# Synthetic Data Gets Real

**Proskauer in Life Sciences Blog** on **June 29, 2021**

As we mentioned in the early days of the pandemic, COVID-19 has been accompanied by a rise in cyberattacks worldwide. At the same time, the global response to the pandemic has accelerated interest in the collection, analysis, and sharing of data – specifically, patient data – to address urgent issues, such as population management in hospitals, diagnoses and detection of medical conditions, and vaccine development, all through the use of artificial intelligence (AI) and machine learning. Typically, AIML churns through huge amounts of real world data to deliver useful results. This collection and use of that data, however, gives rise to legal and practical challenges. Numerous and increasingly strict regulations protect the personal information needed to feed AI solutions. The response has been to anonymize patient health data in time consuming and expensive processes (HIPAA alone requires the removal of 18 types of identifying information). But anonymization is not foolproof and, after stripping data of personally identifiable information, the remaining data may be of limited utility. This is where synthetic data comes in.

A synthetic dataset comprises artificial information that can be used as a stand-in for real data. The artificial dataset can be derived in different ways. One approach starts with real patient data. Algorithms process the real patient data and learn patterns, trends, and individual behaviors. The algorithms then replicate those patterns, trends, and behaviors in a dataset of artificial patients, such that – if done properly – the synthetic dataset has virtually the same statistical properties of the real dataset. Importantly, the synthetic data cannot be linked back to the original patients, unlike some de-identified or anonymized data, which have been vulnerable to re-identification attacks. Other approaches involve the use of existing AI models to generate synthetic data from scratch, or the use of a combination of existing models and real patient data.

While the concept of synthetic data is not new, it has recently been described as a promising solution for healthcare innovation, particularly in a time when secure sharing of patient data has been particularly challenged by lab and office closures. Synthetic data in the healthcare space can be applied flexibly to fit different use cases, and they can be expanded to create more voluminous datasets.

Synthetic data's other reported benefits include the elimination of human bias and the democratization of AI (i.e., making AI technology and the underlying data more accessible). Critically too, regulations governing personal information, such as HIPAA, the EU General Data Protection Regulation (GDPR), and the California Consumer Privacy Act (CCPA), may be read to permit the sharing  and processing of original patient data (subject to certain obligations) such that the resulting synthetic datasets may carry less privacy risk.

Despite the potential benefits, the creation and use of synthetic data has its own challenges. First, there is the risk that AI-generated data is so similar to the underlying real data that real patient privacy is compromised. Additionally, the reliability of synthetic data is not yet firmly established. For example, it is reported that no drug developer has yet relied on synthetic data for a submission to the U.S. Food and Drug Administration because it is not known whether that type of data will be accepted by FDA. Perhaps most importantly, synthetic data is susceptible to adjustment, for good or ill. On the one hand, dataset adjustments can be used to correct for biases imbedded in real datasets. On the other, adjustments can also undermine trust in healthcare and medical research.

As synthetic data platforms proliferate and companies increasingly engage those services to develop innovative solutions, care should be exercised to guard against the potential privacy and reliability risks.

* * * * * *

Check back here for updates on this topic. For more information on how to comply with privacy regulations or defend against regulatory enforcement actions, please contact your Proskauer lawyers.

View Original

**Related Professionals**

- **Colin G. Cabral**
  Partner