predominant country and country-level relevance scores, (iv) clinical main/sub category, (v) question and options, (vi) gold answer, and (vii) the rationale used for score assignment. The examples are intended to illustrate how the scoring rubric is applied in practice, from globally transferable medical knowledge (lower scores) to items that require country-specific legal, policy, or healthcare-system knowledge (higher scores).

Model	en	fr	es	sv	ru	ja	zh	Mean	Range	SD
DeepSeek-V3.1	84.9	84.2	85.2	85.7	81.6	86.2	84.5	84.6	4.5	1.5
DeepSeek-V3.2	85.6	85.1	86.5	86.5	84.4	86.6	85.9	85.8	2.1	0.8
Llama-3.3-70B-Instruct	67.5	65.1	70.2	65.9	68.5	71.5	74.0	69.0	9.0	3.2
Llama-4-Scout-17B-16E-Instruct	71.3	75.1	74.7	76.5	71.8	75.9	73.9	74.2	5.2	2.0
Llama-4-Maverick-17B-128E-Instruct	66.0	66.1	74.7	70.8	65.9	79.6	80.0	71.9	14.1	6.3
Qwen2.5-14B-Instruct	69.0	61.5	63.4	62.0	64.6	63.1	72.5	65.2	11.0	4.1
Qwen2.5-32B-Instruct	67.4	67.5	66.0	67.1	61.3	71.0	73.1	67.6	11.8	3.8
Qwen2.5-72B-Instruct	66.0	64.5	63.3	65.4	62.0	72.7	64.7	65.5	10.7	3.4
Qwen3-4B-Instruct-2507	62.5	60.2	57.3	56.9	58.1	60.9	65.1	60.2	8.2	3.0
Qwen3-4B-Thinking-2507	78.4	77.4	77.2	75.9	68.6	77.4	77.0	76.0	9.7	3.3
Qwen3-8B	76.7	73.7	70.4	73.7	69.0	75.5	70.5	72.8	7.7	2.9
Qwen3-14B	78.3	77.5	77.8	76.9	77.7	79.6	77.1	77.8	2.7	0.9
Qwen3-30B-A3B	79.9	78.3	78.2	78.6	73.4	79.1	77.8	77.9	6.5	2.1
Qwen3-30B-A3B-Instruct-2507	74.6	71.6	70.8	71.4	70.7	73.7	67.0	71.4	7.6	2.5
Qwen3-30B-A3B-Thinking-2507	84.7	83.6	84.0	84.1	82.2	86.2	83.3	84.0	4.1	1.2
Qwen3-32B	81.7	78.9	79.6	79.7	77.4	82.4	83.3	80.4	5.9	2.1
Qwen3-235B-A22B	83.9	83.5	82.6	84.0	80.9	84.8	83.8	83.3	3.9	1.3
Qwen3-235B-A22B-Instruct-2507	80.5	74.4	77.7	78.2	73.2	82.5	79.9	78.0	9.3	3.3
Qwen3-235B-A22B-Thinking-2507	87.4	86.7	87.2	87.1	85.5	89.4	89.3	87.5	3.9	1.4
Qwen3-Swallow-8B-SFT-v0.2	77.4	75.8	75.6	73.5	74.2	76.5	77.1	75.7	4.0	1.5
Qwen3-Swallow-8B-RL-v0.2	79.4	77.8	77.6	75.9	76.8	78.5	79.4	77.9	3.5	1.3
Qwen3-Swallow-30B-A3B-SFT-v0.2	81.0	80.0	80.6	78.5	79.1	81.8	80.7	80.2	3.3	1.1
Qwen3-Swallow-30B-A3B-RL-v0.2	83.0	82.8	83.0	82.1	81.7	83.6	83.2	82.8	1.9	0.7
Qwen3-Swallow-32B-SFT-v0.2	82.4	81.2	81.7	80.0	80.6	82.3	82.8	81.6	2.8	1.0
Qwen3-Swallow-32B-RL-v0.2	84.1	83.8	84.3	82.9	83.1	84.7	85.5	84.1	2.6	0.9
gpt-oss-20b	79.1	78.1	77.4	76.5	75.3	77.3	76.9	77.2	3.8	1.2
gpt-oss-120b	84.3	82.7	83.0	81.9	80.5	83.5	82.2	82.6	3.8	1.2
GPT-OSS-Swallow-20B-SFT-v0.1	80.1	77.5	77.7	75.7	75.2	78.4	75.9	77.2	4.9	1.7
GPT-OSS-Swallow-20B-RL-v0.1	75.9	81.2	81.2	80.3	78.8	82.3	80.1	80.0	6.4	2.1
GPT-OSS-Swallow-120B-SFT-v0.1	84.2	83.1	82.9	81.7	80.5	83.5	81.6	82.5	3.6	1.3
GPT-OSS-Swallow-120B-RL-v0.1	85.0	85.2	85.5	84.6	83.3	86.2	84.3	84.9	2.9	0.9
gemma-3-4b-it	44.4	40.2	36.3	42.0	32.4	41.2	40.5	39.6	12.0	4.0
gemma-3-12b-it	58.2	56.8	52.1	55.2	48.0	58.9	49.1	54.0	10.9	4.4
gemma-3-27b-it	46.8	46.9	55.9	50.1	37.6	46.4	29.6	44.8	26.3	8.6
medgemma-4b-it	56.2	49.7	49.2	47.4	46.8	47.5	46.2	49.0	10.1	3.4
medgemma-27b-it	73.0	67.4	71.1	61.6	65.1	53.7	47.9	62.8	25.2	9.2
Preferred-MedRECT-32B	81.4	75.0	78.8	79.3	77.5	78.9	83.5	79.2	8.4	2.7

Table F.2: Overall accuracy by model and language. Mean, range, and standard deviation are computed across the seven languages.

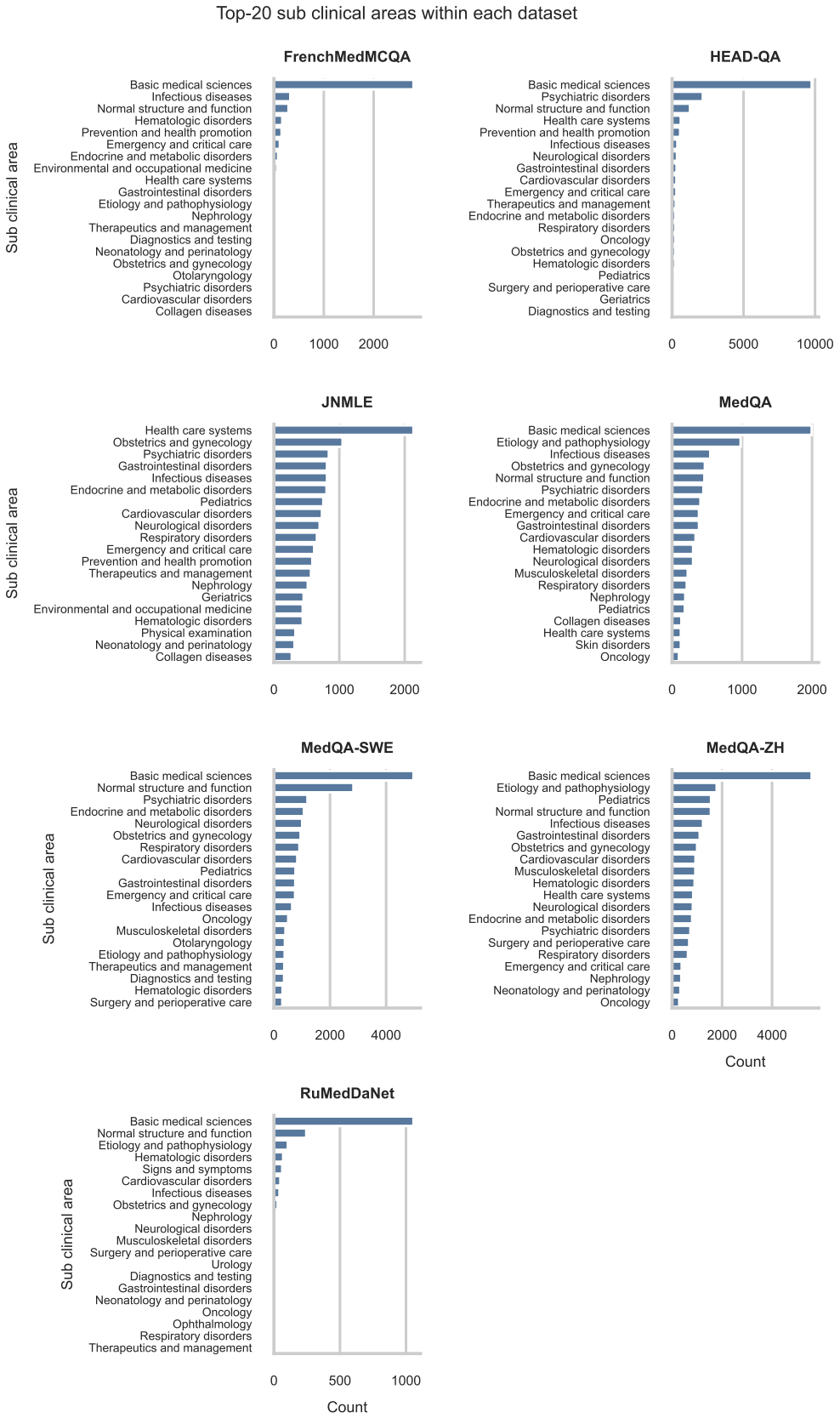

Figure C.2: Top 20 subclinical areas within each source dataset. Subclinical-area labels were derived from source-dataset-specific category annotations and mapped to a standardized vocabulary. Counts represent question-level annotations within each source dataset before multilingual translation.

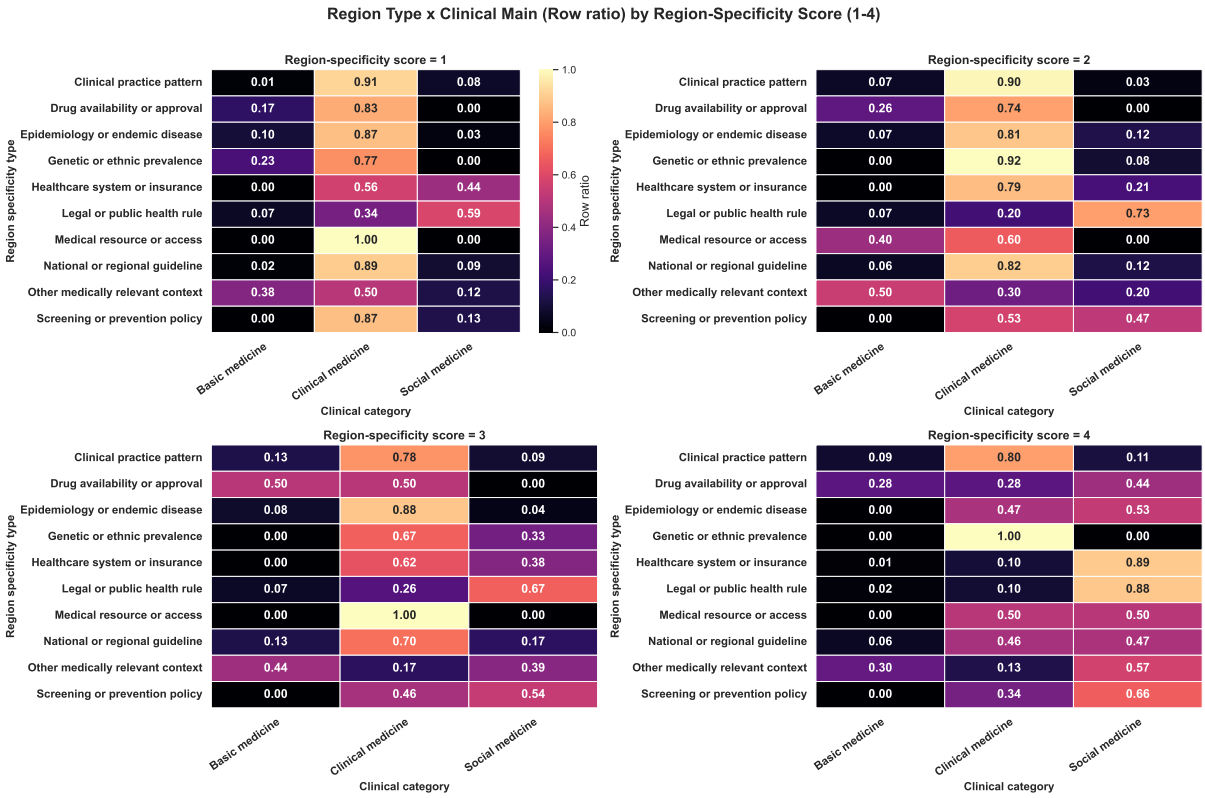

Figure C.3: Distribution of region-specificity types by main clinical category. Heatmaps show row-normalized proportions for each region-specificity type across the three main clinical categories, stratified by region-specificity score from 1 to 4. Each row sums to 1 within a given region-specificity type and score.

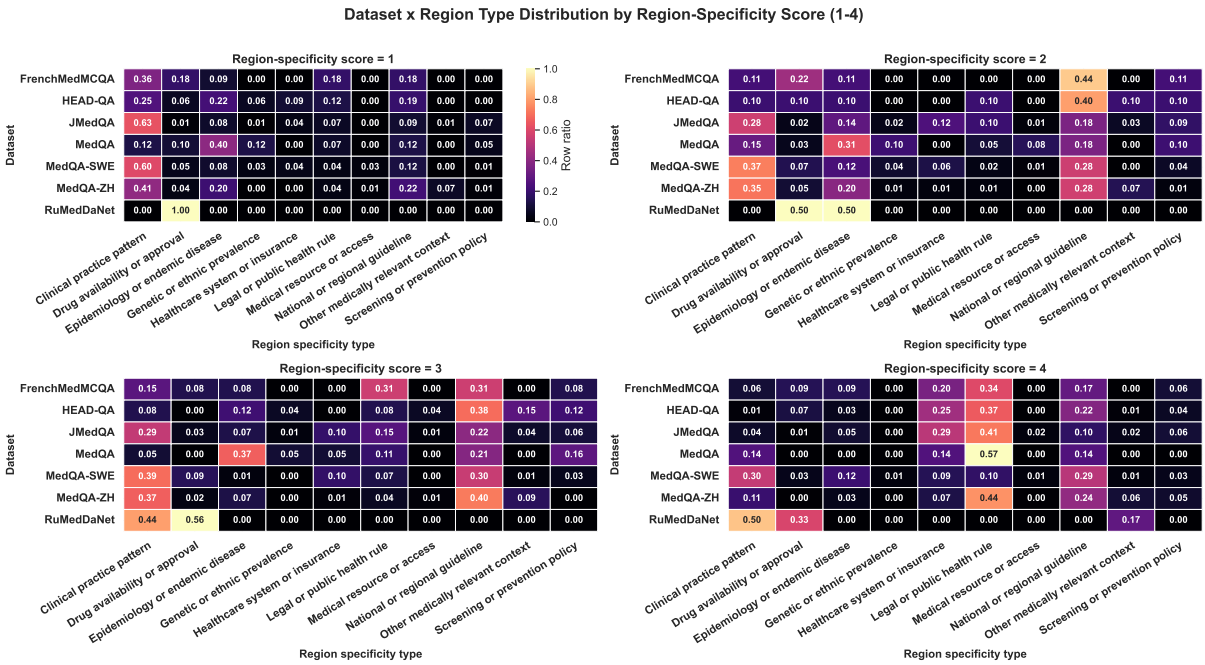

Figure C.4: Distribution of region-specificity types by source dataset. Heatmaps show row-normalized proportions of region-specificity types within each source dataset, stratified by region-specificity score from 1 to 4. Each row sums to 1 within a given dataset and score.

Translation prompt

Role. You are a specialized translation system for medical exam multiple-choice questions.

Task. Translate from source language `{source_language}` to target language `{target_language}`.

Input fields.

- `problem_unique_id`
- `question`
- `options`
- `clinical_area_hint`

Hard requirements.

- Preserve medical meaning exactly, including diagnoses, tests, drugs, numbers, units, age/sex/time, and clinical context.
- Preserve option keys and option order exactly as provided.
- Do not change the answer-key mapping by reordering or relabeling options.
- Choose exactly one clinical area from the allowed list.
- Output JSON only, following the predefined response schema.

Expected JSON output.

```
{
  "question": "translated question",
  "options": {
    "A": "translated option A",
    "B": "translated option B",
    ...
  },
  "clinical_area": "one label from the predefined list"
}
```

Figure D.1: Prompt used for multilingual translation. The prompt preserves medical meaning, answer-option order, and structured output format.

Clinical category classification prompt

Role. You are a medical exam taxonomy classifier.

Task. Classify each item using only the predefined taxonomy.

Evidence. Use the English question and options as the primary evidence.

Main categories.

- `basic_medicine`
- `clinical_medicine`
- `social_medicine`

Hard requirements.

- Choose exactly one `main_category` and one `sub_category`.
- The selected `sub_category` must belong to the selected `main_category`.
- Use only labels from the predefined taxonomy.
- Output JSON only and follow the predefined response schema exactly.

Expected JSON output.

```
{
  "main_category": "one label from the main-category list",
  "sub_category": "one label from the predefined subcategory list"
}
```

Figure D.2: Prompt used for clinical-category classification. The full subcategory taxonomy was provided to the model in the actual prompt.

Region-specificity scoring prompt

Role. You are an expert evaluator of medical multiple-choice questions.

Task. Assess whether a medical question depends on country-specific or region-specific medical context.

Input. You will be given a medical multiple-choice question translated into English, together with answer options. When available, the correct answer and explanation may also be provided.

Primary scoring rubric.

- **0**: Universal medical knowledge with no meaningful country or regional dependence.
- **1**: Mostly universal, with only weak or indirect regional relevance.
- **2**: Mixed, with some advantage from knowing regional epidemiology, practice patterns, guidelines, or healthcare context.
- **3**: Region-specific, where answering likely benefits from knowledge of a specific broad region or country.
- **4**: Strongly region-specific, where the question depends heavily on country- or region-specific knowledge.

Important principles.

- Evaluate the medical reasoning required to answer the question.
- Do not infer the source country from dataset name, source language, formatting, writing style, personal names, or translation artifacts.
- Do not force the question into a single country; multiple broad regions or countries may be relevant.
- If the question is based on universal biomedical knowledge, assign high global relevance and low country/region relevance.
- Personal names, place names, wording, and question format may be considered only if they directly affect the medical knowledge required to answer the question.
- Return only valid JSON. Do not include markdown or explanations outside the JSON.

Expected JSON output.

```
{
  "region_specificity_score": 0-4,
  "medical_universality_score": 0-4,
  "regional_context_level": "none | broad_region | country_specific | multi_region | unclear",
  "broad_region_relevance_scores": {
    "Global": 0-4,
    "East Asia": 0-4,
    "Southeast Asia": 0-4,
    "South Asia": 0-4,
    "Central Asia": 0-4,
    "Middle East and North Africa": 0-4,
    "Sub-Saharan Africa": 0-4,
    "Europe": 0-4,
    "North America": 0-4,
    "Latin America and Caribbean": 0-4,
    "Oceania": 0-4
  },
  "country_relevance_scores": {
    "Japan": 0-4,
    "United States": 0-4,
    "China": 0-4,
    "France": 0-4,
    "Spain": 0-4,
    "Sweden": 0-4,
    "Russia": 0-4,
    "Other": 0-4
  },
  "other_country_or_region_free_text": [],
  "region_specificity_types": [],
  "non_medical_artifact_types": [],
  "exam_artifact_score": 0-4,
  "region_confidence": 0-4,
  "reason_brief": "brief medically relevant rationale"
}
```

Figure D.3: Prompt used for region-specificity scoring. The main analysis used `region_specificity_score`; the remaining fields were collected for interpretability and quality control.

Model	Score 0	Score 1	Score 2	Score 3	Score 4
DeepSeek-V3.2	**87.5**	78.4	71.9	69.9	63.1
DeepSeek-V3.1	**86.4**	76.7	69.8	66.4	61.6
Llama-4-Scout-17B-16E-Instruct	**75.8**	66.1	59.2	55.2	54.4
Llama-4-Maverick-17B-128E-Instruct	**73.5**	64.1	61.3	57.3	50.7
Llama-3.3-70B-Instruct	**70.3**	62.4	59.0	56.7	52.1
Qwen2.5-32B-Instruct	**69.3**	57.9	53.9	53.4	47.1
Qwen2.5-72B-Instruct	**67.2**	55.6	53.9	49.3	43.6
Qwen2.5-14B-Instruct	**66.6**	57.1	54.1	52.2	47.6
Qwen3-235B-A22B-Thinking-2507	**89.2**	80.0	75.4	72.1	63.9
Qwen3-30B-A3B-Thinking-2507	**86.0**	75.9	69.6	66.3	55.8
Qwen3-235B-A22B	**85.2**	75.1	69.2	69.2	58.0
Qwen3-32B	**82.6**	71.3	63.9	61.9	51.5
Qwen3-30B-A3B	**80.1**	68.6	60.6	61.0	47.5
Qwen3-14B	**80.0**	67.3	60.7	59.3	50.4
Qwen3-235B-A22B-Instruct-2507	**79.9**	69.2	63.6	61.5	53.4
Qwen3-4B-Thinking-2507	**78.0**	65.9	63.0	56.0	49.2
Qwen3-8B	**74.8**	62.8	57.0	55.0	47.1
Qwen3-30B-A3B-Instruct-2507	**73.4**	62.9	56.7	54.4	44.2
Qwen3-4B-Instruct-2507	**61.6**	52.2	50.4	46.1	42.6
Qwen3-Swallow-32B-RL-v0.2	**86.2**	75.7	68.1	63.9	55.9
Qwen3-Swallow-30B-A3B-RL-v0.2	**84.9**	74.2	65.8	62.7	52.8
Qwen3-Swallow-32B-SFT-v0.2	**83.7**	73.3	65.5	59.9	52.0
Qwen3-Swallow-30B-A3B-SFT-v0.2	**82.4**	71.8	64.6	61.0	49.8
Qwen3-Swallow-8B-RL-v0.2	**80.1**	67.5	61.1	55.3	50.2
Qwen3-Swallow-8B-SFT-v0.2	**77.8**	66.2	60.2	56.0	49.2
GPT-OSS-Swallow-120B-RL-v0.1	**86.8**	78.0	70.1	67.0	58.8
GPT-OSS-Swallow-120B-SFT-v0.1	**84.5**	74.8	66.8	64.3	54.5
GPT-OSS-Swallow-20B-RL-v0.1	**81.9**	71.9	66.9	60.4	53.0
GPT-OSS-Swallow-20B-SFT-v0.1	**79.2**	69.1	62.6	57.3	50.3
gpt-oss-120b	**84.7**	75.2	67.8	62.6	53.1
gpt-oss-20b	**79.3**	69.9	63.7	57.7	45.6
gemma-3-12b-it	**55.2**	50.1	47.5	41.8	36.9
gemma-3-27b-it	**46.2**	34.6	31.6	32.6	28.6
gemma-3-4b-it	**40.3**	35.9	32.6	33.4	29.9
medgemma-27b-it	**64.4**	57.6	50.6	50.2	40.3
medgemma-4b-it	**49.9**	43.9	44.3	39.6	37.1
Preferred-MedRECT-32B	**81.3**	70.3	65.8	57.6	51.3

Table F.3: Accuracy by region-specificity score under all-language aggregation. For each model, the highest score-specific accuracy is shown in bold, and the lowest is underlined.

Model	Score 0	Score 1	Score 2	Score 3	Score 4
DeepSeek-V3.2	**87.4**	75.8	69.4	66.3	63.9
DeepSeek-V3.1	**86.7**	77.6	68.9	67.5	62.4
Llama-4-Scout-17B-16E-Instruct	**73.2**	62.3	54.1	50.3	46.9
Llama-3.3-70B-Instruct	**69.2**	59.5	53.1	55.6	46.5
Llama-4-Maverick-17B-128E-Instruct	**67.9**	55.1	54.1	52.1	41.1
Qwen2.5-14B-Instruct	**70.8**	58.7	56.6	52.1	46.9
Qwen2.5-32B-Instruct	**69.3**	55.5	51.5	53.9	43.2
Qwen2.5-72B-Instruct	**67.9**	55.3	53.6	46.8	42.4
Qwen3-235B-A22B-Thinking-2507	**89.2**	78.8	73.5	68.6	65.4
Qwen3-30B-A3B-Thinking-2507	**86.7**	76.8	69.4	62.1	56.5
Qwen3-235B-A22B	**85.8**	75.0	68.9	67.5	58.3
Qwen3-32B	**83.7**	74.5	66.8	61.0	52.3
Qwen3-235B-A22B-Instruct-2507	**82.2**	73.0	63.3	63.3	60.2
Qwen3-30B-A3B	**82.1**	71.9	62.2	62.1	49.2
Qwen3-4B-Thinking-2507	**80.7**	67.3	63.3	53.2	47.9
Qwen3-14B	**80.6**	66.5	60.2	58.0	48.8
Qwen3-8B	**79.0**	63.3	60.2	56.8	46.5
Qwen3-30B-A3B-Instruct-2507	**76.7**	65.3	58.7	53.2	46.9
Qwen3-4B-Instruct-2507	**64.3**	53.7	48.0	48.5	39.6
Qwen3-Swallow-32B-RL-v0.2	**86.1**	74.0	68.9	63.3	58.3
Qwen3-Swallow-30B-A3B-RL-v0.2	**85.2**	74.0	65.8	64.5	52.9
Qwen3-Swallow-32B-SFT-v0.2	**84.6**	73.3	65.8	57.4	54.0
Qwen3-Swallow-30B-A3B-SFT-v0.2	**83.1**	73.2	65.3	60.4	51.4
Qwen3-Swallow-8B-RL-v0.2	**81.6**	68.7	61.2	59.2	50.6
Qwen3-Swallow-8B-SFT-v0.2	**79.6**	63.7	64.3	60.4	50.6
GPT-OSS-Swallow-120B-RL-v0.1	**86.8**	77.6	70.9	69.2	60.0
GPT-OSS-Swallow-120B-SFT-v0.1	**86.2**	76.2	67.3	65.1	56.6
GPT-OSS-Swallow-20B-SFT-v0.1	**82.1**	72.8	62.8	58.6	52.5
GPT-OSS-Swallow-20B-RL-v0.1	**77.8**	64.5	64.3	59.8	51.6
gpt-oss-120b	**86.4**	78.0	68.4	63.9	53.8
gpt-oss-20b	**81.1**	73.5	67.3	58.0	47.5
gemma-3-12b-it	**59.7**	52.5	49.0	42.6	35.7
gemma-3-27b-it	**48.5**	35.7	31.1	32.5	27.5
gemma-3-4b-it	**45.6**	38.1	30.6	35.5	29.5
medgemma-27b-it	**74.9**	69.7	58.2	57.4	43.0
medgemma-4b-it	**57.7**	47.5	48.0	42.0	37.9
Preferred-MedRECT-32B	**83.5**	73.5	70.4	59.8	52.9

Table F.4: Accuracy by region-specificity score under English-only evaluation. For each model, the highest score-specific accuracy is shown in bold, and the lowest is underlined.

Model	Score 0	Score 1	Score 2	Score 3	Score 4
DeepSeek-V3.2	**89.5**	82.6	79.1	79.3	77.2
DeepSeek-V3.1	**88.2**	80.6	76.5	75.2	76.3
Llama-4-Maverick-17B-128E-Instruct	**78.6**	71.9	74.0	66.3	63.9
Llama-4-Scout-17B-16E-Instruct	**78.2**	69.5	66.3	60.4	66.7
Llama-3.3-70B-Instruct	**73.2**	66.3	63.3	62.1	61.9
Qwen2.5-32B-Instruct	**72.0**	65.1	59.7	62.7	62.4
Qwen2.5-72B-Instruct	**68.8**	61.5	60.2	59.8	57.0
Qwen2.5-14B-Instruct	**68.4**	60.7	57.6	59.2	56.1
Qwen3-235B-A22B-Thinking-2507	**91.1**	85.2	80.6	78.7	78.1
Qwen3-30B-A3B-Thinking-2507	**87.0**	77.3	76.0	74.0	68.8
Qwen3-235B-A22B	**86.7**	79.4	77.5	75.7	67.8
Qwen3-32B	**84.5**	74.0	67.3	69.2	62.2
Qwen3-235B-A22B-Instruct-2507	**82.8**	76.8	73.0	74.6	69.3
Qwen3-30B-A3B	**81.5**	71.3	63.3	68.6	57.6
Qwen3-14B	**81.3**	70.1	62.2	65.1	61.5
Qwen3-4B-Thinking-2507	**79.8**	69.3	64.3	58.6	57.0
Qwen3-8B	**75.4**	66.3	60.2	61.5	55.3
Qwen3-30B-A3B-Instruct-2507	**73.2**	67.5	62.8	60.4	53.3
Qwen3-4B-Instruct-2507	**62.9**	52.9	48.0	52.1	51.8
Qwen3-Swallow-32B-RL-v0.2	**87.1**	78.8	75.5	68.6	65.6
Qwen3-Swallow-30B-A3B-RL-v0.2	**85.4**	76.0	69.9	72.2	59.2
Qwen3-Swallow-32B-SFT-v0.2	**84.3**	73.8	69.9	65.1	59.2
Qwen3-Swallow-30B-A3B-SFT-v0.2	**83.2**	74.0	66.3	64.5	58.1
Qwen3-Swallow-8B-RL-v0.2	**80.6**	69.7	66.8	58.0	59.4
Qwen3-Swallow-8B-SFT-v0.2	**78.6**	70.1	65.8	59.2	57.8
GPT-OSS-Swallow-120B-RL-v0.1	**86.5**	77.8	73.5	68.0	68.2
GPT-OSS-Swallow-120B-SFT-v0.1	**84.7**	75.0	70.9	68.0	61.1
GPT-OSS-Swallow-20B-RL-v0.1	**80.2**	74.3	68.9	63.9	61.9
GPT-OSS-Swallow-20B-SFT-v0.1	**78.8**	68.3	64.8	58.6	56.6
gpt-oss-120b	**84.8**	75.8	71.9	66.9	60.8
gpt-oss-20b	**78.9**	69.7	63.3	62.7	47.3
gemma-3-12b-it	**53.5**	52.5	50.5	45.0	43.7
gemma-3-27b-it	**44.4**	35.7	31.6	27.8	31.6
gemma-3-4b-it	**41.0**	38.7	36.2	36.7	33.6
medgemma-27b-it	**61.0**	51.1	49.0	49.7	40.6
medgemma-4b-it	**49.0**	44.5	43.4	43.8	40.0
Preferred-MedRECT-32B	**83.3**	74.2	74.0	69.8	61.3

Table F.5: Accuracy by region-specificity score under original-source-language evaluation. For each model, the highest score-specific accuracy is shown in bold, and the lowest is underlined.

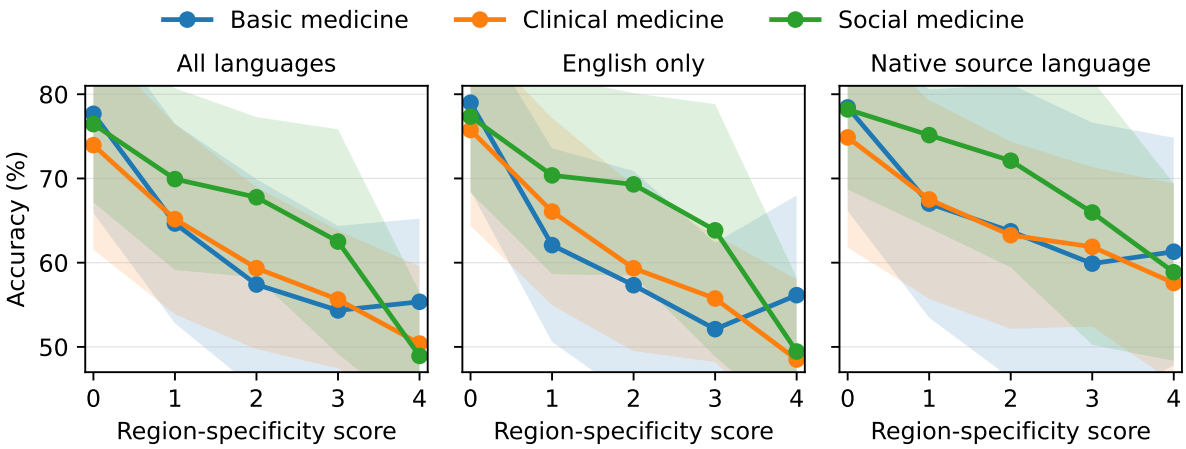

Figure F.1: Accuracy by region-specificity score, stratified by clinical category. Results are shown for the all-language, English-only, and native-source-language evaluation settings. Lines indicate mean accuracy across all evaluated models, and shaded regions indicate one standard deviation. Accuracy declined with increasing region-specificity within all three clinical categories, indicating that the trend is not explained solely by the overrepresentation of social-medicine questions among high-region-specificity items. Native-source-language evaluation partially mitigated the decline, especially in clinical and social medicine.

Family	Size	Adaptation	Stage	Subset	en	es	fr	ja	ru	sv	zh	Mean Δ	Δ SD
Qwen3	8B	Swallow	SFT	High	+5.32	+3.36	+1.68	+4.20	-5.32	+2.24	+5.60	+2.44	+0.96
			RL	High	+5.60	+5.32	+5.04	+5.05	-2.52	+0.00	+8.12	+3.80	+0.61
			SFT	Low	+0.51	+5.26	+2.66	+3.22	+6.59	-0.61	+8.03	+3.67	-1.62
			RL	Low	+3.53	+8.69	+5.06	+5.62	+9.97	+1.49	+11.76	+6.59	-1.36
Qwen3	30B-A3B	Swallow	SFT	High	+5.04	+1.96	+4.20	+4.21	+1.40	+0.56	+3.08	+2.92	+0.95
			RL	High	+5.88	+6.16	+5.60	+3.93	+6.16	+3.08	+3.64	+4.92	-0.58
			SFT	Low	+1.22	+2.82	+1.28	+3.79	+6.96	-1.33	+2.66	+2.49	-1.18
			RL	Low	+1.99	+5.01	+5.11	+6.09	+9.97	+3.83	+6.09	+5.44	-1.93
Qwen3	32B	Swallow	SFT	High	+2.52	+7.00	+5.60	+4.49	+5.60	+1.68	-0.84	+3.72	-0.73
			RL	High	+7.57	+11.20	+8.12	+11.21	+3.08	+10.64	+1.96	+7.68	+1.03
			SFT	Low	+1.49	+1.99	+3.22	+1.22	+2.81	+0.11	+0.41	+1.60	-0.60
			RL	Low	+2.41	+4.09	+6.60	+4.09	+5.06	+3.69	+2.20	+4.02	-0.87
gpt-oss	20B	Swallow	SFT	High	+9.24	+4.48	+5.04	+14.85	+7.84	-0.84	+8.40	+7.00	+2.82
			RL	High	+5.04	+11.77	+5.60	+20.73	+9.24	+12.61	+10.92	+10.84	+3.15
			SFT	Low	+1.12	+0.82	-0.97	+2.00	+0.00	-1.22	-1.02	+0.11	+0.48
			RL	Low	-8.18	+2.61	+2.76	+5.98	+4.81	+3.22	+3.53	+2.11	+2.01
gpt-oss	120B	Swallow	SFT	High	+3.37	+6.45	+3.09	+6.72	+4.76	+2.24	+1.40	+4.01	+0.74
			RL	High	+7.57	+11.49	+8.13	+14.28	+6.44	+7.85	+6.73	+8.93	+2.12
			SFT	Low	+0.00	+0.46	+1.07	+0.56	+1.23	+0.67	-0.10	+0.55	-0.27
			RL	Low	+0.87	+3.53	+2.71	+3.27	+2.81	+3.07	+1.89	+2.59	-0.08
Qwen3	32B	MedRECT	SFT	High	+1.12	+3.36	+2.52	+1.68	-2.80	+1.96	-0.56	+1.04	+0.28
			SFT	Low	+0.11	-1.13	-3.37	-2.36	+0.40	-0.76	+0.31	-0.98	+0.71

Table F.6: Language-wise changes in accuracy on Japan-derived XMedQA questions after Japanese medical adaptation, stratified by region specificity. Values indicate absolute percentage-point differences from the corresponding base model. High and Low denote region-specificity scores of 2–4 and 0–1, respectively. For Swallow models, RL denotes the final model obtained by further reinforcement learning after SFT, although all differences are computed relative to the original base model. Δ SD indicates the change in the standard deviation of accuracy across languages.

REGION-SPECIFIC SCORE 0: UNIVERSAL MEDICAL KNOWLEDGE

FRMED000080_en (FrenchMedMCQA)

Source language: fr | **Region-specific score**: 0
Predominant country: None
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Prevention and health promotion

Question: In 1998, out of 85,000 men aged 60 to 64, 300 cases of myocardial infarction were counted, of which 270 were new cases. Among the new cases, 135 died within the year. The total number of deaths due to this disease for the year is 150. Among the results below, indicate which one represents the incidence rate:

- a: 135/85,000
- b: 300/85,000
- c: 150/85,000
- d: 270/85,000
- e: 135/270

Answer: d
Rationale for score: The question tests the universal epidemiological definition of incidence rate and does not rely on any regional or country-specific medical knowledge.

HEADQA000868_en (HEAD-QA)

Source language: es | **Region-specific score**: 0
Predominant country: None
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Psychiatric disorders

Question: A 33-year-old man with no relevant medical history presents with anxiety and a depressed mood that he has been experiencing for a couple of months, which he relates to difficulties he is having at work and at home. He reports that his life has always been very chaotic, with frequent job changes and difficulties in completing tasks he starts. His work is very disorganized, he makes constant mistakes due to careless oversights, and this leads to problems at work and in his family. He is forgetful, very easily distracted, and is only able to concentrate on things that truly interest him. He has consumed alcohol for years, but in recent months this consumption has become more significant and is causing him problems at work and at home. He also admits to having lost control with slot machine gambling; he plays more and more, has more difficulty stopping, and is neglecting important activities to gamble. Regarding his personality, he describes himself as extroverted, sociable, somewhat stubborn, and having frequent mood swings, shifting from being very euphoric to feeling very sad in a few minutes, several times a day, generally conditioned by the level of stimulation or gratification he has at that moment. Which of the following diagnoses is unlikely in this patient?

- a: Adjustment disorder.
- b: Attention deficit hyperactivity disorder.
- c: Bipolar disorder.
- d: Alcohol abuse.

Answer: c
Rationale for score: The question tests universal psychiatric diagnostic criteria for ADHD, bipolar disorder, alcohol abuse, and adjustment disorder. There is no region-specific medical context.

2024F068_en (JMedQA)

Source language: ja | **Region-specific score**: 0
Predominant country: None
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Infectious diseases

Question: An 80-year-old man presented with fever and difficulty urinating. Present illness: He had been experiencing difficulty urinating for 3 days. He visited the emergency department today due to the onset of fever accompanied by chills and shivering. Past medical history: He takes an ACE inhibitor, a DPP-4 inhibitor, a statin, a tricyclic antidepressant, and an H1 histamine receptor antagonist for hypertension, diabetes, dyslipidemia, depression, and urticaria, respectively. Social history: He was a company employee until age 65 and is currently unemployed. He lives alone. He smoked 10 cigarettes/day for 50 years until age 70 and drinks alcohol occasionally. He keeps a cat. Family history: Father had a cerebral infarction, mother had diabetes. Physical examination: Conscious. Height 170 cm, weight 72 kg. Temperature 38.8 °C. Pulse 112/min, regular. Blood pressure 136/60 mmHg. Respiratory rate 26/min. SpO2 98% (room air). No abnormalities in palpebral or bulbar conjunctiva. No palpable thyroid enlargement or cervical lymphadenopathy. No abnormalities in heart or breath sounds. Lower abdomen is distended but not tender. Liver and spleen are not palpable. Left costovertebral angle tenderness is present. Digital rectal examination reveals a 4 cm prostate with no tenderness. Laboratory findings: Urine: Protein 1+, Glucose 1+, Occult blood (-), sediment shows 1-4 RBC/HPF, >100 WBC/HPF. Blood: RBC 4.61 million, Hb 14.9 g/dL, Ht 44%, WBC 14,300 (band neutrophils 2%, segmented neutrophils 90%, eosinophils 1%, basophils 1%, monocytes 3%, lymphocytes 3%), platelets 120,000, PT-INR 1.18 (ref 0.9-1.1). Biochemistry: Total protein 6.6 g/dL, Albumin 3.9 g/dL, Total bilirubin 1.5 mg/dL, Direct bilirubin 0.4 mg/dL, AST 13 U/L, ALT 15 U/L, LD 219 U/L (ref 124-222), ALP 79 U/L (ref 38-113), γ-GT 30 U/L (ref 13-64), CK 76 U/L (ref 59-248), BUN 26 mg/dL, Creatinine 1.3 mg/dL, Uric acid 4.7 mg/dL, Glucose 120 mg/dL, HbA1c 7.0% (ref 4.9-6.0), Total cholesterol 197 mg/dL, Triglycerides 226 mg/dL, HDL cholesterol 69 mg/dL, LDL cholesterol 83 mg/dL, Na 138 mEq/L, K 4.5 mEq/L, Cl 105 mEq/L. CRP 14 mg/dL. Chest X-ray: Cardiothoracic ratio 51%, no infiltrates in bilateral lung fields. Abdominal ultrasound: Large amount of urine retained in the bladder, no hydronephrosis. A urinary catheter was placed, and 2 sets of blood cultures and a urine culture were collected before admission. Ceftriaxone was started upon admission. The day after admission, both sets of blood cultures were positive for Escherichia coli. The results of the drug susceptibility test are shown in Table 1. The organism detected from the urine culture and its susceptibility results were the same as those from the blood culture. The current serum creatinine is 1.0 mg/dL, and his weight has not changed since admission. It was decided to discontinue ceftriaxone and change the antibiotic. The Cockcroft-Gault formula and Table 2 show the recommended antibiotic dosing based on renal function. Table 1: Drug susceptibility test results:

Ampicillin R, Cefazolin S, Meropenem S. (S: susceptible, R: resistant). Cockcroft-Gault formula: Male: Creatinine clearance = {(140 - age) × weight (kg)} / {72 × serum creatinine (mg/dL)}. Table 2: Recommended antibiotic dosing based on renal function (as defined by the institution). CrCl > 50 mL/min: Ampicillin A (2g q6h), Cefazolin D (2g q8h), Meropenem G (1g q8h). CrCl 10-50 mL/min: Ampicillin B (2g q8h), Cefazolin E (2g q12h), Meropenem H (1g q12h). CrCl < 10 mL/min: Ampicillin C (2g q12h), Cefazolin F (1g q24h), Meropenem I (0.5g q24h). Which of the following is the most appropriate dosing method when switching from ceftriaxone based on Table 2?

- a: A
- b: B
- c: C
- d: D
- e: E

Answer: d
Rationale for score: The question tests universal principles of antibiotic de-escalation based on susceptibility testing and renal dose adjustment using the Cockcroft-Gault formula. There is no region-specific medical knowledge required.

MEDQA001130_en (MedQA)

Source language: en | **Region-specific score**: 0
Predominant country: None
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Basic medicine / Basic medical sciences

Question: Please refer to the summary above to answer this question A cardiologist at another institution is trying to replicate the table in the abstract with a very similar study that he is conducting. After collecting data on changes in plaque volume, LDL-C levels, FEV1/FVC ratio, and ACQ scores from baseline to 12-month follow-up, he is curious if the changes in the aforementioned measurements are statistically significant. Which of the following is the most appropriate statistical method to compare the groups?" "Impact of pulmharkimab on asthma control and cardiovascular disease progression in patients with coronary artery disease and comorbid asthma Introduction: Active asthma has been found to be associated with a more than two-fold increase in the risk of myocardial infarction, even after adjusting for cardiovascular risk factors. It has been suggested that the inflammatory mediators and accelerated atherosclerosis characterizing systemic inflammation may increase the risk of both asthma and cardiovascular disease. This study evaluated the efficacy of the novel IL-1 inhibitor pulmharkimab in improving asthma and cardiovascular disease progression. Methods: In this double-blind, randomized controlled trial, patients (N=1200) with a history of coronary artery disease, myocardial infarction in the past 2 years, and a diagnosis of comorbid adult-onset asthma were recruited from cardiology clinics at a large academic medical center in Philadelphia, PA. Patients who were immunocompromised or had a history of recurrent infections were excluded. Patients were subsequently randomly assigned a 12-month course of pulmharkimab 75 mg/day, pulmharkimab 150 mg/day, or a placebo, with each group containing 400 participants. All participants were included in analysis and analyzed in the groups to which they were randomized regardless of medication adherence. Variables measured included plaque volume, serum LDL-C levels, FEV1/FVC ratio, and Asthma Control Questionnaire (ACQ) scores, which quantified the severity of

asthma symptoms. Plaque volume was determined by ultrasound. Analyses were performed from baseline to month 12. Results: At baseline, participants in the two groups did not differ by age, gender, race, plaque volume, serum LDL-C levels, FEV1/FVC ratio, and ACQ scores ($p > 0.05$ for all). A total of 215 participants (18%) were lost to follow-up. At 12-month follow-up, the groups contained the following numbers of participants: Pulmharkimab 75 mg/d: 388 participants Pulmharkimab 150 mg/d: 202 participants Placebo: 395 participants Table 1: Association between pulmharkimab and both pulmonary and cardiovascular outcomes. Models were adjusted for sociodemographic variables and medical comorbidities. All outcome variables were approximately normally distributed. Pulmharkimab 75 mg/d, (Mean +/- 2 SE) Pulmharkimab 150 mg/d, (Mean +/- 2 SE) Placebo, (Mean +/- 2 SE) P-value Plaque volume (mm3), change from baseline 6.6 ± 2.8 1.2 ± 4.7 15.8 ± 2.9 < 0.01 LDL-C levels, change from baseline -9.4 ± 3.6 -11.2 ± 14.3 -8.4 ± 3.9 0.28 FEV1/FVC ratio, change from baseline 0.29 ± 2.21 0.34 ± 5.54 -0.22 ± 3.21 0.27 ACQ scores, change from baseline 0.31 ± 1.22 0.46 ± 3.25 0.12 ± 1.33 0.43 Conclusion: Pulmharkimab may be effective in reducing plaque volume but does not lead to improved asthma control in patients with a history of myocardial infarction and comorbid asthma. Source of funding: Southeast Institute for Advanced Lung and Cardiovascular Studies, American Center for Advancement of Cardiovascular Health"

- a: Pearson correlation analysis
- b: Chi-square test
- c: Analysis of variance
- d: Kaplan-Meier analysis

Answer: c
Rationale for score: The question tests knowledge of biostatistics, specifically the selection of ANOVA to compare continuous variables across three groups. This is universal scientific knowledge with no regional dependence.

MSWE003031_en (MedQA-SWE)

Source language: sv | **Region-specific score**: 0
Predominant country: None
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Psychiatric disorders

Question: You are treating Frank, a 39-year-old man brought to your clinic by the police. Frank had broken into an apartment where an elderly woman, who does not know him, lives. She perceived the situation as an attempted robbery and called the police. Frank did not react significantly when asked to stop handling her belongings, which he was throwing out of an open window. The woman did not feel threatened; on the contrary, Frank was quite calm and polite, but behaved 'confused' and talked to himself (disorganized). Frank believes he is in his own apartment but does not protest when the police ask him to leave and accompany them to the emergency department. The police perceive Frank as sad and anxious. He is oriented to time, place, and person; he has stated that he feels confused, does not recognize himself, and does not understand what he was doing in the woman's apartment. You check previous psychiatric contacts and find that Frank was treated (for 3 days) 4 years ago for 'confusion' under the diagnosis 'Brief psychotic disorder,' where the death of both parents in a car accident was perceived as the triggering factor. He was treated with haloperidol, after which he improved rapidly and was discharged. Frank had brief contact with social services 15 years ago due

to cocaine abuse, but there seems to be no current abuse. During the conversation, Frank says he feels stressed and sad after his grandmother's passing two months ago. His grandmother meant a lot to him; he mostly lived with her during his upbringing as he had problems in his relationship with his mother. Together with his grandmother, he came to the city two years ago to 'help grandmother, find a job, and find a wife.' During the two months that have passed since her death, he has been sad, felt lonely, and slept extremely sporadically, 'an hour at a time.' However, he claims he has made efforts to find a job and keep the apartment but has not succeeded. For the last two weeks, he has been walking the streets and staying with various friends. The patient denies current drug intake and does not take medication for any specific illness. When asked why he was in an apartment that was not his, he says he went into the apartment because voices told him to do so. He is well aware that he had a similar episode of 'confusion' a few years earlier in connection with his parents' death. Mental status: Awake, oriented, neat appearance, maintains a train of thought, appears unaffected by any drugs, essentially normal emotional and formal contact, help-seeking and cooperative, lowered mood, anxious affects, sad, remorseful and ashamed, no affect lability, somewhat tense and agitated, no formal thought disorders, denies hallucinations, no suicidality, moderately depressive thought content, no grandiosity, insight is perceived as partial. Somatic and neurological examination is normal, as are routine laboratory tests. Your action is to offer the patient inpatient care due to suspected psychosis, which the patient accepts. Because the picture is somewhat unclear and the condition is not in any way dramatic, you decide to wait with medication but order medication for the night, given the patient's report of minimal sleep. The next day, you receive a report that the patient has slept a few hours and feels less confused, and you book an appointment with him for the following day. Frank now changes his story and says he went into the apartment because he thought it belonged to him. He is sincerely embarrassed and thinks the whole thing is embarrassing, but claims he was confused and 'imagined things, really childish things,' and that he deeply regrets it. Contrary to what he said earlier, Frank admits he had a few beers and that he has not had any hallucinations at all, but that he 'was just talking to himself.' Today he feels much better, has no symptoms left, not even depressive ones. When asked what he 'imagined,' Frank says that some people, including himself, can move from one time zone to another, which he has had proof of in recent weeks. But it has also happened before, after his parents' death. As far as he can remember from that time, he became upset, felt confused, talked to himself, and 'wandered in different time zones on the streets for several days before he received psychiatric care.' Frank finds it embarrassing to talk about his time zones but believes that he had to develop these early in life (as a kind of defense) against his mother's psychological and physical abuse. You thus have a patient with diffuse and inconsistent information about auditory hallucinations, 'confusion,' and who believes that time travel is possible, but who at the same time thinks it is nonsense and childish. Without antipsychotic treatment, the patient improves and the depressive mood has decreased significantly. Could something from childhood be the basis for the patient's experiences of 'traveling in different time zones'? Is the loss of loved ones the basis for the two periods of auditory hallucinations (which you still believe he had) with confusion? Could both these factors be sufficient to trigger a psychotic episode? You decide to wait a while longer to start antipsychotic treatment (haloperidol) even though you perceive the patient as dissimulating regarding his auditory hallucinations and that the experience of traveling in different time zones can be considered a delusion. You consider the most likely diagnosis at this stage to be 'Brief psychotic disorder, with significant stress factor' (grandmother's death and homelessness/living on the streets). The patient continues to deny auditory hallucinations and other symptoms but begins to withdraw more and more and appears affected by anxiety. After 10 days on the ward, Frank wants to talk to a nurse about his troubles. He has just informed his aunt about his grandmother's death and now tells about his great grief and continuous auditory hallucinations (which he has had since he was 10 years old). In the last few days, anxiety, depression, irritability, sweating, and shaking in the arms have been added, and he has difficulty sleeping. Contrary to previous information, Frank now admits that he has smoked marijuana daily for the past year (confirmed by urine screening). The reason he denied auditory hallucinations was the fear of being 'locked up' for a long time if it came out that he had

had hallucinations since childhood. The voices have always been male, never commented on what he does, or communicated with him or each other about him. The voices were alternately positive and accusingly negative, but he does not want to go into the exact content. The voices, he firmly believes, are not his own thoughts but come from his brain. The voices have not really caused much trouble. You now have reason to reformulate your working diagnosis and continue the investigation regarding the most likely condition(s) behind the patient's chronic auditory hallucinations. Obviously, the care failed to perform a urine screening upon admission, which would have shortened the time to adequate treatment. Cannabis use disorder seems most likely. You suspect that there may be an underlying real cause for the patient's early-onset auditory hallucinations and his claim that he has developed an ability to 'travel between different time zones,' something the patient does not seem particularly keen to explain. To make a long story short, it turned out that Frank had been systematically beaten by his mother from the age of 10 on the head with a specific sharp lady's shoe and on the body with a reinforced leather belt, which led to numerous hospital visits. He had made a vow never to mention it to anyone. You make a new decision regarding continued pharmacological treatment. The history suggests that similar symptoms responded well to antipsychotic treatment (4 years earlier). Now, depressive symptoms, sleep disturbance, and anxiety have crystallized, in light of probable cannabis withdrawal, which likely maintains or worsens the patient's auditory hallucinations and ongoing depression. He spends hours awake in bed and exhibits moderate to severe depressive syndromes (Depressive syndrome NOS) but without suicidal thoughts. You chose to combine an antipsychotic and a NaSSA (noradrenergic and specific serotonergic antidepressant) preparation, which showed good effect; already after one week of treatment, the patient exhibited only mild depressive symptoms, significantly less anxiety, improved sleep, and an almost complete absence of auditory hallucinations. Based on the information available above and the systematic exclusion of a number of psychiatric comorbid diagnoses, there is one condition left that should primarily be investigated further. Which one?

- a: Borderline personality disorder
- b: Delusional disorder
- c: Malingering
- d: Post-traumatic stress disorder
- e: Childhood schizophrenia

Answer: d
Rationale for score: The question relies on universal psychiatric diagnostic principles, specifically recognizing the presentation of trauma-related disorders (PTSD and dissociative symptoms) following severe childhood abuse. There is no country-specific or region-specific medical knowledge required.

REGION-SPECIFIC SCORE 0: UNIVERSAL KNOWLEDGE WITH MINOR LOCAL ARTIFACTS

FRMED000116_en (FrenchMedMCQA)

Source language: fr | **Region-specific score**: 0
Predominant country: France
Country scores: Japan=0, United States=0, China=0, France=1, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Basic medicine / Basic medical sciences

Question: Which of the following statements is correct? Furosemide (LASILIX® 40mg tablets):

- a: Acts primarily at the distal convoluted tubule
- b: Is inactive in patients with renal failure
- c: Is hypokalemic
- d: Has a delayed and prolonged action
- e: Is indicated for diabetes insipidus

Answer: c
Rationale for score: The question tests universal pharmacological knowledge about furosemide. The inclusion of the specific brand name LASILIX® is a terminology artifact (commonly used in France) but does not change the universal medical reasoning required to answer the question.

HEADQA000710_en (HEAD-QA)

Source language: es | **Region-specific score**: 0
Predominant country: Spain
Country scores: Japan=0, United States=0, China=0, France=0, Spain=1, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Collagen diseases

Question: Question linked to image no. 31. A 26-year-old woman consulted for a 3-week history of fever. The fever was daily, in spikes reaching 38.5ºC and occasionally 39ºC, and was accompanied by pharyngeal pain, myalgias, and arthralgias of large joints. Within a few hours, the fever would subside, either spontaneously or with antipyretics. During the febrile episodes, she presented with a skin rash. The patient reported having had two previous episodes in the two years prior, very similar to the current one, which had not been diagnosed despite thorough investigations. She had not traveled outside of Spain and reported no relevant epidemiological history. Clinical examination during a febrile episode showed the skin rash that can be observed in the image, which disappeared within a few hours. The patient also presented with pharyngeal hyperemia, small and painful bilateral laterocervical lymphadenopathy, and a pole of the spleen was palpable below the left costal arch. To establish the diagnosis, it is expected that various laboratory and imaging studies must be performed. However, which of the following statements seems most appropriate in the current diagnostic phase?

- a: I would request serological tests for hepatitis B and C, human parvovirus B19, Epstein-Barr virus, and a test for human immunodeficiency virus, given that the most likely cause is infectious diseases, especially acute viral infections.
- b: Although it is a diagnosis of exclusion, I would request a serum ferritin level due to the possibility that it is adult-onset Still's disease.
- c: I would request an immunological study including antinuclear antibodies, rheumatoid factor, anti-cyclic citrullinated peptide antibodies, and ANCAs to confirm the suspicion of a systemic rheumatic disease such as rheumatoid arthritis or systemic lupus erythematosus.
- d: I would request a genetic study given that it is most likely an autoinflammatory disease such as familial Mediterranean fever.

Answer: b
Rationale for score: The clinical presentation is classic for Adult-onset Still's disease, and the

diagnostic utility of serum ferritin is universally recognized. The mention of Spain is only to rule out travel-related infections and does not make the question dependent on regional medical knowledge.

2021F060_en (JMedQA)

Source language: ja | **Region-specific score**: 0
Predominant country: Japan
Country scores: Japan=1, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Emergency and critical care

Question: A 22-year-old male was brought in by ambulance following a fall. Present illness: While working at a height of approximately 6 m, he accidentally fell, landing on his lower limbs. He complains of severe lower back pain and left lower limb pain. Past history: None. Social history: Construction worker. Family history: Grandfather has diabetes. Physical examination: Consciousness level JCS I-3. Height 170 cm, weight 60 kg. Body temperature 35.6 °C. Heart rate 128/min, regular. Blood pressure 88/60 mmHg. Respiratory rate 24/min. SpO2 97% (on 10 L/min oxygen via reservoir mask). Capillary refill time is prolonged at 3 seconds. Pupil diameter 4 mm bilaterally, light reflex normal. Facial lacerations present. No oral bleeding. No jugular venous distension. Chest examination shows tenderness in the left anterior chest. Breath sounds are normal bilaterally with no asymmetry. No subcutaneous emphysema. Left lower limb is in external rotation and he refuses to move it due to pain, but no sensory abnormalities are noted. Persistent bleeding is observed from a wound on the left lower leg, which is heavily contaminated with soil. No abnormalities in left toe movement. Follows commands for hand grip, eye opening, and closing. Skin is cold and clammy. Laboratory findings: Urinalysis: Protein (-), glucose (-), ketones (-), occult blood (-), no WBCs in sediment. Blood tests: RBC 3.6 million, Hb 10.8 g/dL, Ht 32%, WBC 12,600, platelets 300,000, PT-INR 1.4 (ref 0.9-1.1), APTT 41.0 s (control 32.2 s), plasma fibrinogen 168 mg/dL (ref 200-400). Blood biochemistry: Total protein 5.0 g/dL, albumin 2.9 g/dL, total bilirubin 0.9 mg/dL, AST 60 U/L, ALT 22 U/L, CK 130 U/L (ref 30-140), BUN 30 mg/dL, creatinine 0.9 mg/dL, blood glucose 124 mg/dL, Na 133 mEq/L, K 4.8 mEq/L, Cl 104 mEq/L. 12-lead ECG: Normal. Imaging: FAST shows no fluid in the pericardial, pleural, or abdominal cavities. Chest X-ray shows left 4th-7th rib fractures and decreased transparency in the left lung field. Pelvic X-ray shows left L5 transverse process fracture, left iliac fracture, left pubic/ischial fracture, and left sacroiliac joint diastasis. Lower leg X-ray shows left tibial shaft fracture. What is the most likely pathophysiology?

- a: Anaphylactic shock
- b: Hemorrhagic shock
- c: Neurogenic shock
- d: Cardiogenic shock
- e: Obstructive shock

Answer: b
Rationale for score: The diagnosis of hemorrhagic shock in a trauma patient with pelvic and long bone fractures is universal medical knowledge. The use of the Japan Coma Scale (JCS) is a regional terminology artifact, but it does not affect the primary medical reasoning required to answer the question.

MEDQA001154_en (MedQA)

Source language: en | **Region-specific score**: 0
Predominant country: United States
Country scores: Japan=0, United States=1, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Nephrology

Question: A 68-year-old man presents to his primary care physician with complaints of increased fatigue and back pain for the past year. His symptoms are progressively getting worse over the past few months despite adequate rest and trial of over the counter analgesics. His back pain is constant and non-radiating in nature. He denies any trauma to his back. His past medical history is insignificant. On examination, there is mild tenderness over the lumbar region in the midline with an absence of paraspinal muscle tenderness. Straight leg raise test is negative. His skin and conjunctiva appear pale. The physician orders basic blood work. His results are provided below. Complete blood count RBC 4.9 million cells/µL Hemoglobin 9.8 g/dL Hematocrit 41% Total leukocyte count 6,800 cells/µL Neutrophils 70% Lymphocyte 26% Monocytes 3% Eosinophil 1% Basophils 0% Platelets 230,000 cells/µL Basic Metabolic Panel Sodium 136 mEq/L Potassium 5.1 mEq/L Chloride 101 mEq/L Bicarbonate 24 mEq/L Albumin 3.6 mg/ dL Urea nitrogen 31 mg/dL Creatinine 2.7 mg/dL Uric Acid 6.7 mg/dL Calcium 12.1 mg/dL Glucose 105 mg/dL Urinalysis shows proteinuria without any RBCs or pus cells. The patient's proteinuria best fits in which of the following category?

- a: Overflow proteinuria
- b: Transient (functional) proteinuria
- c: Orthostatic proteinuria
- d: Isolated proteinuria

Answer: a
Rationale for score: The question tests universal medical knowledge regarding the clinical presentation of multiple myeloma and the pathophysiology of overflow proteinuria. The use of US conventional units (mg/dL, mEq/L) is a minor regional artifact but does not change the universality of the medical reasoning.

MSWE002794_en (MedQA-SWE)

Source language: sv | **Region-specific score**: 0
Predominant country: Sweden
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=1, Russia=0, Other=0
Clinical category: Clinical medicine / Psychiatric disorders

Question: Adam, a single 32-year-old man with no children, reluctantly comes to the psychiatric outpatient clinic with his mother. His mother, who works as a primary care nurse, says that Adam has been completely healthy and well-functioning. Since his youth, he has always been a bit withdrawn and anxious, but he has always had a few good friendships. As far as she knows, he has never used drugs, has not committed any crimes, and has always been cautious with alcohol. Just under a year ago, Adam began to change; he started worrying that he would lose his job as a truck driver. A few months later, he developed severe sleep problems and was often up wandering aimlessly at night.

He was fired from his job about 5 months ago and had expressed at that time that the job was too stressful. Adam then started smoking marijuana 1–2 times per week, something he had never done before. Over the last two months, the symptoms have escalated, with restlessness, worry, and anxiety. His mother says that Adam has taped paper over the windows in his apartment, rarely goes out, and on several occasions, she has understood that he seemed to be talking to himself. When asked about this, Adam said that the neighbors were spying on him and that he could hear them talking about him, even when he was outside the apartment. Adam is oriented to time, place, person, and situation, is relatively coherent, tense, slightly agitated, irritable, and suspicious. He confirms auditory hallucinations. His mood is neutral, and suicidal thoughts are denied. Adam believes his mother does not understand him and maintains that he is not mentally ill, just stressed and tired. Adam wants to go home and wonders if there are any medications for his "sleep and stress." You conclude that it is most likely a schizophreniform syndrome, with typical schizophrenic symptomatology but of too short a duration (about 2–4 months) to meet the criteria for a diagnosis of schizophrenia. The patient opposes admission, and you assess that there is insufficient basis for involuntary admission under the Swedish Mental Health Act (LPT). However, you manage to persuade the patient to accept pharmacological treatment (antipsychotics), followed by a follow-up visit within a couple of weeks. His mother, who has some experience with antipsychotics, wants you to choose a preparation with the lowest risk of tardive dyskinesia and weight gain. Which of the following drugs is the best first-choice antipsychotic for this patient?

- a: Quetiapine
- b: Olanzapine
- c: Risperidone
- d: Haloperidol
- e: Aripiprazole

Answer: e
Rationale for score: The question tests knowledge of the side effect profiles of antipsychotic medications, specifically identifying aripiprazole as having a low risk of weight gain and tardive dyskinesia. This is universal pharmacological knowledge. The mention of the Swedish Mental Health Act is incidental narrative context and does not affect the medical reasoning required to answer the question.

REGION-SPECIFIC SCORE 1

FRMED000427_en (FrenchMedMCQA)

Source language: fr | **Region-specific score**: 1
Predominant country: Other
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=2
Clinical category: Clinical medicine / Infectious diseases

Question: During a trip to Thailand, which of the following parasitic diseases can be contracted?

- a: Malaria
- b: Strongyloidiasis

- c: Schistosomiasis due to Schistosoma haematobium
- d: Ancylostomiasis (hookworm infection)
- e: Schistosomiasis due to Schistosoma mansoni

Answer: a, b, d
Rationale for score: The question tests knowledge of global epidemiology and travel medicine, specifically identifying which parasitic diseases are endemic to Thailand. While it asks about a specific country's epidemiology, this is considered universal medical knowledge required by physicians worldwide for treating and advising travelers.

HEADQA000696_en (HEAD-QA)

Source language: es | **Region-specific score**: 1
Predominant country: Other
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=2
Clinical category: Clinical medicine / Hematologic disorders

Question: Question linked to image no. 17. A 22-year-old young patient is referred from Primary Care for evaluation. He is originally from Jamaica and has been living in Spain for five months. The reason for the consultation with the family doctor was a feeling of tiredness. As medical history, he reports several years of moderate-intensity pain in various joints, without inflammatory signs, which had not been studied in his country. Clinical examination shows pale mucous membranes and a painful ulcerated lesion on the leg, of three years' duration, for which he had not consulted. He did not recall any local trauma. Laboratory tests show Hb 7.6 g/dL, reticulocytes 417,000/uL, Hct 22.8%, leukocytes 13,540/uL, with 51% neutrophils and 36% lymphocytes, platelets 286,000/uL, bilirubin 2.12 mg/dL (direct 0.4 mg/dL), and lactate dehydrogenase 318 U/L (n.v. < 190 U/L). A basic coagulation study is within normal ranges. The chest X-ray is normal, as is a lower extremity Doppler ultrasound. What type of anemia would you suspect before starting the corresponding confirmatory studies?

- a: Sickle cell anemia.
- b: Autoimmune hemolytic anemia.
- c: Anemia of chronic disease.
- d: Iron deficiency anemia.

Answer: a
Rationale for score: The question tests universal medical knowledge regarding the clinical presentation of sickle cell anemia (chronic hemolytic anemia, vaso-occlusive joint pain, and leg ulcers). The mention of Jamaica serves as an epidemiological clue for genetic prevalence, while Spain is merely a setting detail.

2022E042_en (JMedQA)

Source language: ja | **Region-specific score**: 1
Predominant country: Japan
Country scores: Japan=3, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0,

Other=0
Clinical category: Social medicine / Health care systems

Question: An 83-year-old man presented to the emergency department with a chief complaint of dyspnea. Present illness: He has had repeated hospitalizations for respiratory disease for 6 years and is receiving home medical care and home nursing. He is on home oxygen therapy. Dyspnea worsened since last night, and as it did not improve after observation, he was brought to the hospital by a neighbor. Past history: Appendectomy at age 26. Social history: Lives alone. Wife died 3 years ago. Two children both live outside the prefecture. Smoking: 20 cigarettes/day for 58 years, starting at age 20. Quit 5 years ago. Family history: Father died of cerebral hemorrhage at age 80. Physical findings: On arrival, consciousness level is JCS I-1. Difficulty speaking due to dyspnea. Height 160 cm, weight 48 kg. Body temperature 36.4 °C. Pulse 100/min, regular. Blood pressure 124/72 mmHg. Respiratory rate 22/min. SpO2 88% (nasal cannula 1 L/min with portable oxygen cylinder). Diaphoretic. Breath sounds are diminished, and wheezing is heard on expiration. Pursed-lip breathing is observed, and sternocleidomastoid muscles are developed. Laboratory findings: Blood: RBC 4.52 million, Hb 15.3 g/dL, Ht 44%, WBC 9,200. Biochemistry: Total protein 6.4 g/dL, Albumin 2.8 g/dL, AST 36 U/L, ALT 32 U/L, LD 338 U/L (reference 120-245), BUN 25 mg/dL, Creatinine 0.8 mg/dL, Na 140 mEq/L, K 4.0 mEq/L, Cl 102 mEq/L. Arterial blood gas (on 1 L/min nasal cannula oxygen): pH 7.33, PaCO2 56 Torr, PaO2 58 Torr, HCO3- 30 mEq/L. After appropriate treatment during hospitalization, he is to be discharged. Which of the following is least involved in the discharge planning and support for this patient?

- a: Home care physician
- b: Visiting nurse
- c: Clinical laboratory technician
- d: Care manager
- e: Social worker

Answer: c
Rationale for score: The question tests knowledge of multidisciplinary discharge planning. While the core logic is universally applicable (clinical laboratory technicians are not typically involved in discharge planning), the use of the Japan Coma Scale (JCS) and the specific role of a 'care manager' reflect the Japanese healthcare and long-term care insurance systems.

MEDQA000963_en (MedQA)

Source language: en | **Region-specific score**: 1
Predominant country: United States
Country scores: Japan=0, United States=2, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Obstetrics and gynecology

Question: A 19-year-old female presents to the emergency room with 6 days of lower abdominal pain. She states that her symptoms have progressed and she now feels feverish as well. She also notes an episode of nonbloody diarrhea 2 days ago and has been increasingly nauseous. While being evaluated in the emergency room, she vomits three times. The patient has no significant past medical history but was recently treated with ciprofloxacin for an uncomplicated urinary tract infection. She

also notes irregular periods, with a last menstrual period 6 weeks ago. She is a college student who has had 3 sexual partners in the last year with inconsistent use of barrier protection. Her temperature is 100.5°C (38.1°F), blood pressure is 102/68 mmHg, pulse is 97/min, and respirations are 14/min. On exam, the patient is noticeably uncomfortable. There is marked tenderness to palpation in the suprapubic region and right lower quadrant, with voluntary guarding but no rebound. Pelvic exam shows moderate purulent discharge in the vaginal vault, cervical motion tenderness, and a tender adnexal mass on the right side. Which of the following is the best next step in treating this patient's condition?

- a: Intramuscular ceftriaxone and oral doxycycline with outpatient follow-up
- b: Intravenous cefotetan and doxycycline with inpatient admission
- c: Oral vancomycin with outpatient follow-up
- d: Methotrexate with outpatient follow-up

Answer: b
Rationale for score: The clinical presentation of pelvic inflammatory disease (PID) complicated by a tubo-ovarian abscess requires inpatient intravenous antibiotics. While the specific regimen of intravenous cefotetan and doxycycline is heavily emphasized in US CDC guidelines, the general principles of diagnosis and the need for inpatient admission are universally applicable.

MSWE003030_en (MedQA-SWE)

Source language: sv | **Region-specific score**: 1
Predominant country: Sweden
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=2, Russia=0, Other=0
Clinical category: Clinical medicine / Psychiatric disorders

Question: You are treating Frank, a 39-year-old man brought to your clinic by the police. Frank had 'broken into' an apartment where an elderly woman, who does not know Frank, lives. She perceived the situation as an attempted robbery and called the police. Frank did not react significantly when asked to stop handling her belongings, which he was throwing out of an open window. The woman did not feel threatened; on the contrary, Frank was quite calm and polite, but behaved 'confused' and talked to himself (disorganized). Frank believes he is in his own apartment but does not protest when the police ask him to leave and accompany them to the emergency department. The police perceive Frank as sad and anxious. He is oriented to time, place, and person; he has stated that he feels confused, does not recognize himself, and does not understand what he was doing in the woman's apartment. You check previous psychiatric contacts and find that Frank was treated (for 3 days) 4 years ago for 'confusion' under the diagnosis 'Brief psychotic disorder,' where the death of both parents in a car accident was perceived as the triggering factor. He was treated with haloperidol, after which he improved rapidly and was discharged. Frank had brief contact with social services 15 years ago due to cocaine abuse, but there seems to be no current abuse. During the interview, Frank says he feels stressed and sad after his grandmother's passing two months ago. His grandmother meant a lot to him; he mostly lived with her during his upbringing as he had problems in his relationship with his mother. Together with his grandmother, he came to the city two years ago to 'help grandmother, find a job, and find a wife.' During the two months since her death, he has been sad, felt lonely, and slept very sporadically, 'an hour at a time.' However, he claims he has made efforts to find a job and keep the apartment but has not succeeded. For the last two weeks, he has been walking the streets and

staying with various friends. The patient denies current drug intake and does not take medication for any specific illness. When asked why he was in an apartment that was not his, he says he went in because voices told him to. He is well aware that he had a similar episode of 'confusion' a few years earlier in connection with his parents' death. Mental status: Awake, oriented, neat appearance, maintains a train of thought, appears unaffected by any drugs, essentially normal emotional and formal contact, help-seeking and cooperative, lowered mood, anxious affects, sad, remorseful and ashamed, no affect lability, somewhat tense and agitated, no formal thought disorders, denies hallucinations, no suicidality, moderately depressive thought content, no grandiosity, insight is perceived as partial. Somatic and neurological examination are normal, as are routine laboratory tests. Your action is to offer the patient inpatient care due to suspected psychosis, which the patient accepts. As the picture is somewhat unclear, there are several alternatives for acute treatment. Because the picture is somewhat unclear and the condition is not in any way dramatic, you decide to wait with medication but prescribe medication for the night, considering the patient's report of minimal sleep. The next day, you receive a report that the patient has slept a few hours and feels less confused, and you book an appointment with him for the following day. Frank now changes his story and says he went into the apartment because he thought it belonged to him. He is sincerely embarrassed and thinks the whole thing is embarrassing, but claims he was confused and 'imagined things, really childish things,' and that he deeply regrets it. Contrary to what he said earlier, Frank admits that he had had a few beers and that he had no hallucinations at all, but that he 'just talked to himself.' Today he feels much better, has no symptoms left, not even depressive ones. When asked what he 'imagined,' Frank says that some people, including himself, can move from one time zone to another, which he has had proof of in recent weeks. But it has also happened before, after his parents' death. What he can remember from that time is that he became upset, felt confused, talked to himself, and 'wandered in different time zones on the streets for several days before he received psychiatric care.' Frank finds it embarrassing to talk about his time zones but believes that he had to develop these early in life (as a kind of defense) against his mother's mental and physical abuse. You thus have a patient with diffuse and inconsistent information about auditory hallucinations, 'confusion,' and who believes that time travel is possible, but who at the same time thinks it is nonsense and childish. Without antipsychotic treatment, the patient improves and the depressive mood has decreased significantly. Could something from childhood be the basis for the patient's experiences of 'traveling in different time zones'? Is the loss of loved ones the basis for the two periods of auditory hallucinations (which you still believe he had) with confusion? Could both these factors be sufficient to trigger a psychotic episode? You decide to wait a while longer to start antipsychotic treatment (haloperidol) even though you perceive the patient as dissimulating regarding his auditory hallucinations and that the experience of traveling in different time zones can be considered a delusion. You consider the most likely diagnosis at this stage to be 'Brief psychotic disorder, with significant stress factor' (grandmother's death and homelessness/street life). The patient continues to deny auditory hallucinations and other symptoms but begins to withdraw more and more and appears affected by anxiety. After 10 days on the ward, Frank wants to talk to a nurse about his problems. He has just informed his aunt about his grandmother's death and now tells about his great grief and continuous auditory hallucinations (which he has had since he was 10 years old). In the last few days, anxiety, depression, irritability, sweating, and tremors in his arms have been added, and he has difficulty sleeping. Contrary to previous information, Frank now admits that he has smoked marijuana daily for the past year (confirmed by urine screening). The reason he denied auditory hallucinations was the fear of being 'locked up' for a long time if it came out that he had had hallucinations since childhood. The voices have always been male, never commented on what he does, or communicated with him or each other about him. The voices were alternately positive and alternately of an accusing negative nature, but he does not want to go into the exact content. The voices, he firmly believes, are not his own thoughts but come from his brain. The voices have not really caused much trouble. You now have reason to reformulate your working diagnosis and continue the investigation regarding the most likely condition(s) behind the patient's chronic auditory hallucinations. Obviously, the care unit failed to perform a urine screening at admission, which would

1036

have shortened the time to adequate treatment. Cannabis use disorder seems most likely. You suspect that there may be an underlying real cause for the patient's early-onset auditory hallucinations and his claim that he has developed an ability to 'travel between different time zones,' something the patient does not seem very keen to explain. To make a long story short, it turned out that Frank had been systematically beaten by his mother from the age of 10 on the head with a specific sharp lady's shoe and on the body with a reinforced leather belt, which led to numerous hospital visits. He had made a promise never to mention it to anyone. You make a new decision regarding continued pharmacological treatment. The history suggests that similar symptoms responded well to antipsychotic treatment (4 years earlier). Now, depressive symptoms, sleep disturbance, and anxiety have crystallized, in light of probable cannabis withdrawal, which likely maintains or worsens the patient's auditory hallucinations and ongoing depression. He spends hours awake in bed and exhibits moderate to severe depressive syndromes (Depressive syndrome NOS) but without suicidal thoughts. What should be your preferred choice when it comes to pharmacological treatment?

- a: SSRI + Zopiklon (till natten)
- b: Bediol (torkat extrakt av Cannabis Sativa; THC/CBD (licenspreparat))
- c: Risperidon + Mirtazapin
- d: Oxascand
- e: Suboxone sublingualt (låg dos; subterapeutisk i jämförelse med opiatsubstitution)

Answer: c
Rationale for score: The clinical reasoning for treating psychotic symptoms with an antipsychotic (risperidone) and comorbid depression/insomnia with a sedating antidepressant (mirtazapine) is based on universal psychiatric principles. The use of local brand names (Oxascand) and regulatory terms (licenspreparat) reflects a Swedish context but does not fundamentally change the universal medical knowledge required to answer the question.

REGION-SPECIFIC SCORE 2

FRMED000135_en (FrenchMedMCQA)

Source language: fr | **Region-specific score**: 2
Predominant country: France
Country scores: Japan=0, United States=0, China=0, France=3, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Infectious diseases

Question: Regarding Taenia saginata taeniasis, only one of these statements is false, which one?

- a: It is the most common taeniasis in France
- b: It is caused by the presence of a trematode in its adult form in humans
- c: The responsible worm is a heteroxenous parasite
- d: The infective form for humans is transmitted by undercooked beef
- e: Niclosamide (TREDEMINE®) is the first-line treatment

Answer: b
Rationale for score: The core medical fact required to identify the false statement (Taenia saginata

is a cestode, not a trematode) is universal. However, the distractors rely on knowledge of French epidemiology (most common taeniasis in France) and a specific drug brand name available in France, giving an advantage to those familiar with the French medical context.

HEADQA000433_en (HEAD-QA)

Source language: es | **Region-specific score**: 2
Predominant country: Spain
Country scores: Japan=0, United States=0, China=0, France=0, Spain=3, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Prevention and health promotion

Question: According to the health promotion and prevention strategy in the National Health System, what recommendations should be made regarding diet? Indicate the INCORRECT one:

- a: Limit the consumption of high-calorie drinks/foods, considering their habitual daily consumption to be unhealthy.
- b: Consume 5 or more servings of fruit and vegetables per day.
- c: Reduce salt intake (< 5 g/day).
- d: Substitute unsaturated fats with saturated fats.

Answer: d
Rationale for score: The question explicitly references a specific National Health System's health promotion and prevention strategy (likely Spain's 'Sistema Nacional de Salud'). However, the incorrect option (substituting unsaturated fats with saturated fats) is universally recognized as unhealthy, making the core medical knowledge highly universal despite the regional framing.

2018B042_en (JMedQA)

Source language: ja | **Region-specific score**: 2
Predominant country: Japan
Country scores: Japan=3, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Geriatrics

Question: A 76-year-old woman presents with a chief complaint of shortness of breath. Present illness: She began noticing shortness of breath one year ago, and for the past three months, it has occurred after walking for about 10 minutes. Three days ago, she caught a cold, which worsened her shortness of breath to the point where she could not move, so she was brought by her daughter to the respiratory outpatient clinic of a general hospital. Past medical history: She has been attending multiple medical institutions for diabetes, hypertension, chronic heart failure (NYHA II), osteoarthritis of the knee, osteoporosis, and insomnia. For the past six months, her visits to these institutions have been inconsistent. Social history: She lives with her daughter. Her daughter is at work during the day. She can eat, use the toilet, and dress herself, but housework and going out are difficult, and her daughter assists her with bathing. Smoking history: 15 cigarettes/day for 45 years. No alcohol history. Physical examination: Conscious and alert. Height 158 cm, weight 42 kg. Temperature 36.6 °C. Pulse 104/min, regular. Blood pressure 120/76 mmHg. Respiratory rate 28/min. SpO2 93% (room

air). Skin is normal. No abnormalities in palpebral or bulbar conjunctiva. No thyroid enlargement or lymphadenopathy in the neck, and no jugular venous distension. Use of accessory respiratory muscles is prominent. No abnormalities in heart sounds. Wheezes are heard in both lung fields, but no crackles. Abdomen is flat and soft. No edema in the extremities. Revised Hasegawa's Dementia Scale score is 27/30. Laboratory findings: Chest X-ray shows pulmonary hyperinflation, but no infiltrates or pulmonary congestion. Cardiothoracic ratio is 53%. Chest CT shows low attenuation areas throughout the lung fields. After 4 days of outpatient treatment with oral corticosteroids and inhaled beta-agonists, her acute respiratory symptoms improved, and SpO2 reached 96% (room air). However, a nurse reported that the patient said, 'I don't intend to quit smoking, but do I really have to keep coming to the hospital?' Which of the following is the most appropriate response to the patient at this time?

- a: Mandate attendance at a smoking cessation clinic.
- b: Refer her to a primary care physician and encourage regular follow-up visits.
- c: Ask in detail about the situation of relatives who do not live with her.
- d: Encourage continued visits to all medical departments where she has a history of treatment.
- e: Explain that she should not easily visit the general hospital even if symptoms recur.

Answer: b
Rationale for score: The core medical concept of consolidating care for an elderly patient with multimorbidity is universal. However, the scenario reflects a healthcare system with free access where patients often visit multiple specialists independently, and explicitly uses the Revised Hasegawa's Dementia Scale, which is standard in Japanese clinical practice.

MEDQA000762_en (MedQA)

Source language: en | **Region-specific score**: 2
Predominant country: United States
Country scores: Japan=0, United States=3, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Gastrointestinal disorders

Question: A 39-year-old woman comes to the physician because of fever, generalized fatigue, and chills for 1 week. She has been having dull right-side abdominal pain during this period. She has nausea. She has no history of travel outside the United States. She has type 1 diabetes mellitus. Current medications include insulin and an oral contraceptive. She appears ill. Her temperature is 40°C (104°F), pulse is 118/min, and blood pressure is 106/70 mm Hg. On examination, the liver is palpated 2–3 cm below the right costal margin and is tender. There are decreased breath sounds over the right lung base. The remainder of the examination shows no abnormalities. Laboratory studies show: Hemoglobin 14.1 g/dL Leukocyte count 17,100/mm3 Segmented neutrophils 74% Eosinophils 1% Lymphocytes 20% Monocytes 5% Serum Glucose 79 mg/dL Creatinine 1.1 mg/dL Total bilirubin 0.9 mg/dL Alkaline phosphatase 180 U/L AST 69 U/L ALT 100 U/L A right upper quadrant ultrasound is shown. Which of the following is the most appropriate next step in management?"

- a: Discontinue the oral contraceptive
- b: Obtain hepatoiminodiacetic acid scan
- c: Perform endoscopic retrograde cholangiopancreatography

- d: Perform percutaneous drainage "

Answer: d
Rationale for score: The question explicitly states the patient has no travel history outside the United States. This epidemiological clue is used to rule out an amebic liver abscess (which is endemic in many tropical regions and typically treated with antibiotics alone) and point toward a pyogenic liver abscess (which requires percutaneous drainage). While the core medical management is universal, the diagnostic reasoning relies on knowing that the US is not endemic for amebiasis.

MSWE003027_en (MedQA-SWE)

Source language: sv | **Region-specific score**: 2
Predominant country: Sweden
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=3, Russia=0, Other=0
Clinical category: Clinical medicine / Psychiatric disorders

Question: You meet Frank, 39, who has been brought to your clinic by the police. Frank has 'broken into' an apartment where an elderly woman, who does not know Frank, lives. She perceived the situation as an attempted robbery and called the police. Frank did not react significantly when asked to stop picking up her belongings, which he threw out of an open window. The woman did not feel threatened; on the contrary, Frank was quite calm and polite, but behaved 'confused' and talked to himself (disorganized). Frank believes he is in his own apartment, but does not protest when the police ask him to leave the apartment and accompany them to the emergency department. The police perceive Frank as sad and anxious. He is oriented to time, place, and person; he has said that he feels confused and that he does not recognize himself, and that he does not understand what he was doing in the woman's apartment. You check previous psychiatric contacts and find that Frank was treated (for 3 days) 4 years ago due to 'confusion', under the diagnosis 'Brief psychotic disorder', where the death of both parents in a car accident was perceived as the triggering factor. He was then treated with T. haloperidol, after which he improved rapidly and was discharged. Frank had a brief contact with social services 15 years ago due to cocaine abuse, but there seems to be no current abuse. During the conversation, Frank says he feels stressed and sad after his grandmother's passing two months ago. His grandmother meant a lot to him; he usually lived with her during his upbringing when he had problems in his relationship with his mother. Together with his grandmother, he came to the city two years ago to 'be able to help grandmother, find a job, and find a wife'. During the two months that have passed since her death, he has been sad, felt lonely, and slept very sporadically, 'an hour at a time'. However, he says he has made efforts to find a job and keep the apartment, but has not succeeded. For the last two weeks, he has been walking the streets and staying with various friends. The patient denies current drug intake and does not take medication for any specific illness. When asked why he was in an apartment that was not his, he says he went into the apartment because voices told him to do so. He is well aware that he had a similar episode of 'confusion' a few years earlier in connection with his parents' death. Mental status: Awake, oriented, neat appearance, maintains a train of thought, appears unaffected by any drugs, essentially normal emotional and formal contact, help-seeking and cooperative, lowered mood, anxious affects, sad, remorseful and ashamed, no affect lability, somewhat tense and agitated, no formal thought disorders, denies hallucinations, no suicidality, moderately depressive thought content, no grandiosity, insight into illness is perceived as partial. Somatic and neurological examination results are normal, as are routine laboratory tests. Your action is to offer the patient inpatient care due to suspected psychosis, which the patient accepts. Since the clinical picture is somewhat unclear, there are several alternatives for acute treatment. What is the most reasonable measure regarding pharmacological treatment?

- a: T. haloperidol 1 mg 2 × 1 in combination with T. sertraline 50 mg 1 × 1
- b: No treatment except T. propiomazine 25 mg 1-2 at night as needed
- c: T. haloperidol 1 mg 2 × 1
- d: T. sertraline 50 mg 1 × 1
- e: T. mirtazapine 30 mg 1 × 1

Answer: b
Rationale for score: The clinical scenario tests the management of a brief psychotic episode, which relies on universal psychiatric principles. However, the options include propiomazine, a hypnotic drug primarily used in Sweden, and use Swedish prescription abbreviations ('T.' for tablet), which gives an advantage to those familiar with Swedish clinical practice and drug availability.

REGION-SPECIFIC SCORE 3

FRMED000034_en (FrenchMedMCQA)

Source language: fr | **Region-specific score**: 3
Predominant country: France
Country scores: Japan=0, United States=1, China=0, France=4, Spain=1, Sweden=1, Russia=0, Other=0
Clinical category: Social medicine / Prevention and health promotion

Question: Among the following statements, select the correct one. Hepatitis B vaccination is neither recommended nor mandatory for:

- a: Infants
- b: Healthcare personnel
- c: Adolescents
- d: Adults
- e: The household or partner of an individual infected with the hepatitis B virus or a chronic carrier of the HBs antigen

Answer: d
Rationale for score: Hepatitis B vaccination policies vary significantly by country. While vaccination for infants, healthcare workers, and high-risk contacts is universally recommended or mandatory, routine vaccination for general adults is not recommended in some countries (like France), whereas others (like the United States) now recommend it for all adults up to age 59. Thus, the correct answer depends on specific national public health guidelines.

HEADQA000797_en (HEAD-QA)

Source language: es | **Region-specific score**: 3
Predominant country: Spain
Country scores: Japan=0, United States=0, China=0, France=0, Spain=4, Sweden=0, Russia=0,

Other=0
Clinical category: Social medicine / Prevention and health promotion

Question: At the International Vaccination Center of Bilbao, a study is intended to be conducted to determine if the incidence and characteristics of adverse effects appearing in their travelers after the administration of the yellow fever vaccine correspond to the evidence available in the literature. What type of study is this?

- a: Phase III safety clinical trial.
- b: Case-control study.
- c: Pre-authorization study.
- d: Post-authorization observational study of prospective follow-up (EPA).

Answer: d
Rationale for score: While the epidemiological concept of a post-marketing observational study is universal, the question relies on the specific Spanish regulatory acronym 'EPA' (Estudio Posautorización) to identify the correct study design, making it highly specific to Spain's medical and regulatory framework. The mention of Bilbao is an additional surface cue.

2025C064_en (JMedQA)

Source language: ja | **Region-specific score**: 3
Predominant country: Japan
Country scores: Japan=4, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Health care systems

Question: A 58-year-old man was brought to the emergency department by ambulance due to fever and impaired consciousness. Present illness: He had a fever of 38.0 °C yesterday, took over-the-counter antipyretics, and stayed home. This morning, he was unresponsive in bed, prompting his family to call an ambulance. Past history: Left nephrectomy and splenectomy due to traffic trauma at age 21. Social history: Office worker, non-smoker. Drinks 350 mL of beer twice a week. Lives with his 52-year-old wife and two sons in their 20s. No pets. Vaccination history is unknown. Family history: Father died of myocardial infarction at age 78. No significant findings in cohabiting family members. Physical examination: Consciousness level JCS III-200. Height 170 cm, weight 62 kg. Temperature 38.8 °C. Heart rate 128/min, regular. Blood pressure 76/52 mmHg. Respiratory rate 30/min. SpO2 91% (on 10 L/min oxygen via reservoir mask). Skin is moist with marked diaphoresis. No abnormalities in conjunctiva or head/neck. Heart and breath sounds are normal. Abdomen is flat and soft, liver not palpable. Surgical scar on the left flank. Dark purple discoloration on both toes. No edema in lower legs. Laboratory findings: RBC 4.9 million, Hb 13.2 g/dL, Ht 39%, WBC 2,000 (myelocytes 5%, metamyelocytes 13%, band neutrophils 38%, segmented neutrophils 41%, eosinophils 0%, basophils 0%, monocytes 1%, lymphocytes 2%), platelets 62,000, PT-INR 1.4 (ref 0.9-1.1), fibrinogen 426 mg/dL (ref 186-355), D-dimer 30 μg/mL (ref $\leq$1.0). Biochemistry: Total protein 5.9 g/dL, albumin 2.8 g/dL, total bilirubin 2.7 mg/dL, direct bilirubin 1.8 mg/dL, AST 197 U/L, ALT 149 U/L, LD 351 U/L (ref 124-222), ALP 109 U/L (ref 38-113), γ-GT 60 U/L (ref 13-64), amylase 44 U/L (ref 44-132), CK 460 U/L (ref 59-248), BUN 40 mg/dL, creatinine 2.1 mg/dL, glucose 98 mg/dL, Na 130 mEq/L, K 3.4 mEq/L, Cl 101 mEq/L, Ca 7.5 mg/dL, lactate 28 mg/dL (ref 5-20). CRP 30 mg/dL. Arterial blood gas (10 L/min O2): pH 7.33, PaCO2 28 Torr, PaO2

84 Torr, HCO3- 18 mEq/L. Blood cultures were drawn, intravenous access was secured, and fluid resuscitation and antibiotics were initiated, but hypotension persisted. The patient died on the 4th day of hospitalization despite intensive care. The family consented to a pathological autopsy. Which of the following is correct?

- a: Return the body to the family after the autopsy is completed.
- b: Dispose of all instruments used in the autopsy.
- c: Report the performance of the autopsy to the police.
- d: Medical personnel involved in the autopsy should take prophylactic antibiotics.
- e: Issue the death certificate after the pathological autopsy report is completed.

Answer: a
Rationale for score: The clinical scenario describes Overwhelming Post-Splenectomy Infection (OPSI), which is universal medical knowledge. However, the question focuses on post-mortem administrative procedures, testing knowledge of national laws regarding death certification, police reporting for unnatural deaths, and autopsy regulations, which are specific to Japan's legal framework. The use of the Japan Coma Scale (JCS) is also a country-specific terminology artifact.

MEDQA000375_en (MedQA)

Source language: en | **Region-specific score**: 3
Predominant country: United States
Country scores: Japan=0, United States=4, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Clinical medicine / Infectious diseases

Question: A 50-year-old man presents to his physician with chronic cough, aching joints in his hips and lower back, and malaise over the past 2 months. He describes himself as being "generally healthy" before the cough began. Past medical history includes hypertension. The patient takes clopamide and a multivitamin daily. His parents are both well and living in a senior living facility. He does not smoke and only drinks alcohol occasionally. During a review of symptoms, the patient described a hiking trip in and around the desert near Phoenix, Arizona, 4 months ago. At the office, his temperature is 38.6°C (101.4°F), heart rate is 102/min, respirations are 20/min, and blood pressure is 120/82 mm Hg. A focused chest exam reveals mild fremetus and dullness with percussion on the right side. A chest X-ray shows a right-sided consolidation of the lower right lung, and a chest CT shows an irregular opacity measuring 3.8 cm x 3.0 cm in the sub-plural region of the right middle lobe, a small right-sided pleural effusion, and mild right-sided hilar adenopathy. A lung biopsy is performed to rule out cancer and reveals necrotic granulomatous inflammation with multinucleated giant cells and spherules with endospores in the surrounding tissues. The laboratory tests show the following results: Hemoglobin 12.9 mg/dL Leukocyte count 9,300/mm3 Platelet count 167,000/mm3 Erythrocyte sedimentation rate 43 mm/hr Periodic acid-Schiff and silver methenamine Positive Acid-fast stain Negative Which of the following is the most likely diagnosis?

- a: Histoplasmosis
- b: Mycetoma
- c: Blastomycosis
- d: Valley fever

Answer: d
Rationale for score: The diagnosis of Coccidioidomycosis (Valley fever) relies heavily on the epidemiological clue of travel to an endemic area, specifically the desert Southwest of the United States (Phoenix, Arizona), along with the classic histological finding of spherules with endospores.

MSWE002795_en (MedQA-SWE)

Source language: sv | **Region-specific score**: 3
Predominant country: Sweden
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=4, Russia=0, Other=0
Clinical category: Clinical medicine / Psychiatric disorders

Question: Adam, a single 32-year-old man with no children, reluctantly comes to the outpatient psychiatric clinic with his mother. His mother, who works as a primary care nurse, says that Adam has been completely healthy and well-functioning. Since his youth, he has always been a bit withdrawn and anxious, but he has always had a few good friendships. As far as she knows, he has never used drugs, has not committed any crimes, and has always been careful with alcohol. Almost a year ago, Adam began to change; he started worrying that he would lose his job as a truck driver. A few months later, he developed severe sleep problems and was often up wandering aimlessly at night. He was fired from his job about 5 months ago and had expressed that the job was too stressful. Adam then started smoking marijuana 1–2 times per week, something he had never done before. Over the last two months, the symptoms have escalated, with restlessness, worry, and anxiety. His mother says that Adam has taped paper over the windows in his apartment, rarely goes out, and on several occasions, she has understood that he seemed to be talking to himself. When asked about this, Adam said that the neighbors were spying on him and that he could hear them talking about him, even when he was outside the apartment. Adam is oriented to time, place, person, and situation, is relatively coherent, tense, slightly agitated, irritable, and suspicious. He confirms auditory hallucinations. His mood is neutral, and he denies suicidal thoughts. Adam believes his mother does not understand him and maintains that he is not mentally ill, just stressed and tired. Adam wants to go home and wonders if there are any medications for "sleep and stress." You conclude that it is most likely a schizophreniform syndrome, with typical schizophrenic symptomatology but of too short a duration (about 2–4 months) to meet the criteria for a diagnosis of schizophrenia. The patient opposes admission, and you assess that there is insufficient basis for admission under the Compulsory Psychiatric Care Act (LPT). However, you manage to persuade the patient to accept pharmacological treatment (antipsychotics), followed by a follow-up visit within a couple of weeks. His mother, who has some experience with antipsychotics, wants you to choose a preparation with the lowest risk of tardive dyskinesias and weight gain. You therefore choose T. aripiprazole, which is prescribed in a low initial dose (5 mg/day) and titrated weekly to the target dose (15 mg/day). The patient fails to attend the follow-up visit. His mother calls after four weeks. Especially during the last week, Adam has worsened, no longer goes out during the day, sleeps little, barely eats or drinks, and walks around town for several hours every night. He is increasingly irritated with the neighbors. On several occasions, he has said that "he will put an end to them once and for all." His mother believes there is a hunting weapon in the house. You assume he has not been taking his medication. What is the best management now?

- a: Since he is perceived as threatening, you make a report to the police authority that the patient cannot be trusted with firearms
- b: You act quickly due to the somewhat threatening situation and therefore call the patient to your clinic the very next day

- c: Increase the dose according to plan, prescribe T. zopiclone 7.5 mg at night for sleep, and ask the mother to dispense the medication
- d: Since the patient lives 40 km outside the city, you contact the nearest health center by phone and ask them to call the patient for an examination as soon as possible
- e: Arrange for secure transport (police transport) to the psychiatric clinic for assessment

Answer: e
Rationale for score: The question explicitly mentions the Swedish Compulsory Psychiatric Care Act (LPT) and the correct management involves arranging police transport for an involuntary psychiatric assessment, which aligns with Swedish legal and clinical practice for acute, dangerous psychiatric patients.

REGION-SPECIFIC SCORE 4

FRMED000155_en (FrenchMedMCQA)

Source language: fr | **Region-specific score**: 4
Predominant country: France
Country scores: Japan=0, United States=0, China=0, France=4, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Health care systems

Question: A patient on multiple medications presents, about ten days after starting treatment with CYMÉVAN® (ganciclovir) at a dose of 4 capsules of 250 mg three times a day, with a high fever accompanied by edema of the lower limbs. The prescriber is wondering if this condition is attributable to the administered product. Who can perform the imputability assessment?

- a: The AFSSAPS
- b: The pharmaceutical manufacturing laboratory
- c: The laboratory's medical representative
- d: The regional pharmacovigilance center
- e: The poison control center

Answer: d
Rationale for score: The question asks about the specific entity responsible for performing an imputability assessment for an adverse drug reaction, referencing AFSSAPS and regional pharmacovigilance centers, which are specific to the French pharmacovigilance and healthcare system.

HEADQA000414_en (HEAD-QA)

Source language: es | **Region-specific score**: 4
Predominant country: Spain
Country scores: Japan=0, United States=0, China=0, France=0, Spain=4, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Health care systems

Question: In which of the following strategies developed by the Ministry of Health, Social Services and Equality is the objective included to promote the development of individualized nursing care plans adapted to the needs of each patient, specifying that reference be made, at least, to the following aspects: Prevention of falls and associated injuries, Prevention of pressure ulcers, Prevention of healthcare-associated infections, Safety in the physical restraint of patients who require it, Prevention of malnutrition and dehydration, especially in elderly patients, Prevention of aspiration, and Prevention and control of pain in adults and children, attending to the needs and preferences of patients and their caregivers?

- a: Health Promotion and Disease Prevention Strategy of the National Health System.
- b: Strategy for the Management of Chronicity of the National Health System.
- c: Patient Safety Strategy of the National Health System.
- d: Diabetes Strategy of the National Health System.

Answer: c
Rationale for score: The question explicitly tests knowledge of a specific national strategy developed by the Ministry of Health, Social Services and Equality for the National Health System, which refers directly to the Spanish healthcare system and its national guidelines.

2023B044_en (JMedQA)

Source language: ja | **Region-specific score**: 4
Predominant country: Japan
Country scores: Japan=4, United States=0, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Health care systems

Question: A 38-year-old man presents with complaints of loss of appetite, lower back pain, and back pain. Present illness: He has experienced heartburn, loss of appetite, and lower back pain for the past 6 months. Two months ago, he developed back pain and dyspnea, and visited a local clinic. An upper gastrointestinal endoscopy revealed advanced gastric cancer in the body of the stomach, and he was admitted for further evaluation and treatment. Although he has lower back pain, his ADL is independent and he is ambulatory. Past history: None. Social history: Lives with his wife and two elementary school-aged daughters. Occupation: Company employee, sales position involving frequent travel. Smoking: 20 cigarettes/day since age 20, but quit 2 weeks ago. Alcohol: Occasional. Family history: None. Physical examination: Conscious. Height 176 cm, weight 63 kg. Temperature 36.4 °C. Pulse 92/min, regular. BP 120/64 mmHg. Respiratory rate 18/min. SpO2 97% (room air). No abnormalities in palpebral or bulbar conjunctiva. Heart and breath sounds are normal. Abdomen is flat and soft; liver and spleen are not palpable. Laboratory findings: RBC 3.9 million, Hb 9.1 g/dL, Ht 39%, WBC 6,300, Platelets 260,000. Biochemistry: Total protein 5.0 g/dL, Albumin 3.0 g/dL, Total bilirubin 0.8 mg/dL, AST 16 U/L, ALT 18 U/L, LD 184 U/L (ref 120-245), ALP 200 U/L (ref 38-113), Creatinine 1.0 mg/dL, Na 141 mEq/L, K 4.0 mEq/L, Cl 101 mEq/L, Ca 8.5 mg/dL. Chest X-ray showed scattered masses in both lungs. Abdominal CT and lumbar MRI showed metastatic bone tumors in the L1 and L3 vertebrae. No liver metastasis. The patient was informed of his condition in detail, and chemotherapy and radiation therapy were planned. The patient consulted, saying, 'I am worried about quitting my job because I don't want to cause trouble for the company. The treatment costs are high, and I am extremely worried about how I will support my two elementary school-aged daughters in the future.' The patient received chemotherapy and radiation therapy for the metastatic

lumbar bone tumors, was discharged home after 1 month, and returned to work. He is scheduled to continue outpatient chemotherapy. Which system can this patient utilize?

- a: Long-term care insurance
- b: Disability pension
- c: High-cost medical expense benefit system
- d: Medical expense support system for independence (self-reliance support medical care)
- e: Designated intractable disease medical expense subsidy system

Answer: c
Rationale for score: The question requires knowledge of specific social security and health insurance systems unique to Japan, such as the High-cost medical expense benefit system, Long-term care insurance age requirements, and the Designated intractable disease medical expense subsidy system.

MEDQA000308_en (MedQA)

Source language: en | **Region-specific score**: 4
Predominant country: United States
Country scores: Japan=0, United States=4, China=0, France=0, Spain=0, Sweden=0, Russia=0, Other=0
Clinical category: Social medicine / Health care systems

Question: A 67-year-old man is brought to the emergency department with sudden onset of slurred speech. The patient's eldest daughter says that he couldn't move his arm during the episode. Past medical history is significant for hypertension and a hemorrhagic stroke 6 years ago, which was treated surgically. After admission, the patient gets increasingly worse and loses consciousness. A noncontrast CT of the head shows a subarachnoid hemorrhage (SAH). The patient is taken to the operating room, where the SAH evacuated, and the vessel is repaired. Postoperatively, the patient is unresponsive. Several days later, a repeat CT of the head shows an enlarging aneurysm in a different vessel in the brain. The patient's daughter is asked to consent to a non-emergent, life-saving operation since the patient is in a comatose state. She does not consent to the operation even though the physician stresses it would be life-saving. Upon further discussion, the physician finds out that the patient was abusive and neglectful to his children, and the daughter wants "to live in peace." The patient has no written advanced directive. Which of the following is the most appropriate course of action?

- a: Follow the daughter's wishes to withhold the operation
- b: Refer this case to the court
- c: Contact another family member for consent
- d: Report the daughter to the police

Answer: c
Rationale for score: The question relies heavily on the US legal and ethical framework for surrogate decision-making, where next-of-kin are legally required to provide or refuse consent for incapacitated patients. In many other jurisdictions (such as the UK), physicians make medical decisions based on the patient's best interests, and family members do not have the automatic legal authority to consent or refuse life-saving treatment unless they hold a specific power of attorney.

MSWE001423_en (MedQA-SWE)

Source language: sv | **Region-specific score**: 4
Predominant country: Sweden
Country scores: Japan=0, United States=0, China=0, France=0, Spain=0, Sweden=4, Russia=0, Other=0
Clinical category: Clinical medicine / Cardiovascular disorders

Question: A 70-year-old man was brought by ambulance after fainting and becoming unresponsive while sitting at the kitchen table. According to his wife, he had no seizures or urinary incontinence. The man is otherwise healthy except for hypertension treated with a calcium channel blocker. He does not smoke and has no known neurological disease. He reports that for over a week he has had sudden dizzy spells, occurring without warning and unrelated to physical activity. Physical examination: General condition: Awake and alert, able to describe the event. Blood pressure: 120/80 mmHg in both arms. Heart and lungs: No murmurs or adventitious sounds. The ECG recording showed third-degree AV block. He requires a DDD pacemaker. While observing the patient, you notice something you have not seen before: his jugular veins fill irregularly and then collapse, unrelated to pulse or respiration. The irregular filling and collapse of the jugular veins, unrelated to pulse or respiration, is due to the right atrium occasionally contracting against a closed tricuspid valve. Isoprenaline can be used to normalize the heart rhythm. A pacemaker is implanted without complications. Follow-up ECG shows that the pacemaker is functioning correctly, sensing P-waves in the atrium and subsequently pacing the right ventricle. The patient is now to be discharged home. He is a forest owner and holds a driver's license for both passenger cars and heavy timber trucks. What advice should you give regarding driving?

- a: The man may drive both passenger cars and timber trucks immediately
- b: The man may drive a passenger car immediately but should wait 2 weeks before driving the timber truck
- c: After pacemaker implantation, the man may not drive either a passenger car or a timber truck
- d: The man must book an appointment for new driving tests before he is allowed to drive
- e: The man may never again drive a passenger car or a timber truck

Answer: b
Rationale for score: The question asks for specific advice regarding driving restrictions for passenger cars and heavy trucks after pacemaker implantation. These restrictions are governed by national transport authority regulations, making the correct answer highly dependent on country-specific legal and public health rules.