International Journal of Information and Cybersecurity

Int. J. Inf. Cybersec.-2020

A Stagewise Framework for Implementing AI Privacy Models to Address Data Privacy and Security in Cancer Care

Shivansh Khanna
School of Information Sciences, University of Illinois at Urbana-Champaign

Ishank Khanna
Sri Aurobindo Medical College and PG Institute, Indore, India

Shraddha Srivastava
School of Information Sciences, University of Illinois at Urbana-Champaign

Vedica Pandey
Sri Aurobindo Medical College and PG Institute, Indore, India

Abstract

Cancer patient data is not only highly sensitive but also incredibly diverse, containing genetic, clinical, and personal information. This diversity poses challenges in privacy and security, which traditional privacy models may not adequately address. This study introduces a stagewise framework for implementing AI privacy models designed to address the challenges of data privacy and security in cancer care. The framework unfolds across six stages. The initial stage, data collection, focuses on data anonymization and masking. This step is for safeguarding personally identifiable information (PII), where sensitive details are replaced with fictional yet plausible data in preliminary datasets. As the framework progresses to the data aggregation stage, it uses federated learning and privacy-preserving record

linkage (PPRL). These methods enable the integration of decentralized data from varied sources, such as different hospitals, without compromising individual identities. In the data analysis stage, differential privacy and secure multi-party computation (SMC) are employed. These techniques ensure that the analysis of aggregated data does not reveal individual patient details. Stage four of model training emphasizes using synthetic data and homomorphic encryption, necessary for training AI models with reduced privacy risks and enabling training on encrypted data. Data Sharing/Reporting, the fifth stage, includes k-anonymity and homomorphic encryption to maintain the confidentiality of shared or reported data. The final stage, Ongoing Monitoring and Updating, reiterates the continuous application of differential privacy and federated learning, essential for updating models with new data without infringing on privacy.

Keywords: *Cancer Care, Data Privacy, Privacy Models, Security, Stagewise Framework*

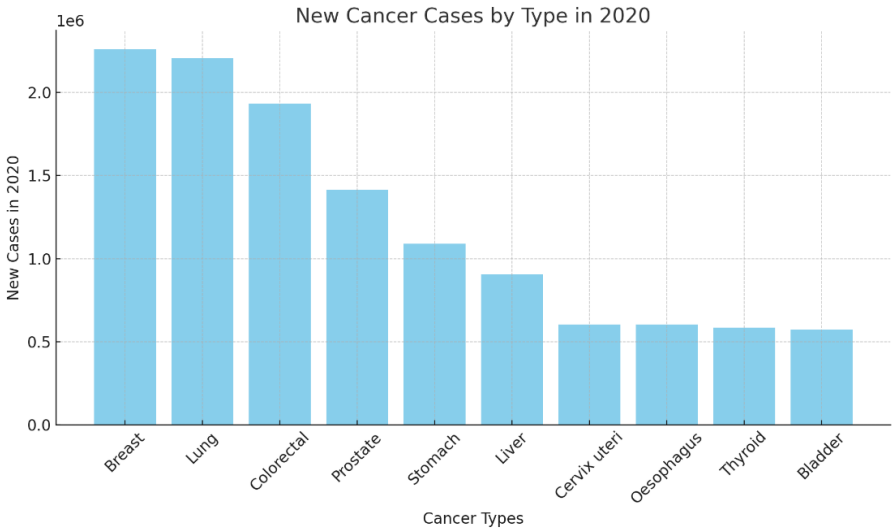
Introduction

In 2020, the United States witnessed a significant impact of cancer, with 1,603,844 new cases reported and 602,347 deaths due to the disease. This translates to 403 new cases and 144 deaths per 100,000 people as shown in figure 1. The data from 2020 is the most recent available regarding cancer incidence. This escalating global burden underscores the urgent need for cancer prevention, a paramount public health challenge in the 21st century. On a global scale, breast and lung cancers emerged as the most prevalent, each accounting for 12.5% and 12.2% respectively of the total new cases diagnosed in 2020. Following closely, colorectal cancer was the third most common, with 1.9 million new cases, making up 10.7% of the total new cases. It is important to note that all statistical figures presented exclude non-melanoma skin cancer from the total percentage of all cancers.

Cancer patient data contain a range of information types, including genetic, clinical, and personal details. Genetic data, for example, is essential for understanding the specific mutations that might be driving a patient's cancer. This information can guide the selection of targeted therapies that are more likely to be effective based on the patient's unique genetic makeup. However, genetic data also raises privacy concerns, as it can reveal information about an individual's risk for various diseases, potentially impacting not just the patient but also their relatives.

Clinical data in cancer patient records includes diagnosis, treatment history, imaging, and lab test results. This information is used by healthcare providers to track the progress of the disease and the effectiveness of treatments. It also forms the basis for much of the ongoing research into new cancer treatments and understanding of the disease. However, the management of such data presents challenges in terms of ensuring accuracy, privacy, and security. The diverse nature of clinical data, reflecting the wide range of cancer types and patient responses to treatments, adds to the complexity of its use and management.

***Figure 1**. Cancer cases in the United States by types*



Data source: wcrf.org

The diversity of data types and sources presents unique challenges in maintaining privacy and security, which are not adequately addressed by traditional privacy models (Al-Issa et al., 2019; Tschider, 2019). This diversity spans a wide range of data, including detailed medical histories, genetic information, treatment responses, and lifestyle factors. Each type of data has its own set of privacy concerns (Saxena, 2020). For instance, genetic data not only affects the individual but also has implications for their family members, potentially revealing hereditary cancer risks. Medical histories and treatment records are highly sensitive and can impact a

patient's life beyond their health, such as their insurability and employment prospects. Traditional privacy models in healthcare are often designed around more straightforward patient-provider confidentiality and may not fully encompass the complexities and sensitivities involved in cancer-related data, especially when this data is shared for research or with third-party healthcare providers.

Additionally, the methods through which cancer-related data is collected, stored, and shared further complicate privacy and security concerns. With the integration of digital technologies in healthcare, such as electronic health records (EHRs), wearable health devices, and telemedicine, there is a continuous and pervasive collection of patient data. This omnipresent data collection leads to massive data repositories that, if not properly secured, are vulnerable to breaches and unauthorized access. Traditional privacy models in healthcare often rely on consent and limited data sharing principles, but these may fall short in the face of complex, interconnected digital systems where data flows are continuous and multifaceted. The security of cancer patient data is also a critical issue, as cyber threats become more sophisticated and capable of exploiting vulnerabilities in healthcare IT systems. Moreover, the international sharing of cancer-related data for research and collaboration purposes introduces additional challenges, given the variability in data protection laws across different countries. This global dimension of data sharing in cancer care calls for a nuanced understanding of privacy and security that transcends traditional models and addresses the intricate and sensitive nature of cancer patient data in a digital and interconnected world.

The escalating concerns around data privacy in various sectors have necessitated the establishment of robust regulatory frameworks, such as the General Data Protection Regulation (GDPR) in the European Union, the China Cyber Security Law, and the California Consumer Privacy Act (CCPA) in the United States. These regulations represent a significant shift towards empowering individuals with greater control and rights over their personal data. For instance, the GDPR, one of the most comprehensive data protection laws globally, grants individuals several critical rights including the right to access their data, the right to have incorrect data corrected, the right to have their data erased under certain conditions, and the right to object to certain types of processing, including automated decision-making and profiling. These provisions aim to address the power imbalance between data collectors and individuals, ensuring that personal data is handled transparently and with due respect for privacy. Similarly, laws like the CCPA and China's Cyber

Security Law have put forth their own sets of rules and guidelines, mandating businesses to disclose their data collection practices, obtain consent from consumers before data collection, and provide options for consumers to opt-out of data sharing.

However, while these regulations are a step in the right direction, they also present challenges in implementation and compliance, especially for organizations operating on a global scale. The GDPR, for instance, has extraterritorial applicability, meaning it applies to any organization dealing with EU residents' data, regardless of where the organization is based. This global reach requires companies around the world to reassess and often overhaul their data handling practices to ensure compliance. Moreover, the differences between various privacy laws, like the CCPA and GDPR, create a complex legal landscape for international businesses. For example, the CCPA includes specific provisions about selling personal information, which are not as explicitly addressed in the GDPR (Barrett, 2019). Additionally, China's Cyber Security Law places a strong emphasis on data localization, which poses challenges for multinational companies that are accustomed to storing and processing data globally. These complexities highlight the evolving nature of data privacy regulations and the need for continuous adaptation and vigilance by organizations to keep pace with these changes and adequately protect individuals' privacy rights.

Data security remains a concern in various sectors accentuated by recent high-profile data leaks and the increasing risk of inference attacks, where sensitive information can be deduced from seemingly innocuous data. The complexity of these security challenges is further amplified in environments that involve the transfer of large volumes of data across multiple institutions, such as in healthcare, finance, and research. Recognizing these challenges, innovative approaches to enhance data security are emerging. One notable example is the concept of ‘Federated Learning’ introduced by Google in 2016. This approach represents a paradigm shift in data handling for machine learning and AI applications. Instead of the traditional method of pooling data into a central repository for model training, federated learning allows for the training of algorithms on decentralized devices or servers. This means sensitive data no longer needs to be transferred to a central institution, significantly reducing the risks associated with data transmission and storage. This method not only addresses privacy concerns but also allows for the development of robust, versatile models that benefit from a diverse range of data sources without compromising data security.

Healthcare organizations manage vast amounts of sensitive data, essential for delivering effective care. This data includes patient medical records, treatment plans, and personal health information, all of which require stringent security measures. Despite the critical nature of this data, many healthcare organizations often grapple with inadequate technical support and minimal security infrastructure. This gap in data security makes the healthcare industry vulnerable to data breaches, which are frequently reported and publicly disclosed. Cyber attackers employ sophisticated data mining methods to extract sensitive information, leading to significant privacy violations and other associated risks. Implementing effective security measures in healthcare is a complex and ongoing process. As security technologies evolve, so do the tactics used by cybercriminals to bypass these controls. The continuous advancement of hacking techniques raises the stakes in the cybersecurity arms race, requiring healthcare institutions to perpetually update and strengthen their security postures to safeguard patient data against ever-evolving threats.

Existing Privacy and security models

Differential Privacy

Differential Privacy aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its entries (Chavez et al., 2019). Mathematically, a randomized function (K) gives (ϵ) - differential privacy if for all datasets $(D_1) and (D_2)$ differing on at most one element, and all ($S \subseteq \text{Range}(K)$),

$$[\mathbf{Pr}[\boldsymbol{K}(\boldsymbol{D_1}) \in \boldsymbol{S}] \leq \boldsymbol{e^{\epsilon}} \times \mathbf{Pr}[\boldsymbol{K}(\boldsymbol{D_2}) \in \boldsymbol{S}]]$$

Here, (ϵ) is a non-negative parameter that controls the privacy guarantee, with smaller values providing stronger privacy.

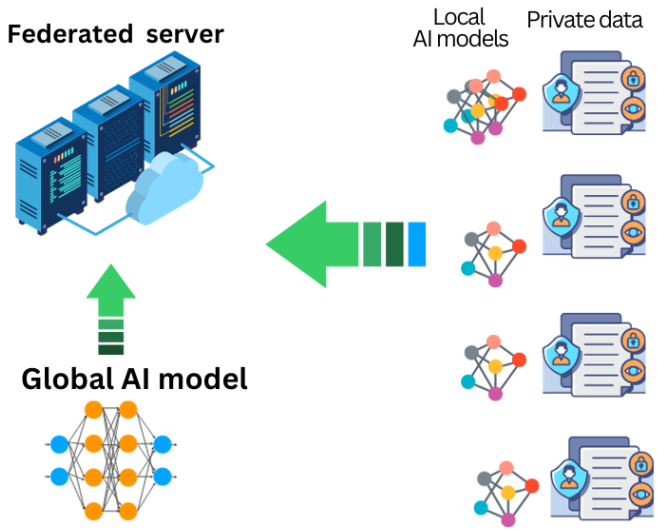
Federated Learning

Federated Learning is a machine learning setting where the goal is to train a model across multiple decentralized edge devices or servers holding local data samples, without exchanging them (Bai et al., 2019). While there's no single mathematical formula that defines Federated Learning, it generally involves solving optimization problems like:

$$[\min_{\theta}\left[F(\theta)=\sum_{k=1}^{K}p_kF_k(\theta)\right]]$$

where $(F_k(\theta))$ is the local objective function for the $(k)-th$ device, (θ) represents the model parameters, and (p_k) is the relative size of the (k)-th dataset.

Figure 2. *A centralized server and connections to various local servers or devices (such as hospitals) which hold local data sets. These local entities train models on their data and then send model updates to the central server where they are aggregated. This approach maintains data privacy and security since the raw data does not leave its original location.*



Source: Author

Homomorphic Encryption

Homomorphic Encryption is a form of encryption that allows computation on ciphertexts, generating an encrypted result which, when decrypted, matches the

result of operations performed on the plaintext. The mathematical specifics depend on the scheme, but generally, if ($\mathbf{Enc}(\cdot)$) denotes the encryption function, and ($\oplus$) and ($\otimes$) represent some operations, then for plaintexts (a) and (b),

$$[\mathbf{Dec}\big(\mathbf{Enc}(\boldsymbol{a}) \oplus \mathbf{Enc}(\boldsymbol{b})\big) = \boldsymbol{a} + \boldsymbol{b}]$$

$$[\mathbf{Dec}\big(\mathbf{Enc}(\boldsymbol{a}) \otimes \mathbf{Enc}(\boldsymbol{b})\big) = \boldsymbol{a} \times \boldsymbol{b}]$$

Secure Multi-party Computation (SMC)

SMC allows parties to jointly compute a function over their inputs while keeping those inputs private (Bayatbabolghani & Blanton, 2018; Du & Atallah, 2001). In its simplest form, if two parties have private inputs (x) and (y), they want to compute a function ($f(x, y)$) without revealing (x) or (y) to each other. The computation is done using cryptographic protocols, often involving homomorphic encryption or secret sharing.

Data Anonymization

This is a process in which personally identifiable information from data sets is removed or obscured. The purpose is to ensure that the individuals whom the data describe remain anonymous. This is crucial for protecting privacy and complying with data protection regulations. Data anonymization techniques can include data masking, pseudonymization, generalization, and more, with the aim of preserving the data's utility while safeguarding individual identities.

Synthetic Data Generation

This involves creating artificial data that is not derived from real-world events but is generated by algorithms or simulations. Synthetic data is designed to be similar to actual data in terms of statistical properties. This technique is often used when actual data is limited, sensitive, or unavailable. It's valuable for training machine learning models, testing systems, and ensuring privacy, as it does not correspond to real individuals.

Privacy-Preserving Record Linkage (PPRL)

PPRL is a technique used to integrate or link records from different sources while preserving the privacy of the individuals whose data is being linked (Boyd et al., 2015). This process involves matching records that relate to the same entity across different databases without revealing sensitive information contained in the records.

Techniques used in PPRL include hashing, encryption, and the use of secure multi-party computation to ensure that personally identifiable information is not disclosed during the linkage process (Verykios & Christen, 2013). PPRL is critical in areas like healthcare and research, where data from different sources needs to be combined for analysis but privacy must be maintained (Vatsalan et al., 2017).

K-anonymity

K-anonymity is a property attained by a dataset when the information for each person contained in the release cannot be distinguished from at least ($\boldsymbol{k-1}$) individuals whose information also appears in the release (Vijayarani & Tamilarasi, 2010). A dataset is said to satisfy k-anonymity if for every record there are ($\boldsymbol{k-1}$) other records that are indistinguishable from it in terms of certain 'quasi-identifier' attributes (Domingo-Ferrer et al., 2006; El Emam et al., 2009).

Stagewise framework

these AI privacy models can be applied in a stagewise manner in in cancer care, when dealing with sensitive datasets. Applying these models in stages allows for a addressing different aspects of data security and privacy at various points in the data processing and analysis pipeline.

Stage 1: Data Collection

In the initial stage of data collection in cancer care, data anonymization involves systematically stripping away personally identifiable information (PII) from patient records. This process transforms the data into a state where individual patients cannot be identified directly or indirectly. For instance, names, addresses, and other direct identifiers are removed or altered. Additionally, care is taken with indirect identifiers, like specific medical procedures or unique treatment regimes, which might be combined with external data to re-identify a patient. In cancer care research, where patient data is crucial for understanding disease patterns and treatment outcomes, anonymization ensures that researchers can analyze comprehensive datasets without compromising individual privacy. However, the challenge lies in doing so without losing the data's utility for research purposes. Therefore, sophisticated techniques are employed to maintain the balance between data utility and privacy, ensuring that the anonymized data remains valuable for scientific discovery and the advancement of medical knowledge.

Table 1. Challenges in maintaining the privacy and security of sensitive patient data in various aspects of cancer care

Data Type	Description	Privacy and Security Challenges
Patient History and Demographics	Basic information such as age, gender, family medical history, and personal medical history.	Risk of identity theft and discrimination due to the sensitivity of personal information.
Diagnostic Information	Data from tests and procedures used to diagnose cancer, like imaging tests, biopsies, and blood tests.	Potential for misuse if accessed without authorization due to detailed health information.
Cancer Specific Information	Details about the type of cancer, its stage, grade, location, and, where applicable, genetic markers and hormone receptor status.	Risk of genetic discrimination and privacy breaches due to the highly sensitive and specific health data.
Treatment Information	Types of treatments received (surgery, chemotherapy, radiation, etc.) and the patient's response to these treatments.	Sensitive healthcare data could be misused in contexts like insurance and employment.
Side Effects and Complications	Recording of any side effects or complications from treatment, and overall well-being of the patient.	Health data that could be stigmatizing if mishandled or disclosed improperly.
Follow-up Data	Ongoing data collection post-treatment to monitor health, looking for signs of recurrence or managing chronic issues.	Long-term storage of data increases the risk of unauthorized access and data breaches.
Patient-Reported Outcomes	Information on the patient's quality of life, including aspects such as physical, emotional, and social well-being.	Contains personal information that could lead to social stigma if disclosed.
Research and Clinical Trials Data	For patients in clinical trials, specific data collection according to the study protocol.	Data sharing in research contexts raises concerns about re-identification and consent, especially in genetic research.

Data masking, on the other hand, involves replacing sensitive data elements with fictitious but realistic counterparts. In the context of cancer care, data masking is employed in the early stages of data handling, where the exact patient details are not necessary. For example, real patient demographic information might be replaced with fictional but demographically similar data. This technique is beneficial when data needs to be shared with parties such as software developers working on healthcare applications or third-party analysts conducting preliminary research. The primary goal is to make the data realistically usable for operational and developmental purposes without exposing actual patient details.

Stage 2: Data Aggregation

In Stage 2: Data Aggregation, the application of Federated Learning and Privacy-Preserving Record Linkage (PPRL) offers a practical solution to the challenges of handling sensitive health data across multiple sources. Federated Learning operates

by distributing the AI model training process across various data sources, such as different hospitals or research centers. Each participating institution trains a local model on its dataset and only shares model updates, not the sensitive data itself. This method significantly enhances data privacy, as the raw patient data remains within the confines of its original location. For example, in a study involving multiple cancer centers, each center can develop its part of the model based on its patient data. These individual model updates are then combined to improve the overall AI model, without the need for direct access to patient data from other centers.

Table 2. Various aspects of data aggregation in cancer care and the corresponding challenges in maintaining privacy and security

Data Aggregation Aspect	Description	Privacy and Security Challenges
Consolidation of Data Sources	Combining data from various sources such as patient records, diagnostic tests, treatment information, and follow-up data.	Risk of mismatching or misinterpreting data; challenges in ensuring consistency and accuracy across different data systems.
Standardization and Normalization	Ensuring data from different sources is compatible and standardized for analysis. This includes normalizing varying formats, scales, and terminologies.	Difficulty in maintaining data integrity and consistency during the standardization process; potential loss of data detail or context.
Data Cleaning and Quality Checks	Removing inaccuracies and duplicates to ensure the reliability of the aggregated data set.	Risk of data distortion or loss during cleaning processes; balancing thoroughness of cleaning with preservation of data integrity.
Data Integration	Integrating various types of data (clinical, genomic, imaging, etc.) to create a comprehensive view of patient information and treatment outcomes.	Complexity in integrating diverse data types while preserving privacy and confidentiality; managing large volumes of sensitive data.
Anonymization and De-identification	Removing or encrypting identifiable information to protect patient privacy while allowing data to be used for research and analysis.	Challenges in completely anonymizing data without losing critical information; risk of re-identification in certain cases.

Data Warehousing and Storage	Storing the aggregated data in a secure and accessible manner, often using data warehouses.	Ensuring security and confidentiality in data storage; risk of data breaches and unauthorized access in large, centralized data repositories.
Data Accessibility and Sharing	Making aggregated data available for healthcare providers, researchers, and sometimes patients, while controlling access rights and permissions.	Balancing the need for data accessibility with the need to protect sensitive information; managing permissions and access controls.
Compliance with Regulations	Adhering to legal and ethical standards such as HIPAA in the U.S., GDPR in Europe, and other local data protection regulations.	Ensuring continuous compliance with evolving legal and ethical standards; managing differences in regulations across regions.

Privacy-Preserving Record Linkage (PPRL) complements Federated Learning by enabling the linkage of patient records from different databases without exposing individual identities. PPRL techniques, such as hashing or encryption, transform patient identifiers in such a way that records can be matched for comprehensive analysis while preserving anonymity. For instance, when combining cancer patient records from different regional databases, PPRL ensures that researchers can identify and analyze trends across broader populations without the risk of revealing personal information. This is especially important in cases where a patient's data might be spread across multiple institutions, providing a more complete picture for analysis without compromising privacy.

The combination of Federated Learning and PPRL in Data Aggregation not only adheres to privacy regulations but also opens doors for more robust and diverse datasets in AI-driven cancer research. This approach allows for a more comprehensive analysis of data from varied populations and geographies, enhancing the potential for developing more accurate and generalizable AI models in cancer care. For example, by aggregating and analyzing data from diverse demographic groups, AI models can be trained to recognize patterns and treatment responses that are specific to subpopulations, leading to more personalized and effective cancer treatments. This stage, therefore, is not about exaggerating the importance of the methods, but rather about leveraging their specific capabilities to address the unique challenges posed by the sensitive nature of health data in cancer research.

Stage 3: Data Analysis

The integration of advanced data protection techniques like Differential Privacy and Secure Multi-party Computation (SMC) into cancer care represents a advancement in the way we handle sensitive medical information. These techniques are considered important in cancer care, where patient data is not only extremely sensitive but also immensely valuable for research and treatment development.

Table 3. key elements of the data analysis stage in cancer care, along with the associated privacy and security challenges

Data Analysis Aspect	Description	Privacy and Security Challenges
Statistical Analysis	Application of statistical methods to understand trends, correlations, and patterns in the aggregated data.	Risk of misinterpretation of data leading to privacy concerns, especially if the analysis results are made public.
Predictive Modeling	Using data to create models that predict outcomes such as treatment responses or disease progression.	Potential for biases in the model affecting patient privacy, especially if the model inadvertently reveals sensitive personal information.
Genomic Data Analysis	Analysis of genetic data to understand cancer genetics, treatment responses, and predispositions.	Genetic data is highly sensitive; there's a risk of unauthorized access leading to privacy breaches and genetic discrimination.
Treatment Efficacy Evaluation	Assessing the effectiveness of different cancer treatments based on patient data.	Challenges in anonymizing patient data while maintaining the integrity and usefulness of the analysis.
Comparative Effectiveness Research	Comparing the effectiveness, benefits, and harms of different treatment options.	Privacy risks in aggregating and analyzing data from diverse populations and treatment scenarios.
Machine Learning and AI Analysis	Using advanced algorithms and machine learning techniques to uncover insights from complex and large datasets.	AI and machine learning models can inadvertently expose sensitive information; challenges in ensuring the models do not compromise patient privacy.

Pharmacogenomics	Studying how genes affect a person's response to drugs, to develop effective, safe medications and doses that will be tailored to a person's genetic makeup.	Risk of privacy breaches with genetic data used in pharmacogenomic studies; potential for misuse of genetic information.
Real-World Evidence Analysis	Analysis of data derived from a variety of sources, including electronic health records, insurance claims, and patient-generated data, to inform treatment and policy decisions.	Challenges in maintaining privacy when analyzing data from diverse and potentially identifiable sources; balancing data utility with confidentiality requirements.

Differential Privacy technique attempts to safeguard patient confidentiality during data analysis. In cancer care, vast amounts of data are collected from various sources including clinical trials, patient records, and genetic information. This data is used by for researchers and healthcare providers, as it can lead to breakthroughs in understanding and treating cancer. However, there's a significant risk of compromising patient privacy when such large datasets are analyzed. This is where Differential Privacy comes in. It provides a way to gain insights from the aggregated data while ensuring that individual patient details cannot be inferred. By adding controlled noise to the data or using algorithms that limit the impact of any single data point, Differential Privacy ensures that the results of the analysis are useful from a research and treatment perspective, without revealing any individual's sensitive information. This approach is beneficial in cancer research, where detailed patient data can reveal insights about disease progression and treatment efficacy.

Secure Multi-party Computation (SMC) further enhances data security when multiple entities are involved in cancer research and treatment. In many instances, effective cancer care and research involve collaboration between different hospitals, research institutions, and pharmaceutical companies. Each of these entities holds a piece of the puzzle — be it unique patient data, proprietary treatment techniques, or specialized research findings. SMC allows these diverse parties to compute collaboratively on combined data sets without actually exposing their individual data to each other. For instance, if two hospitals are trying to determine the effectiveness of a new treatment, they can jointly analyze their data using SMC techniques to reach a conclusion, all without actually sharing the sensitive patient data with one

another. This method is incredibly powerful in cancer care, where collaboration is key, but the privacy and security of patient data are paramount. By enabling secure, collaborative analysis, SMC paves the way for more comprehensive and effective cancer treatment strategies while rigorously protecting patient privacy.

Stage 4: Model Training

Synthetic Data Generation is employed to mitigate the privacy risks associated with the use of real patient data. In the context of cancer care, where data sensitivity is high, synthetic data serves as a viable alternative. This technique involves generating artificial data that mimics the statistical properties of real patient data. The advantage here is twofold: it allows for the training of AI models without compromising patient privacy, and it alleviates concerns about data confidentiality breaches. However, it's important to note that while synthetic data can be valuable for training purposes, it may not always capture the complexity and nuances of real patient data, which could impact the model's accuracy and applicability in real-world scenarios.

Table 4. Challenges associated with the model training stage in cancer care,

Model Training Aspect	Description	Privacy and Security Challenges
Selection of Training Data	Identifying and selecting relevant datasets for training predictive models, such as patient records, treatment outcomes, and genetic data.	Ensuring representative and unbiased data selection without compromising individual patient privacy.
Feature Engineering	Creating predictive features from raw data, which involves choosing which aspects of the data are important for the model to learn.	Risk of including sensitive features that could lead to the identification of individuals in the dataset.
Model Choice and Development	Deciding on the type of machine learning model (e.g., neural networks, decision trees) that is best suited for the cancer care application.	Selecting models that do not overfit to sensitive or identifiable features in the training data.
Algorithm Training	Training the chosen model on the selected data, which involves adjusting the model parameters to improve accuracy and performance.	Ensuring the training process does not compromise data privacy when using external or cloud-based computing resources.
Validation and Testing	Evaluating the model's performance using a separate	Maintaining data confidentiality during validation, especially when

	dataset to ensure its accuracy and reliability.	using external datasets for testing model robustness.
Bias and Fairness Assessment	Assessing the model for potential biases (e.g., towards certain patient demographics) and ensuring fairness in predictions.	Identifying and mitigating biases that could lead to privacy concerns or unequal treatment of certain patient groups.
Model Interpretability	Ensuring the model's decisions and predictions can be understood and interpreted by human experts, especially in critical healthcare decisions.	Balancing the complexity of models with the need for interpretability to avoid misinterpretation that could affect patient privacy and treatment decisions.
Deployment Readiness Evaluation	Assessing the model's readiness for deployment in a clinical setting, including its integration with existing healthcare systems.	Ensuring that the model's deployment does not introduce new privacy vulnerabilities in clinical care settings.

Homomorphic Encryption, on the other hand, is applied when the use of real patient data is indispensable. In some cancer research scenarios, the specificity and richness of real patient data are crucial for the development of accurate and effective AI models. Homomorphic Encryption allows for the encryption of patient data in such a way that it can still be used for computations and training AI models. The data remains encrypted throughout the process, ensuring that patient privacy is maintained even when real data is in use. However, this approach has its limitations, as it can be computationally intensive and may lead to longer processing times, which could be a drawback in time-sensitive research or treatment scenarios.

Stage 5: Data Sharing/Reporting

K-anonymity is a technique applied in cancer care data management to enhance the privacy of individuals within larger datasets. This method transforms and generalizes the data in such a way that any given record is indistinguishable from at least $(k-1)$ other records concerning certain identifying attributes. In practical terms, when a dataset is k-anonymized, an individual's data cannot be isolated from at least $(k-1)$ other individuals' data, effectively masking their identity within the group. For example, in a dataset of patient treatment outcomes, k-anonymity would ensure that any specific patient's information is indistinct and blends with the data of other patients. This approach is used when datasets are shared for research or reporting purposes. It allows for the useful dissemination of data while protecting

individuals from being identified, even if an intruder has access to other sources of information. K-anonymity addresses the risk of re-identification in cancer care data, a field where patient confidentiality is as important as the insights drawn from the data.

Table 5. data sharing and reporting stage in cancer care, with a focus on the associated privacy and security challenges without speculative elements

Data Sharing/Reporting Aspect	Description	Privacy and Security Challenges
Research Findings Dissemination	Publishing results from cancer research, including clinical trials, observational studies, and meta-analyses.	Avoiding the inadvertent release of identifiable patient information in research publications.
Clinical Data Reporting	Reporting treatment outcomes, side effects, and patient experiences to healthcare entities and oversight bodies.	Ensuring patient anonymity and data security when transferring data electronically.
Data Exchange Between Institutions	Sharing patient data for collaborative treatment or referrals between hospitals, clinics, and specialists.	Securing data during transfer and ensuring it is only accessed by authorized personnel.
Public Health Data Reporting	Providing anonymized data to public health agencies for tracking cancer trends, outcomes, and public health planning.	Preventing re-identification of individuals from large datasets used in public health studies.
Interaction with Insurance Entities	Reporting patient data for insurance claims processing and coverage determination.	Safeguarding sensitive health information against unauthorized access or use in insurance decisions.
Patient Access to Their Data	Patients accessing their own health data through electronic health records or patient portals.	Protecting online platforms from breaches, ensuring patients can securely access only their own data.
Collaborative Research Data Sharing	Exchanging data with other research entities for joint studies or analysis.	Implementing secure data sharing agreements and practices to protect patient

		confidentiality in a research context.
Regulatory Compliance Reporting	Submitting data to regulatory bodies for compliance with healthcare regulations and standards.	Ensuring data is shared in a compliant manner, respecting regulations like HIPAA or GDPR, depending on the jurisdiction.

Homomorphic encryption represents a significant advancement in the way sensitive data is handled. This is also true in cancer care. This form of encryption enables data to be encrypted and then processed or analyzed while still encrypted, producing an encrypted result that, when decrypted, matches the result of operations performed on the unencrypted data. In cancer care, where data is often shared for analysis or used to train AI models, homomorphic encryption allows for the sharing of results or model outputs without ever exposing the raw data. For instance, an AI model could be trained on encrypted patient data, and the results of this training, such as predictive models for treatment outcomes, could be shared across institutions without compromising the confidentiality of the underlying patient data. This method is useful for collaborative research and analysis in cancer care, as it provides a way for multiple entities to benefit from shared insights while maintaining the utmost data privacy.

Stage 6: Ongoing Monitoring and Updating

Differential privacy, when applied in the context of updating models with new data in cancer care, offers a practical solution for maintaining patient privacy. This technique is relevant as healthcare data is regularly updated with new patient information, which could potentially compromise individual privacy if not handled carefully. Differential privacy works by adding a certain amount of statistical noise to the data or the model's output, making it difficult to identify individual patient information from the aggregated data. This method is valuable in longitudinal studies or ongoing treatment efficacy research, where maintaining the anonymity of patients in the face of new data is crucial. It ensures that the addition of new patient data to an existing dataset does not significantly increase the risk of identifying any individual patient, thereby providing a consistent level of privacy protection.

Table 6. Key elements of the ongoing monitoring and updating stage in cancer care, along with the associated privacy and security challenges

Monitoring and Updating Aspect	Description	Privacy and Security Challenges
Continuous Data Collection	Regular collection of patient health data post-treatment, including health status, side effects, and recurrence of cancer.	Ensuring the continuous collection of data is secure and maintains patient confidentiality.
Data Quality Assurance	Ongoing verification and validation of the data being collected to ensure its accuracy and relevance.	Balancing the need for accurate, up-to-date information with the protection of sensitive patient data.
Model Re-evaluation and Tuning	Periodically re-evaluating and adjusting predictive models based on new data to ensure their accuracy and effectiveness.	Keeping predictive models updated without compromising data security, especially when integrating new data.
System Security Updates	Regularly updating IT systems, software, and security protocols to protect against new vulnerabilities and threats.	Staying ahead of evolving cybersecurity threats to protect sensitive health data.
Compliance with Evolving Regulations	Adapting to changes in legal and regulatory standards related to patient data privacy and security.	Keeping data handling and reporting practices aligned with changing regulations and standards.
Patient Data Review and Feedback	Providing patients with regular updates on their health status and any changes in their treatment or care plan.	Ensuring patient data is shared securely and is accessible only to authorized individuals.
Real-Time Health Monitoring	Using wearable devices or digital health platforms to monitor patients' health indicators in real-time.	Protecting data collected from wearable devices from unauthorized access or breaches.
Feedback Integration and System Adaptation	Integrating patient and clinician feedback to improve data collection and analysis processes.	Safely incorporating feedback into systems without exposing them to new privacy or security risks.

Federated learning offers a more conservative approach to model updating in cancer care, especially when considering the privacy of patient data. This method allows for the development and improvement of predictive models using data from multiple sources, without the need to directly share the data itself. In practice, federated learning enables different healthcare institutions to contribute to a collective model without exposing their individual patient data. This is relevant in cancer research,

where pooling data can lead to more robust and accurate models. By using federated learning, each participating entity can train the model locally on their dataset and only share the model updates, rather than the data, thereby reducing the risk of privacy breaches. This approach is beneficial for collaborative efforts in cancer research and treatment development, where sharing insights is important, but patient privacy must be rigorously protected.

Conclusion

The six-stage framework proposed in this study offers a pragmatic approach to safeguarding patient data throughout its entire lifecycle, from collection to ongoing monitoring and updating. Compliance with relevant laws and regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, ensures that patient data is handled in a manner that respects both privacy and confidentiality. This compliance is not just a legal requirement but also a trust-building element between patients and healthcare providers. Each stage of data collection, analysis, sharing, and reporting must adhere to these regulations to protect sensitive patient information. This involves implementing appropriate security measures, ensuring patient consent for data use, and maintaining transparency in data handling practices. For cancer care providers and researchers, navigating these regulations can be complex, especially when dealing with cross-border data sharing or collaborative research, but it is essential for maintaining ethical standards and public trust.

Data integrity is another critical factor in managing patient data in cancer care. Maintaining the accuracy and usefulness of data throughout its lifecycle is vital for effective diagnosis, treatment planning, and research. This involves ensuring that data is not only collected and recorded accurately but also maintained and updated with the same level of diligence. Inaccuracies or inconsistencies in data can lead to incorrect treatment decisions or flawed research outcomes. As such, regular audits, validation processes, and updates are necessary to ensure that the data remains reliable and relevant. Moreover, as new data is integrated, it is important to ensure that it aligns with existing datasets in terms of format and quality. In cancer care, where treatment decisions can be life-altering, the importance of data integrity cannot be overstated.

Balancing efficiency and privacy in the use of patient data is a delicate and ongoing challenge in cancer care. On one hand, efficient use of data can lead to significant advancements in treatment and research, providing insights that can improve patient

outcomes and care. On the other hand, this needs to be balanced with the imperative to protect patient privacy. Each step in the handling of patient data involves a trade-off between making the data useful for healthcare providers and researchers, and keeping it secure and private. This balance requires careful consideration of how data is accessed, shared, and used. It often involves employing advanced technologies and methodologies, such as data anonymization and secure data sharing protocols, to ensure that patient privacy is not compromised in the pursuit of efficiency. As technology evolves and the volume of data increases, this balance will continue to be a key consideration in the ethical and effective management of patient data in cancer care.

When integrating various privacy-preserving techniques in the management of sensitive healthcare data, such as in cancer care, it is worth recognizing that these methods can interact in complex ways. Each technique, whether designed to anonymize, encrypt, or otherwise secure data, operates under its own set of principles and affects the data differently. When these techniques are combined, their interactions can potentially alter the data's usability and privacy protection in unforeseen ways. For instance, a method that adds noise to data for privacy might conflict with another method that compresses data for efficient storage, resulting in either compromised privacy or reduced data quality. This complexity necessitates rigorous testing and validation of each combination of techniques to ensure they work harmoniously without counteracting each other's benefits. Such validation should assess not only the effectiveness of privacy protection but also how the combination impacts the data's utility for research and clinical applications.

References

Al-Issa, Y., Ottom, M. A., & Tamrawi, A. (2019). eHealth Cloud Security Challenges: A Survey. *Journal of Healthcare Engineering*, *2019*, 7516035.

Bai, Z., Yang, R., & Liang, Y. (2019). Mental task classification using electroencephalogram signal. *ArXiv Preprint ArXiv:1910.03023*.

Barrett, C. A. D. (2019). ARE THE EU GDPR AND THE CALIFORNIA CCPA BECOMING THE DE FACTO GLOBAL STANDARDS FOR DATA PRIVACY AND PROTECTION? *Scitech Lawyer*, *15*(3), 24–29.

Bayatbabolghani, F., & Blanton, M. (2018). Secure Multi-Party Computation. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2157–2159.

Boyd, J. H., Randall, S. M., & Ferrante, A. M. (2015). Application of Privacy-Preserving Techniques in Operational Record Linkage Centres. In A. Gkoulalas-Divanis & G. Loukides (Eds.), *Medical Data Privacy Handbook* (pp. 267–287). Springer International Publishing.

Chavez, A., Koutentakis, D., Liang, Y., Tripathy, S., & Yun, J. (2019). Identify statistical similarities and differences between the deadliest cancer types through gene expression. *ArXiv Preprint ArXiv:1903.07847*.

Domingo-Ferrer, J., Solanas, A., & Martinez-Balleste, A. (2006). Privacy in statistical databases: K-anonymity through microaggregation. *2006 IEEE International Conference on Granular Computing*. 2006 IEEE International Conference on Granular Computing, Atlanta, GA, USA. https://doi.org/10.1109/grc.2006.1635915

Du, W., & Atallah, M. J. (2001). Secure multi-party computation problems and their applications: a review and open problems. *Proceedings of the 2001 Workshop on New Security Paradigms*, 13–22.

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., & Bottomley, J. (2009). A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association: JAMIA*, *16*(5), 670–682.

Saxena, A. K. (2020). Balancing Privacy, Personalization, and Human Rights in the Digital Age. *Eigenpub Review of Science and Technology*, *4*(1), 24–37.

Tschider, C. A. (2019). The healthcare privacy-artificial intelligence impasse. *Santa Clara High Tech. LJ*. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/sccj36&section=22

Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In A. Y. Zomaya & S. Sakr (Eds.), *Handbook of Big Data Technologies* (pp. 851–895). Springer International Publishing.

Verykios, V. S., & Christen, P. (2013). Privacy-preserving record linkage. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *3*(5), 321–332.

Vijayarani, S., & Tamilarasi, A. (2010). K-Anonymity Techniques-A Review. *International Journal of Computer Science and Application*. https://www.researchgate.net/profile/Vijayarani-Mohan/publication/329363217_K-Anonymity_Techniques_-_A_Review/links/5e44f3ee299bf1cdb924ea7c/K-Anonymity-Techniques-A-Review.pdf