Unlocking the Potential of AI in Healthcare: Overcoming Clinical Data Access Challenges

Introduction

The integration of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare holds immense potential to revolutionize patient care, improve diagnostics, and enhance treatment plans. However, a significant roadblock remains: access to real-world Electronic Health Record (EHR) and Electronic Medical Record (EMR) data. Enterprises and customers are understandably cautious about sharing Protected Health Information (PHI) and Personally Identifiable Information (PII) due to privacy concerns.
This white paper explores innovative strategies to overcome these challenges and unlock the full potential of AI in healthcare.

This white paper explores four innovative strategies to overcome these data access challenges:

1. Generating high-quality synthetic EHR/EMR data that maintains the statistical properties of real data while preserving patient privacy
2. De-identifying and obfuscating sensitive information in real EHR/EMR data to enable its use in model training
3. Leveraging federated learning to train models on decentralized data across multiple organizations without the data leaving their systems
4. Clinical Domain specific LLM can be fine-tuned on small labeled EMR / EHR datasets for and still achieve high performance by leveraging the medical knowledge from pre-training. This reduces the need for large annotated datasets.
5.

We also discuss methods to rigorously evaluate the effectiveness of synthetic data in terms of data fidelity, utility for analytics tasks, and privacy preservation. By adopting these cutting-edge approaches, healthcare organizations can unlock the power of their data assets for AI applications while ensuring regulatory compliance and safeguarding patient confidentiality.

The Promise and Challenge of Healthcare AI

The increasing digitization of health records and the rapid advancement of AI/ML techniques have opened up exciting possibilities to derive insights from the vast troves of real-world data contained in EHRs and EMRs. From early disease detection and diagnosis to personalized treatment recommendations and drug discovery, AI models trained on EHR data have the potential to transform many aspects of clinical practice and medical research.
However, a major impediment is the sensitive nature of the PHI and PII contained in medical records. Privacy regulations like HIPAA and GDPR place strict constraints on the use and disclosure of such data. This leads to a catch-22 situation - while large, diverse, and

representative datasets are needed to train robust and generalizable AI models, healthcare data is often siloed and inaccessible due to valid privacy concerns. Navigating this tradeoff between data utility and privacy is a key challenge.

Strategies to Overcome Data Access Challenges

1. Synthetic EHR/EMR Data

Synthetic data generation is a transformative approach that addresses privacy concerns while enabling robust AI model training. Techniques like Generative Adversarial Networks (GANs) can create high-quality synthetic data that mirrors the statistical properties of real data. This method ensures patient confidentiality while providing valuable data for AI development.

- **EHR-Safe Framework:** The EHR-Safe framework has demonstrated the ability to generate realistic synthetic EHR data, maintaining high fidelity and privacy. This approach allows for robust model training without compromising patient confidentiality.
    - Learn more about EHR-Safe: EHR-Safe
- **EHR-M-GAN:** This model synthesizes mixed-type timeseries EHR data, capturing the complex temporal dynamics in patient trajectories. It has shown superior performance in generating high-fidelity synthetic data, enhancing AI algorithm development in resource-limited settings.
    - Read about EHR-M-GAN: EHR-M-GAN

Using synthetic data, AI models can be trained without privacy risks. The generated data can also augment real datasets to improve model robustness.

2. Data Obfuscation/De-identification

Data obfuscation and de-identification involve removing or masking sensitive information in EHR/EMR data. This process allows the use of real data for AI model training while safeguarding patient privacy. However, it must be meticulously executed to ensure the data remains useful for its intended purpose.

- **SaaS based De-identification Tools:** SaaS based offerings from Google and AWS such as -
    - Google Healthcare API does at the dataset level as well as FHIR object level - https://cloud.google.com/healthcare-api/docs/how-tos/deidentify
    - Amazon Comprehend Medical's DetectPHI operation that lets detect PHI information - https://docs.aws.amazon.com/comprehend-medical/latest/dev/textanalysis-phi.html
- **Advanced De-identification Tools:** Tools developed by institutions like the Mayo Clinic have shown impressive results in transforming identifiable data into plausible surrogates, enabling large-scale data use while safeguarding privacy.
    - Explore automated de-identification: Automated De-identification

- Differential privacy techniques can be applied when aggregating EHR data to provide mathematical guarantees of privacy.

However, de-identification must be balanced against preserving data utility. Over-sanitization can strip out important signals and impact model performance.

3. Federated Learning

Federated learning offers a decentralized approach to model training. By training models across multiple organizations without the data ever leaving their systems, federated learning leverages real-world data while maintaining privacy. This method has been successfully implemented in various healthcare settings, allowing for the creation of comprehensive and diverse models.

- Truveta Language Model (TLM): The TLM is a multi-modal LLM trained on a vast collection of medical records, achieving over 90% accuracy on clinical tasks. This model exemplifies the potential of federated learning in healthcare.
    - Discover the Truveta Language Model: Truveta Language Model

4. Domain specific LLM

Adapting BERT for the medical domain: Several studies have developed domain-specific BERT models by pre-training them on large medical datasets. Examples include BioBERT (pre-trained on biomedical literature), ClinicalBERT (pre-trained on clinical notes), BlueBERT (pre-trained on PubMed abstracts and clinical notes), and Med-BERT (pre-trained on structured EHR data). These models capture the unique language and concepts in medical data better than general-domain BERT.
Domain-specific BERT models help overcome this by enabling transfer learning - they can be fine-tuned on small labeled datasets for specific medical NLP tasks and still achieve high performance by leveraging the medical knowledge from pre-training. This reduces the need for large annotated datasets.

- **BioBERT (pre-trained on biomedical literature):**
    - Paper: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506
    - GitHub repo: https://github.com/dmis-lab/biobert
    - Pre-trained weights: https://github.com/naver/biobert-pretrained
- ClinicalBERT (pre-trained on clinical notes):
    - Hugging Face model card: https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
    - GitHub repo: https://github.com/kexinhuang12345/clinicalBERT
    - Paper: https://arxiv.org/abs/1904.05342
- BlueBERT (pre-trained on PubMed abstracts and clinical notes):
    - Hugging Face model cards:
      https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-24_H-1024_A-16
      https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-12_H-768_A-12
    - GitHub repo: https://github.com/ncbi-nlp/bluebert
    - Paper: https://arxiv.org/abs/1906.05474

- Med-BERT (pre-trained on structured EHR data):
    - Paper: https://www.nature.com/articles/s41746-021-00455-y
    - GitHub repo: https://github.com/ZhiGroup/Med-BERT
    - Pre-print: https://arxiv.org/abs/2005.12833

Measuring the Effectiveness of Synthetic EHR Data

When using synthetic EHR data for developing AI applications, it's important to evaluate its quality and effectiveness. This involves assessing three key aspects:

- **Fidelity:** Fidelity measures how closely the synthetic data resembles the real data. If the synthetic data has high fidelity, it should be difficult to distinguish it from real data. This can be evaluated using statistical similarity metrics like the Kolmogorov-Smirnov (KS) test and visualization techniques like t-SNE.
- **Utility:** Utility assesses how well the synthetic data can be used for specific analytical tasks, such as training machine learning models. This involves comparing the performance of models trained on synthetic data to those trained on real data using metrics like AUC-ROC and precision-recall curves. It also includes evaluating the synthetic data's ability to support downstream tasks like disease prediction and clinical decision support.
- **Privacy:** Privacy metrics ensure that the synthetic data does not reveal any real patient's identity. This can be evaluated by assessing the risk of membership inference attacks, where an adversary tries to determine if a specific individual's data was used to train the synthetic data model. Lower risk scores indicate better privacy preservation. Another approach is to ensure that the synthetic data generation process adheres to differential privacy standards, which provide mathematical guarantees of privacy.

4. Comprehensive Evaluation Frameworks

Several comprehensive evaluation frameworks have been proposed to assess synthetic EHR data across multiple dimensions:

- SynEval: A multi-faceted evaluation framework that integrates data fidelity, utility, and privacy evaluation with a comprehensive set of metrics. This framework provides a holistic assessment of synthetically generated data and helps in making informed decisions regarding their deployment in real-world scenarios.
- Systematic Benchmarking Framework: This framework involves generating multiple synthetic datasets using different models, assessing each dataset on various metrics, and ranking the models based on their performance. This approach helps in identifying the most suitable synthetic data generation model for specific use cases.

5. Call to Action

Navigating the challenges of accessing real-world EHR/EMR data requires innovative strategies. By leveraging synthetic data, advanced de-identification techniques, and federated learning, we can unlock the full potential of AI in healthcare while ensuring patient privacy. These solutions not only address privacy concerns but also enable the development of robust AI models that can drive forward the future of medical AI.

Citations

https://www.linkedin.com/pulse/challenges-opportunities-implementing-ai-electronic-health-records-235xf
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6821018/
https://www.mckinsey.com/industries/life-sciences/our-insights/real-world-data-quality-what-are-the-opportunities-and-challenges
https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01768-6
https://www.sciencedirect.com/science/article/pii/S2405632423000331
https://www.medicaldevice-network.com/features/lack-of-trust-in-ai-led-electronic-health-systems-remains/
https://www.decentriq.com/article/6-challenges-in-advancing-real-world-data-use-in-healthcare-and-their-solutions
https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-022-08882-7
https://www.scnsoft.com/healthcare/ehr/artificial-intelligence
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9437583/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/
https://www.researchgate.net/publication/347447047_Integration_of_Artificial_Intelligence_in_electronic_health_records_Impacts_and_challenges
https://demigos.com/blog-post/ehr-emr-interoperability/
https://www.sciencedirect.com/science/article/pii/S1532046421003099
https://www.covingtondigitalhealth.com/2020/07/ehr-interoperability-public-health-benefits-privacy-considerations/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10718098/
https://dl.acm.org/doi/10.1145/3653297
https://www.tebra.com/theintake/practice-operations/legal-and-compliance/privacy-concerns-with-ai-in-healthcare
https://academic.oup.com/bioinformatics/article/36/4/1234/5566506
https://huggingface.co/bionlp/bluebert_pubmed_uncased_L-24_H-1024_A-16
https://www.nature.com/articles/s41746-021-00455-y