# Beyond the Firewall: Can AI Medical Data Be Hacked?

Rasit Dinc

*Rasit Dinc Digital Health & AI Research*

Published: February 20, 2023 | Medical Imaging AI

DOI: 10.5281/zenodo.17997604

## Abstract

The integration of Artificial Intelligence AI into healthcare promises a revolution in diagnostics, treatment planning, and patient management. From sophisti...

The integration of Artificial Intelligence (AI) into healthcare promises a revolution in diagnostics, treatment planning, and patient management. From sophisticated image analysis to personalized medicine, AI systems process vast quantities of highly sensitive patient data. This raises a critical question for professionals and the public alike: **Can AI medical data be hacked?** The answer is unequivocally yes, but the threat extends far beyond traditional data breaches. The vulnerabilities inherent in AI systems introduce new, complex attack vectors that target not just the data itself, but the very logic and output of the AI models.

## The Foundation of Risk: Data at Rest and in Transit

The most familiar form of hacking involves the compromise of the underlying infrastructure where AI training data and patient records are stored. AI models are only as secure as the databases they are trained on and the electronic health record (EHR) systems they interact with. A typical healthcare data breach, often resulting from phishing, ransomware, or unpatched vulnerabilities, can expose millions of patient records. While this is a conventional cybersecurity risk, the sheer volume and sensitivity of the data required for AI—including genetic sequences, high-resolution medical images, and detailed clinical histories—make these systems a prime target. The financial impact is significant, with healthcare consistently reporting the highest average cost per data breach across all industries [1].

## The Evolving Threat: Adversarial Attacks on Model Integrity

A more insidious threat to AI medical data is the **adversarial attack**, which targets the integrity of the AI model's decision-making process. These attacks exploit the subtle mathematical weaknesses in machine learning algorithms, particularly deep neural networks. An adversarial example is a piece of input data that has been intentionally and minimally perturbed to cause the model to make a confident, yet incorrect, prediction [2].

In a clinical setting, this can have life-threatening consequences. Researchers have demonstrated that an imperceptible, carefully calculated perturbation—often invisible to the human eye—can be added to a medical image, such as an MRI or a photograph of a mole. This manipulation can flip a diagnosis from "benign" to "malignant" with high confidence, or vice versa [2]. The motivation for such attacks is not always financial; it could be to commit fraud, sabotage a competitor's model, or even inflict harm. Furthermore, the difficulty of updating complex medical IT systems and the inherent ambiguity in some medical diagnoses make it challenging to detect these subtle, malicious perturbations in real-time.

## The New Frontier: Prompt Injection and Logic Manipulation

The rise of large language models (LLMs) and vision-language models (VLMs) in clinical decision support has introduced a third, cutting-edge vulnerability: **prompt injection**. This attack vector manipulates the AI's instructions or logic by embedding malicious, often hidden, commands within the input data.

For example, a prompt injection attack could involve embedding sub-visual text commands within a medical image or a patient's electronic chart. When the VLM processes this input, the hidden command overrides the model's safety protocols, causing it to generate a harmful or incorrect output, such as a misdiagnosis in oncology [3]. This is a fundamental security flaw because the attack is performed by simply interacting with the model's input interface, without needing to access its internal parameters or the underlying infrastructure. The goal is to exfiltrate private data, evade model guardrails, or corrupt clinical decisions.

## Mitigating the Risk: A Multi-Layered Defense

Addressing the hackability of AI medical data requires a comprehensive, multi-layered defense strategy that goes beyond conventional firewalls. This strategy must include:

1. **Data Security:** Implementing robust encryption, access controls, and de-identification techniques for all training and inference data. 2. **Model Robustness:** Developing and deploying AI models that are specifically trained to be resilient against adversarial examples and data poisoning. 3. **Input Validation:** Establishing strict validation and sanitization protocols for all data inputs, especially for LLMs and VLMs, to detect and neutralize prompt injection attempts.

The security landscape for AI in healthcare is rapidly evolving. The question is no longer *if* AI medical data can be hacked, but *how* the industry will adapt to defend against these increasingly sophisticated and model-specific threats. For more in-depth analysis on this topic, the resources at [www.rasitdinc.com] (https://www.rasitdinc.com) provide expert commentary.

**

## *References*

*[1] IBM. (2023).* Cost of a Data Breach Report 2023*. [https://www.ibm.com/security/data-breach/] (https://www.ibm.com/security/data-breach/) [2] Finlayson, S. G., et al. (2019). Adversarial attacks on medical machine learning: Emerging vulnerabilities demand new conversations.* Science*, 363(6433), 1287-1289. [https://pmc.ncbi.nlm.nih.gov/articles/PMC7657648/] (https://pmc.ncbi.nlm.nih.gov/articles/PMC7657648/) [3] Clusmann, J., et al. (2025). Prompt injection attacks on vision language models in oncology.* Nature Communications*, 16(1), 1239. [https://www.nature.com/articles/s41467-024-55631-x] (https://www.nature.com/articles/s41467-024-55631-x)

---

**Rasit Dinc Digital Health & AI Research**

https://rasitdinc.com