# UNCERTAINTY AND THE MEDICAL INTERVIEW

## TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS

Jakob D. Havtorn

DTU Compute
Department of Applied Mathematics and Computer Science

# Outline of Part

- Introduction

## Healthcare

*Healthcare is the improvement of health via the prevention, diagnosis, treatment, amelioration or cure of disease, illness, injury, and other physical and mental impairments in people.*

# Medical dialogue

General practitioner

Emergency room

Health insurance

Hospital

Specialist doctor

Psychiatrist

Physiotherapist

Emergency services

# Medical dialogue

1. Failure of communication is a leading cause of medical error contributing to two out of three adverse events [4].
2. Between 9% and 16.6% of all hospital admissions had preventable adverse outcomes (AU, UK, NZ, DK) [15].

# Types of dialogue

- Context: Controlled / Chaotic
- Domain: Specialized / General
- Person: Nurse, doctor, midwife, caregiver, psychiatrist, insurance professional
- Purpose: Triaging, diagnosis, treatment, follow-up, documentation, coding, billing

# Part I

## Unsupervised Out-of-Distribution Detection

# Outline of Part

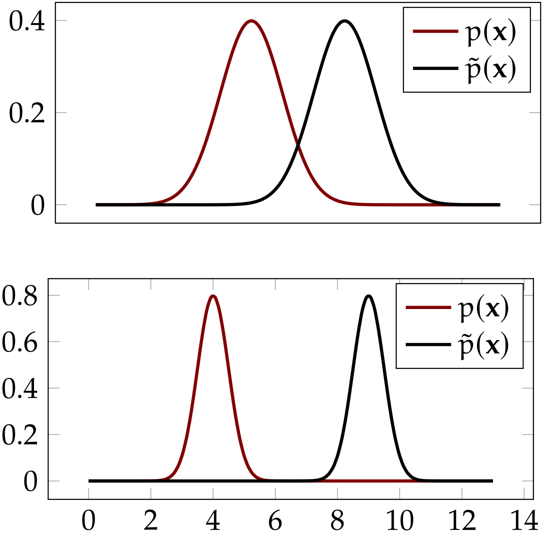
- Out-of-distribution detection
- Latent variable models
- Identifying the issue
- The $\mathcal{L}^{>k}$ likelihood bound
- Likelihood ratio

# Defining OOD detection

Out-of-distribution (OOD) detection is about enabling models to distinguish the training data distribution $p(\mathbf{x})$ from any other distribution $\tilde{p}(\mathbf{x})$.

We are concerned with doing this on a per-observation basis, i.e. answering the question:

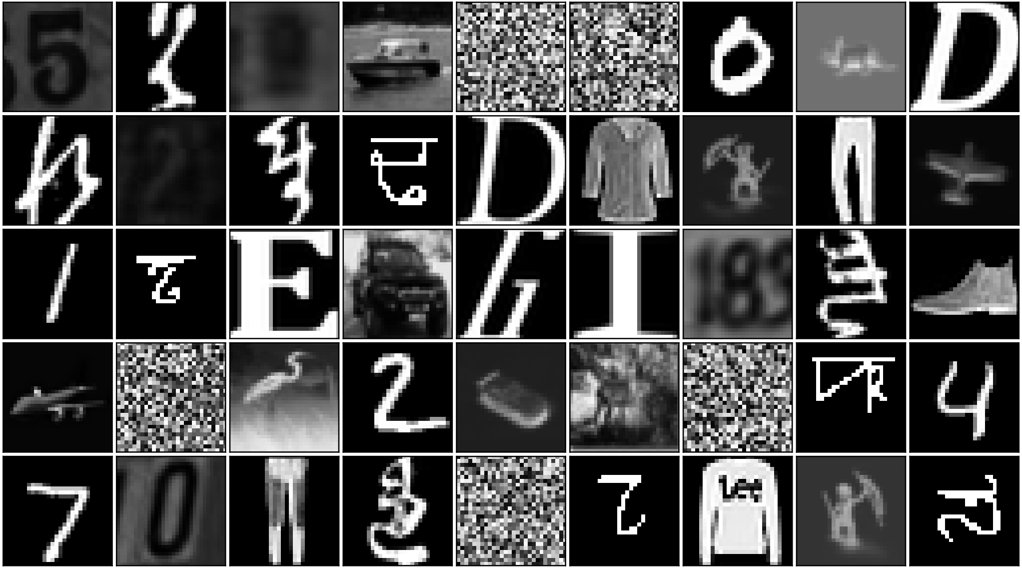
> "Was $\mathbf{x}$ sampled from $p(\mathbf{x})$ or not?"



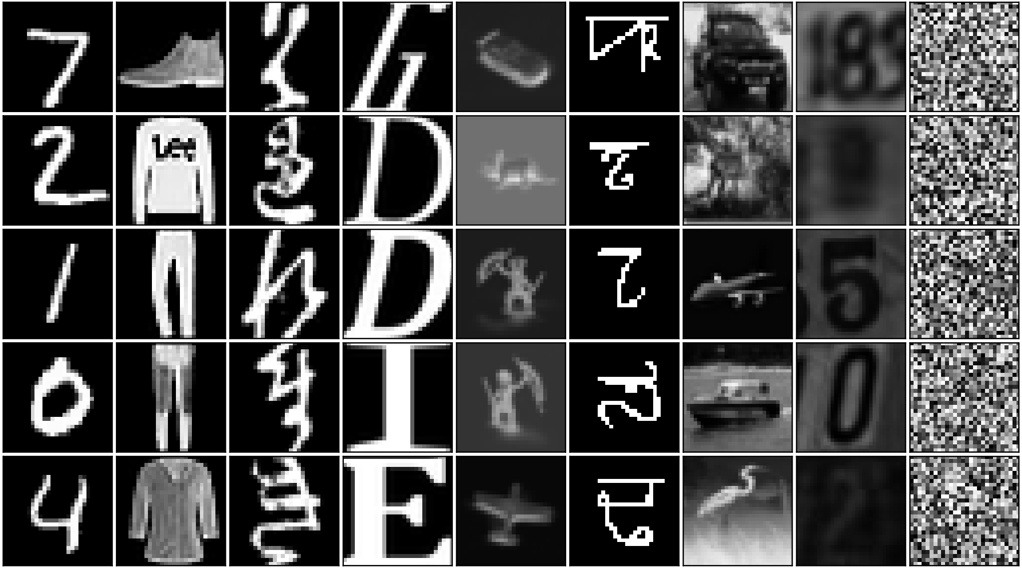
# Problem and Contributions

- Deep generative models often fail at OOD detection task when using their likelihood estimate as the score function [9] by, perhaps surprisingly, assigning **higher likelihoods** to the OOD data.
- Contributions:
  - We present a fast and fully unsupervised method for OOD detection competitive with the state-of-the-art
  - We provide evidence that out-of-distribution detection fails due to learned low-level features that generalize across datasets.

# In distribution?



# Out of distribution?



# Hierarchical VAE

We choose the hierarchical VAE as our model [2, 3].

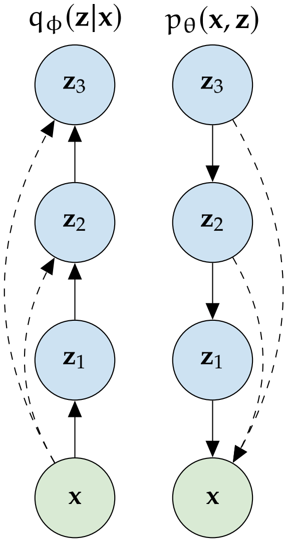
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z})d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$$

Specifically we use

1. a three-layered hierarchical VAE with bottom-up inference and deterministic skip-connections for both inference and generation.

$$\begin{aligned} \text{Generative model:} \quad & p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}_1)p_\theta(\mathbf{z}_1|\mathbf{z}_2)p(\mathbf{z}_3), \\ \text{Inference model:} \quad & q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{z}_2|\mathbf{z}_1)q_\phi(\mathbf{z}_3|\mathbf{z}_2). \end{aligned}$$

2. a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [8].



## What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction likelihood}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}_{\text{regularization penalty}} \,. \tag{1}$$

The first term is high if the data is well-explained by $\mathbf{z}$.

The second term we can rewrite as,

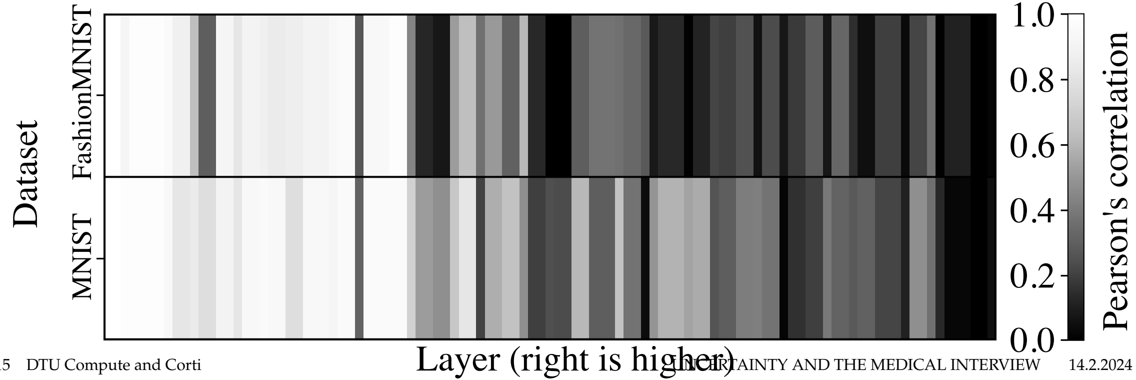
$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] . \tag{2}$$

The absolute log-ratios grow with $\dim(\mathbf{z}_i)$ since the log probability terms are computed by summing over the dimensionality of $\mathbf{z}_i$.

# What do the lowest latent variables code for?

Absolute Pearson correlations between data representations in all layers of the inference network of a hierarchical VAE trained on FashionMNIST and of another trained on MNIST.

Correlation computed between the representations of the two different models given the same data, FashionMNIST (top) and MNIST (bottom).



# An alternative likelihood bound, $\mathcal{L}^{>k}$

An alternative version of the ELBO that only partially uses the approximate posterior can be written as [8]

$$\mathcal{L}^{>k}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}>k)q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}_{>k})}{q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\right] \tag{3}$$

Here, we have replaced the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely

$$p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})$$

This bound uses the conditional prior for the lowest latent variables in the hierarchy.

# Likelihood ratios

We can use our new bound to compute the score used in a standard likelihood ratio test [1].

$$\mathrm{LLR}^{>k}(\mathbf{x}) \equiv \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x})\,. \tag{4}$$

We can inspect what this likelihood-ratio measures by considering the exact form of our bounds.

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right), \\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{KL}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right)\,. \end{aligned} \tag{5}$$

In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} \mathrm{LLR}^{>k}(\mathbf{x}) = &-D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &+ D_{KL}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right)\,. \end{aligned} \tag{6}$$

# Importance sampling the ELBO

The well-known importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples, $S \to \infty$ [5],

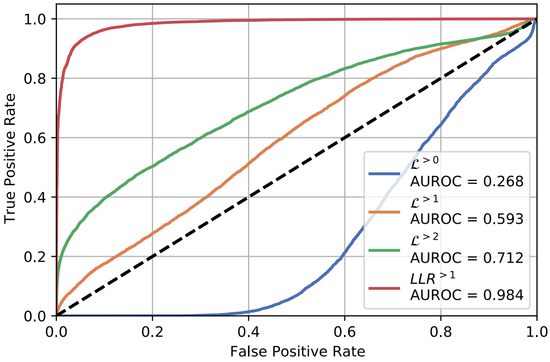
$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N} \sum_{s=1}^{S} \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] \leqslant \log p_\theta(\mathbf{x}) \,, \tag{7}$$

Consequently, by importance sampling the ELBO, the associated KL-divergence associated vanishes and our likelihood ratio reduces to the KL-divergence associated with $\mathcal{L}^{>k}$.

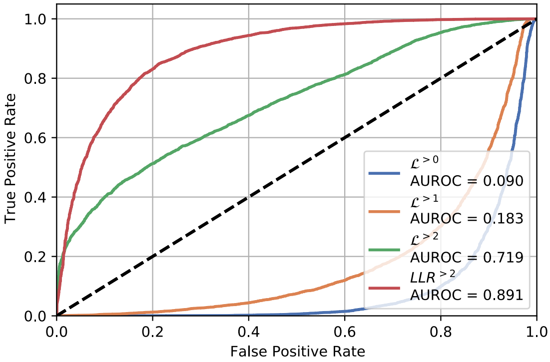
$$\mathrm{LLR}_S^{>k}(\mathbf{x}) \to D_{\mathrm{KL}}(p(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))\,. \tag{8}$$

We can now see that $\mathrm{LLR}_S^{>k}(\mathbf{x})$ performs OOD detection based on the top-most latent variables.

# Results with $\mathrm{LLR}^{>k}$




(a) FashionMNIST HVAE evaluated on MNIST




(b) CIFAR10 BIVA evaluated on SVHN

# Results with $\mathrm{LLR}^{>k}$

The score has good performance across many datasets.

| OOD dataset | Metric | AUROC↑ | AUPRC↑ | FPR80↓ |
|---|---|---|---|---|
| **Trained on CIFAR10** | | | | |
| SVHN | $\mathrm{LLR}^{>2}$ | 0.811 | 0.837 | 0.394 |
| CIFAR10 | $\mathrm{LLR}^{>1}$ | 0.469 | 0.479 | 0.835 |
| **Trained on SVHN** | | | | |
| CIFAR10 | $\mathrm{LLR}^{>1}$ | 0.939 | 0.950 | 0.052 |
| SVHN | $\mathrm{LLR}^{>1}$ | 0.489 | 0.484 | 0.799 |

| OOD dataset | Metric | AUROC↑ | AUPRC↑ | FPR80↓ |
|---|---|---|---|---|
| **Trained on FashionMNIST** | | | | |
| MNIST | $\mathrm{LLR}^{>1}$ | 0.986 | 0.987 | 0.011 |
| notMNIST | $\mathrm{LLR}^{>1}$ | 0.998 | 0.998 | 0.000 |
| KMNIST | $\mathrm{LLR}^{>1}$ | 0.974 | 0.977 | 0.017 |
| Omniglot28x28 | $\mathrm{LLR}^{>2}$ | 1.000 | 1.000 | 0.000 |
| Omniglot28x28Inverted | $\mathrm{LLR}^{>1}$ | 0.954 | 0.954 | 0.050 |
| SmallNORB28x28 | $\mathrm{LLR}^{>2}$ | 0.999 | 0.999 | 0.002 |
| SmallNORB28x28Inverted | $\mathrm{LLR}^{>2}$ | 0.941 | 0.946 | 0.069 |
| FashionMNIST | $\mathrm{LLR}^{>1}$ | 0.488 | 0.496 | 0.811 |
| **Trained on MNIST** | | | | |
| FashionMNIST | $\mathrm{LLR}^{>1}$ | 0.999 | 0.999 | 0.000 |
| notMNIST | $\mathrm{LLR}^{>1}$ | 1.000 | 0.999 | 0.000 |
| KMNIST | $\mathrm{LLR}^{>1}$ | 0.999 | 0.999 | 0.000 |
| Omniglot28x28 | $\mathrm{LLR}^{>1}$ | 1.000 | 1.000 | 0.000 |
| Omniglot28x28Inverted | $\mathrm{LLR}^{>1}$ | 0.944 | 0.953 | 0.057 |
| SmallNORB28x28 | $\mathrm{LLR}^{>1}$ | 1.000 | 1.000 | 0.000 |
| SmallNORB28x28Inverted | $\mathrm{LLR}^{>1}$ | 0.985 | 0.987 | 0.000 |
| MNIST | $\mathrm{LLR}^{>2}$ | 0.515 | 0.507 | 0.792 |

# Part II

## Medical Applications

# Outline of Part

- A Retrospective Study on Machine Learning-Assisted Stroke Recognition for Medical Helpline Calls

## Stroke

- Stroke is a leading cause of disability and death worldwide [12, 7, 6].
- Effective treatment is time-sensitive, and an optimal outcome is more likely when treatment is administered within the first four and a half hours from stroke onset [11, 10].
- The gateway to ambulance transport and hospital admittance is through prehospital telehealth services, including emergency medical call centers, nurse advice call lines, and out-of-hours health services.
- In the pre-hospital setting, the use of mobile stroke units has made it possible to deliver advanced treatment faster [13, 14].
- As the mobile stroke unit is only dispatched to patients with a suspected stroke, the impact of mobile stroke unit is directly influenced by accurate call-taker recognition of stroke [13, 14].
- Call-takers who can rapidly and accurately recognize stroke are therefore crucial in facilitating prompt care in both pre-hospital and in-hospital settings.

# Thank you for your attention

# Bibliography I

[1] Adolf Buse. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982), pp. 153–157.

[2] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. Banff, AB, Canada, 2014. arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114.

[3] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. International Conference on Machine Learning. Vol. 32. Beijing, China: PMLR, Jan. 16, 2014, pp. 1278–1286. URL: http://proceedings.mlr.press/v32/rezende14.pdf (visited on 08/12/2018).

# Bibliography II


[4] Amy J Starmer et al. "Changes in Medical Errors after Implementation of a Handoff Program". In: *New England Journal of Medicine* 371.19 (2014), pp. 1803–1812.

[5] Yuri Burda, Roger Grosse, and Ruslan R. Salakhutdinov. "Importance Weighted Autoencoders". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. San Juan, Puerto Rico, 2016, p. 8. URL: `https://arxiv.org/abs/1509.00519` (visited on 10/04/2017).

[6] Mira Katan and Andreas Luft. "Global Burden of Stroke". In: *Seminars in Neurology*. Vol. 38. 02. Thieme Medical Publishers, 2018, pp. 208–211.

# Bibliography III


[7] Hmwe Hmwe Kyu et al. "Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 359 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (2018), pp. 1859–1922.

[8] Lars Maaløe et al. "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Vancouver, Canada, Feb. 6, 2019, pp. 6548–6558. URL: `http://arxiv.org/abs/1902.02102` (visited on 03/19/2019).

# Bibliography IV

[9] Eric Nalisnick et al. "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. New Orleans, LA, USA, 2019. arXiv: `1810.09136`. URL: `http://arxiv.org/abs/1810.09136` (visited on 10/02/2019).

[10] Guillaume Turc et al. "European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke". In: *Journal of Neurointerventional Surgery* 11.8 (2019), pp. 535–538.

[11] Eivind Berge et al. "European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke". In: *European Stroke Journal* 6.1 (2021), pp. I–LXII.

# Bibliography V


[12] GBD 2019 Stroke Collaborators et al. "Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019". In: *The Lancet Neurology* 20.10 (2021), pp. 795–820. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(21)00252-0.

[13] Praveen Hariharan et al. "Mobile Stroke Units: Current Evidence and Impact". In: *Current Neurology and Neuroscience Reports* 22.1 (2022), pp. 71–81.

[14] Babak B Navi et al. "Mobile Stroke Units: Evidence, Gaps, and next Steps". In: *Stroke* 53.6 (2022), pp. 2103–2113.

[15] Niki Carver, Vikas Gupta, and John E. Hipskind. "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514. URL: http://www.ncbi.nlm.nih.gov/books/NBK430763/ (visited on 02/13/2024).