1 AI Ethics in Healthcare

1.1 Healthcare Data Privacy

The realm of data ethics encompasses a delicate and particularly sensitive domain, with healthcare data uniquely positioned as a result of being sensitive in nature. Unlike other types of data such as financial or social media, medical records contain permanent and unchangeable personal information regarding an individual (such as diagnosis, family history, and/or genetic makeup) that, when publicized, could have a severe and adverse effect on an individual (discrimination or employment-related issues, such as wrongful denial of insurance coverage). The regulations that govern this area are outlined in two major jurisdictions or regulatory frameworks.

1.1.1 HIPAA (US)

Protected Health Information (PHI) is any identifiable individual health information that is maintained by a covered entity, as defined by U.S. HIPAA laws. The HIPAA regulations specify several types of administrative, physical, and technical safeguards, such as access controls, audit trails, and encryption during transmission of PHI. In addition, there is a requirement to perform a formal risk assessment prior to allowing an AI system to process PHI and to notify the appropriate parties of a data breach within 60 days or sooner depending on state notification laws.

1.1.2 GDPR (EU)

The GDPR (EU) regards health data as a “special category of data” and therefore it needs explicit and distinctive consent plus a legal basis under Articles 7 and 8 for processing. Article 17’s right to be forgotten creates tension with AI systems because an individual can request the destruction of their data, however, the model(s) built with that data will have retained the learned parameters of the destroyed data and there is thus no true destruction of the data; i.e., technical or otherwise, renders the software less certain than the individual might think about whether their data has been destroyed or not.

1.1.3 Anonymization challenges

The challenges of anonymizing data complicate both frameworks. Just removing identifying information from your data like your subjects' names, dates and other personal identifiers will not protect your subjects' identities. Research has shown that using only three of these types of pseudo-identifying factors (zipcode, birthdate, and sex) can be used to identify 87% of the

population in the U.S. In some cases, identifying factors for patients diagnosed with rare diseases will include either the rare disease itself or an unusual combination of medications or geographic location that can be identified from the "de-identified" data collected by hospitals and health care systems.

1.2 Algorithmic Bias in Medical AI

There are three primary types of bias in the literature that have been identified as influencing the validity of medical models. The first type of bias is caused by the under-representation of minority populations in the clinical training data. The clinical datasets used to create these models are usually generated at large urban academic hospitals that provide services to primarily older white, insured patients. Therefore, the models created from these datasets will be poorly calibrated for individuals of other minority races. The second source of bias is caused by the geographic and/or socio-economic distribution of patients represented in the clinical datasets, resulting in very few patients without regular access to healthcare being included in the electronic health record system and therefore in the data collected to create predictive models.

In a circulated example of a real-life case study of the Optum Risk Stratification Algorithm, according to research published in Obermeyer et al.'s (2019) study published at science, the algorithm uses healthcare costs as an indirect measure of health need. Unintentionally, the algorithm incorporated a racial bias when determining whether to assign a lower risk score to black patients than to white patients with the same degree of illness, therefore effectively denying 46% of black patients' high-risk care management program access relative to white patients of equal illness severity.

The two strategies for mitigating this risk: one at the technical level (i.e. implementing demographic parity constraints and using disaggregated evaluation metrics for protected subgroups) requires that all models undergo fairness-aware training and this must be part of the evaluation process prior to clinical use; while, the second strategy is that AI tools must be approved by a diverse group of clinically trained individuals who will review their proposed use, along with ethicists, patient advocates and community members that represent the affected population prior to procurement. The second strategy creates accountability for AI providers that purely technical solutions cannot do.

1.3 Ethical Decision Framework

- Informed Consent — Has meaningful consent been obtained?
- Data Minimization — Are we collecting only what is necessary?
- Bias Audit — Has the model been evaluated across all demographic subgroups?
- Transparency — Can the model's decision be explained to a clinician?
- Accountability — Is there a named human responsible for each AI decision?
- Ongoing Monitoring — Is model performance reviewed after deployment?