Student Name: Joseph Chan

Email: Joseph.chan@berkeley.edu

Capstone Project Title: Optimizing Clinical Trial Recruitment

Capstone Team Names and Emails

Eric Shen: shen.y.eric@ischool.berkeley.edu

Simone Ong: simoneong@ischool.berkeley.edu

Matthew Zhang: matthewzhang@ischool.berkeley.edu

Darren lo: darrenlo@ischool.berkeley.edu

Jonathan Luo: jonnyluo@ischool.berkeley.edu

Capstone instructor emails:

kwetzel@ischool.berkeley.edu

zonakostic@ischool.berkeley.edu

Link to slides:

https://docs.google.com/presentation/d/1RM8Y9mF_rIyFpMvqqrvjoM6Y9DOAdAfEMd497Bty8eo/edit?usp=sharing

Transcript: (Shared google link, apologies zoom cloud did not record)

https://drive.google.com/file/d/1BKXtdfjWJ3pluG7kT3Yr9niKRftCo6Am/view

Capstone Audit

1.0 - Project Summary

The 'Optimizing Clinical Trial Recruitment' capstone project aims to create a machine learning powered service, with an associated web interface, that allows end users to find and accurately match eligible patients to clinical trials. The motivation for the project is centred around two key challenges. Firstly, the pharmaceutical industry's need to efficiently complete trials for product commercialization. Secondly and more crucially, the significant disparity between patients' interest in participating in clinical trials and actual enrolment onto clinical trials.

The project is divided into three key components that work together to facilitate the end user experience. The first is to leverage large language models to extract key patient information from patients notes. At the present time, the project harnesses the OpenAI API but there are plans to build and utilise their own LLM. The second is to create a model that accurately matches this extracted patient information to clinical trials listed on clinicaltrials.gov. The final part binds these two parts through a web interface that is potentially accessed via login. This web interface provides end users the ability to upload data and return a list of clinical trials that most accurately match a patient's eligibility criteria.

During training and evaluation, the team is planning to use the publicly available MIMIC-IV and TREC synthetic medical datasets (Johnson et al., 2023; Roberts et al., 2022). Relevant patient information will be summarised and extracted. This will be matched to clinical trials on the publicly available clinicaltrials.gov website. This website lists many of the clinical trials running around the world. Initially the team aims to provide a match based on the accuracy of a patient's extracted details against the eligibility criteria listed on potential clinical trials. However, the team plans to enhance this matching model and the associated quantitative metrics as the project develops.

The project is primarily aimed at empowering patients to look for clinical trials and lower the barrier to access trials. However, the team also envisages that clinicians will also be end users of the system.

At the time of the audit, many of the components of the project were in their preliminary stages and we discussed aspects of the project both currently and as the team anticipates how they will exist in deployment. The team has also not developed a clear privacy policy nor a terms of service policy at the time of the audit interview.

This audit will provide a privacy impact assessment (PIA), discuss ethical and fairness issues and create a model card based on the discussion with the team.

2.0 - Privacy Impact

The aspects of privacy for the current project are expansive as the project aims to use different data sources in a variety of ways that change throughout the project from development to production. Here we dissect the project through the lens of several privacy frameworks then offer clear concrete steps to protect privacy.

To begin, Mulligan's analytic provides a powerful framework to understand key components or privacy (Mulligan et al., 2016). Mulligan's dimensions of theory and protection help us frame and contextualize the basis of privacy as well as the target of privacy protection. Viewing it through this lens, the object of privacy is control of personal information and in this case, the emphasis must be on sensitive health information, which concretely is stored and used as medical data. The justification of privacy, as viewed through Mulligan's dimension of theory, is to preserve the dignity and individual liberty of the subject (Mulligan et al., 2016). For this project, the subject is primarily patients who would be using the clinical trials application. The exemplar, or archetypal examples of privacy violation in this context are innumerable in healthcare. The cost to patients and organisations when privacy is violated is substantive. Breaches in privacy lead to significant harms, including patient distress, erosion in trust in healthcare and significant economic costs (Pool et al., 2024).

2.1 - Data collection and use during development

To further understand what harms could arise, as seen by Solove's Taxonomy of harms, we must first understand what data is being collected, how that data is being used and how that data might be stored and shared. For this project, the initial data being used during training and evaluation of models and in development comes from publicly available datasets. For patient data, this comes from the MIMIC-IV dataset and from the TREC synthetic dataset (Johnson et al., 2023; Roberts et al., 2022). The MIMIC-IV set is one of the most extensively used datasets in healthcare research and has already been deidentified according to HIPAA safe harbor provisions (Johnson et al., 2023). In addition to safe harbor provisions, users of this dataset must agree to not make attempts at re-identifying patients. The TREC dataset on the other hand is a fully synthetic dataset which does not contain any real patient information (Roberts et al., 2022). Finally, the clinical trial data comes from clinicaltrials.gov (NIH, n.d.). This is a publicly accessible register of clinical trials which have been uploaded by investigators. This register poses negligible risk to end users of the project as they do not interface with it. These datasets are the backbone of the project, providing the basis for training and evaluating models that will then be used in production.

As each of these datasets are publicly available, the privacy expectations of stakeholders in this context change and the free open use of these data is far more acceptable. Nissenbaum's contextual integrity defines privacy through the idea of information norms and the contexts by which information flows (Nissenbaum, 2011). Although these datasets are open access and de-identified/synthetic, there still exist norms in these contexts. For example, agreeing not to identify patients as per the terms of agreement when using MIMIC-IV data should be respected. Indeed, failure to do so would lead to aggregation and identification harms identified by Solove's taxonomy of harms(Solove, 2005). Concretely, steps to mitigate these harms could comprise of steps to exclude prompts sent to a third-party large language model (LLM) that could identify individuals and to minimize aggregation where possible.

2.2 - Privacy concerns during deployment

As the project moves from testing and training to production and deployment, the collection, processing and use of data changes substantially. Here end users, which could include both patients and clinicians, upload patient documents into the web portal where they are processed and used for matching. The team also expands further, describing an idealised product where users may have to create an account to use the service. The audit will focus on these aspects now, as there is a plethora of risks that can arise when using actual patient data. It should be noted that the team has not made a final decision as to whether the entirety of the aforementioned datasets will be used or whether a subset will be used. How the data will be split into training/validation and testing sets has also not yet been finalised. Thus, the exact technical specifications in terms of features, hyperparameters and model tuning are not known for this specific audit. As the final configuration and workflow of the end product is unknown, this audit will address both the idealised versions of the end product as well as alternatives that were raised by the team.

2.3 - Data collection during deployment

It is anticipated that in production, that real patient data will be uploaded onto the website where it will be processed by an LLM. De-identification of patient data had not been adequately considered when interviewing the team. This is a necessary step as data storage prior to processing, even if temporary, can lead to risks of data insecurity and accessibility as described by Solove (Solove, 2005). These harms, when realized, can lead to further widespread harms of information dissemination including breach of confidentiality, disclosure and exposure. Additionally, legal jurisdictional concerns may arise depending on who the end user is and, whether the end user is based. For example, HIPAA may become more applicable if the end user is a clinician and privacy measures encompassed within GDPR becomes essential if the service anticipates users from the European Union (McGraw & Mandl, 2021). Concrete steps to address the privacy concerns would be to require full anonymization of patient data, considering both HIPAA safe harbor requirements as well as potentially GDPR requirements. In the most ideal scenario, anonymization would be achieved prior to document upload. The simplest way to achieve anonymization would be to ask end users to remove all identifiers on the uploaded documents themselves, however this might not be realistic in real world deployment. Alternatively, a client-side LLM or NLP-powered anonymization tool could be used to process the data before upload. However, it is important to acknowledge that such a tool is unlikely to be perfectly accurate without human review. The team should communicate to end users that de-identification is not guaranteed to be 100% effective.

Moreover, de-identification should always be coupled with rigorous cybersecurity measures, such as strong encryption of records and compartmentalization of data to mitigate risks.

2.4 - Data processing during deployment

In terms of data processing, the team aims to process the patient data by summarising and extracting appropriate fields to then match to clinical trials. The team's current workflow for summarisation uses OpenAI's API to receive the data and prompt, before summarizing patient data. The team indicated that OpenAI does not store information. However, a cursory review of OpenAI's platform documents suggests that while there are options for patients using sensitive data, companies need to specifically request for zero data retention (OpenAI, n.d.). This expands the possibilities of violating privacy if care is not taken to consider third party data sharing. Specifically, by potentially breaching confidentiality, disclosure and exposure of information, and potential secondary use – all harms identified by Solove's Taxonomy (Solove, 2005). Moreover, the norms of appropriateness, as viewed through Nissenbaum's contextual integrity, need to be considered carefully here (Nissenbaum, 2011). Most patients would not expect their health data to be shared except where explicit consent has been provided, and they would expect that this data be secure from breaches. Moreover, patients would likely not expect OpenAI to retain or reuse health information. Concrete steps if the team chooses to use OpenAI in its final model is to review data retention policies of OpenAI, opt for zero data retention explicitly and to list OpenAI within the teams' trials platforms terms of services and privacy policy. These steps also abide to FIPPs 'transparency' principles and alleviates potential risks of exclusion or failing to let people know about handling of their data as identified by Solove's Taxonomy (DHEW, 1973; Solove, 2005).

In the alternate, idealised scenario, where the team has indicated they would build their own LLM. Privacy risks are not necessarily mitigated. First, regardless of where the LLM is hosted – either locally or on the cloud, the team now needs to consider potential risks of insecurity. There may also be transnational regulations such as the General Data Protection Regulation (GDPR) that need to be navigated(McGraw & Mandl, 2021). The team has identified the need to implement modern cybersecurity measures, and these would indeed mitigate some of these risks. Further concrete steps would include careful scoping of both security risks, considering transnational laws and by extension any regulations that would extend from there. Similar risks of harm from Solove's Taxonomy seen in the OpenAI scenario are still applicable here. The team needs to be transparent with end users about where their data is stored, how long it is stored and how it will be processed even if it is done 'locally'. Implementing these would mitigate risks of data breaches. This in turn would reduce harm to end users, protect their dignity and avoiding legal and financial ramifications of a data breach.

The next part of the project involves matching patients summarised data against the eligibility criteria set by clinical trials extracted from clinicaltrials.gov. In the base scenario, the simplest model will be an accuracy score based on the number of matching criteria. In the idealised production level model, the team suggests that matching could involve using a model that relies on embeddings of patient data and clinical trials. Regardless of method of match, there are potential privacy risks associated with both scenarios. These relate primarily to storage of summarised data which we identified earlier. Here we see further justification for early and immediate de-identification of patient data. During discussions, it was also initially planned that each session would be cleared immediately after use. However, on discussing the potential benefits of longitudinal storage to enable patients to match new trials – the team considered this potentially as a highly beneficial new feature. In this new idealised situation, the risks of insecurity to privacy are amplified as are risks of information dissemination and potential risks of disclosure and exposure. Concrete steps to mitigate these risks are broadly covered by the team's plan to implement modern cybersecurity standards.

2.5 - Secondary use of data and data linkage

In this specific project, the team initially considered the project to have a cache that clears every session as the primary method to avoid data harms associated with data storage and to help mitigate secondary data use concerns. However, on questioning there were several avenues of future development that could bypass this strategy and lead to harms arising from both secondary data use and aggregation.

First, there should be consideration of whether data and results used during deployment should be used to improve the model. The team did not address this explicitly during our interview, but this needs to be carefully considered in the context of medical data. If patient data is used for purposes beyond the initial intent of just matching to trials, then patients should be informed transparently. From Nissenbaum's lens, the norms of information flow would suggest that in this context, many patients would likely accept this secondary data use if adequately informed. Indeed, there is good evidence in the literature to support this idea (Hutchings et al., 2021). Secondary data harms arise when consent and notice is not given because this violates an end users' dignity, creates architectural harms from feeling powerless and violates trust and expectations on data use. Solove highlights these harms and argues that such harms arise from the idea of power imbalance (Solove, 2005).

An additional issue that we identified relates to the creation of user accounts and potential data linkage harms from aggregation. During our interview, the team highlighted that people may be allowed to create user accounts. They could not provide a definite answer as to whether this would be a barrier to access. However, one harm that they had not considered was that patient data could be potentially re-identified if the account details were linked to the associated uploaded patient data. For example, if a name was required for an account, that name could reidentify patients' data which had been de-identified. This links back to the idea of aggregation harms where the whole becomes greater than the sum of its parts and creates dignitary harms as discussed by Solove (Solove, 2005). Similarly, it violates the norms of appropriateness because end users have a reasonable expectation that data will not be aggregated and thus allow privacy to be violated (Nissenbaum, 2011). Tangible steps to avoid these harms would involve anonymizing user accounts to the bare minimum details required for the platform, ensuring best cybersecurity practices, compartmentalizing of data types and minimizing data linkage where possible. These steps would provide a multilayered approach that would reduce data breaches and minimize harm to end users even if there is a data breach. Implementing these protective measures would also reinforce users trust in the platform.

2.6 – Privacy protection for special groups and special data

Privacy protection is particularly important for special groups and special types of data because of the harms that can arise when those groups are not protected. One overarching concern is how protected health information is handled on a web platform such as the current project. As the project's platform collects, stores and potentially transmits sensitive health information - this platform is potentially subject to HIPAA regulations (Tovino, 2025). We would strongly suggest seeking legal advice if this platform does indeed go live as the implications of this are beyond the scope of this audit.

The one special group that we must highlight is children as a group. Although the current projects intended scope is for adults, there are ways in which out of scope use could occur with children. For example, depending on the final setup of the platform, children could inadvertently access the platform if there is no login account barrier that requires verification. It is not inconceivable that children, particularly older children who are unwell with a life limiting illness, might search for clinical trials. Under the Children's Online Privacy Protection Act (COPPA) there are extra protections that apply to children under 13. COPPA requires clear privacy policies, parental consent and limits data collection (*Research Handbook on Privacy and Data Protection Law*, 2022). If the platform is accessible in Europe or by a European, this also puts the platform under the GDPR requirements which have implications for children under the age of 16 (*Research Handbook on Privacy and Data Protection Law*, 2022). One concrete step to prevent use by children would be to mandate account setup with verification. However, such a step might discourage users and undermine the team's mission to empower patients. A less rigorous but viable alternative would be to place a pop-up screen that verifies users age in a similar way to pop-up barriers placed on gambling websites. In addition, the team should consider automatic detection of age in the uploaded patient documents such that the pipeline will stop processing data and immediately purge the data from the cache. Finally, it would be prudent to again seek legal advice to ensure compliance with appropriate legislation across different jurisdictions. These concrete steps must be carefully considered in the context of the platform. Implementing these steps would help protect the dignity of children and protect the project from significant legal ramifications as well.

3.0 – Additional ethical concerns

There are some additional ethical concerns that arise from this project. First, there are aspects of data collection and processing that could be perceived as medical research. If secondary use of data generates insights that are then used to improve the system, these aspects of the platform would likely benefit from review by an institutional review board. The Belmont principles offer two fundamental ideas of justice and beneficence which are relevant if the project is viewed through the lens of biomedical research (DHEW, 1978). These ideas advocate for the burdens of research to be distributed fairly and for the idea that the research should be of benefit to patients. If new insights are found during improvement of the platform's models that could be of benefit to users, withholding these insights could be considered unethical. This project offers challenging ethical questions that the team should consider carefully.

Second, although no medical opinion is provided, the platform does provide matches to clinical trials. These matches, regardless of intent, will influence some patient's choices regarding their healthcare that might not otherwise have occurred. Whilst the platform could empower patients to make decisions, this 'empowerment' might also lead to unintended consequences without the adequate guidance of physicians. Clinicaltrials.gov is a platform that does not consider the scientific validity of listed trials (NIH, n.d.). Thus, patients may be potentially making decisions that are not be in their best interests given the context of their clinical situation. The team should also consider that the idea of a clinical trial might bring false hope and lead to financial toxicity that might not have occurred if patients did not have access to the platform. Whether these choices should lead to increased gatekeeping, in terms of how easily the platform can be accessed, needs to be carefully considered prior to full deployment.

3.1 - Fairness

Finally, there should be some consideration of fairness. At the present time, the team has not fully considered whether their models are fair for all users who might access their platform. The projects primary source of patient data during training is the MIMIC-IV dataset which is heavily skewed towards a Caucasian, elderly population from Boston (Mohammed et al., 2023). Training their models on this population might bias the model and result in certain demographic groups such as minorities and younger adults failing to match as well. The team has not considered unitary or intersectional quantitative analyses of demographic subgroups yet. This audit recommends that the performance of the model be considered in terms of age, gender and racial subgroups to assess fairness. Limitations of the model or unfairness should be transparent to users on the platform. Future models should consider additional intersectional analyses such as subgroups of disease that may be underrepresented.

4.0 - Actionable steps for ensuring privacy

The project is in its early stages of development, but the team has already carefully considered some steps to try to mitigate privacy risks. For example, they have chosen to use publicly available synthetic and de-identified data sets for training and evaluation. They also talk about using modern cybersecurity measures and try to avoid data storage as much as possible. For these steps, they should be commended. However, moving forward into development there are significant risks that can arise. Here we will provide some additional concrete steps to mitigate privacy risks and impacts.

One excellent document that provides key principles for helping mitigate privacy concerns is the Fair Information Practice Principles (FIPPs)(DHEW, 1973). Here the team has already taken steps to implement principles such as limiting data collection and data security. This audit makes several additional recommendations directly addressing FIPPs principles.

The first principle to address is that of notice and awareness. This is the idea that organizations should inform individuals about their information practices, disclose what data is collected and how it will be used (DHEW, 1973). This audit strongly recommends that a clear notice of data use and collection be displayed as a popup screen that must be acknowledged when end users access the website. This notice should entail a brief but accurate summary of the terms of service and privacy policy. It should also provide links to easily access the full policy documents. By extension, the team should create clear terms of service and privacy policy documents. These need to be easily accessible, avoid jargon and provide sufficient detail for users. These documents should explain how data is collected, processed, used, stored and shared with any third parties if applicable. The documents should also be easily accessible from the web platform.

Our second recommendation addresses the principle calling for choice and consent and the ability to opt-in and opt-out of options where possible and appropriate (DHEW, 1973). Given that the projects platform uses highly sensitive health data, this audit strongly recommends that consent be obtained when an end user uses the platform. The consent should appear prior to any collection or use of the platform. This consent should be easily understood, informative and provide options that relate to secondary data use. There should also be a revocation of consent if applicable and recourse if this is not enforced.

The third recommendation addresses the principle that individuals should be able to access their data and review what data is held in regard to that individual (DHEW, 1973). If future deployed versions retain data beyond the immediate session, we strongly urge that users be able to access their data and contest any accuracy issues. Failing to do so could result in harm with incorrect matches or bias.

By implementing all these steps, end users will gain trust in the platform and be empowered to know that they are in control of their data and information flows. Enforcement of these principles as well as accountability is essential for privacy to be maintained. There should be mechanisms in force that allow enforcement of privacy principles and for redress if these principles are violated. Information regarding how patients can seek redress should be clearly available on the platform. These steps will reassure users that their concerns are taken seriously if violations do take place.

5.0 – Conclusion

This capstone project represents an innovative solution of leveraging machine learning models to empower patients to match to clinical trials. However, such a system needs to be well considered in many contexts including privacy, ethical, legal and fairness principles. The project team has already made significant progress in addressing some of these concerns and this audit expands on some key elements that if implemented will protect both end users and the project team from potential harms.

Model Card – Optimizing Clinical Trials Recruitment

Model Details

- Developed by Capstone Team from MIDS
- Model date: Ongoing, Developed Q1 2025
- Model Version: In development
- First part of model is large language model, unknown architecture
- Second part of model is an embedding space matching model

Intended use

- For adult patients who want to match to a clinical trial, may include all medical conditions or a subset (to be decided).
- Secondary use is for clinicians who want to match their patients to trials
- Intended to provide a list of highest scoring trials according to set criteria.
- Not intended for children (< 18 years)
- Not intended to provide a medical opinion

Factors

- Relevant factors: Based on published demographic data about MIMIC-IV dataset (Johnson et al., 2023), relevant factors include gender, age, ethnicity, type of disease. Environment factors include insurance status and location of patient.
- Evaluation factors: Not yet reported as project in development but should include all of the relevant factors.

Metrics

- Evaluation metrics initially include simple accuracy score provided by % of clinical trial criteria matching extracted patient data.
- Top three matches will be provided to end users per search.
- Further evaluation metrics not fully considered yet as project continues to be in development.

Quantitative Analyses

- No quantitative analyses available.

Training data

- Primarily MIMIC-IV dataset training data split. Details can be seen here and here. Dataset is a de-identified set of patient data from a single urban hospital in Boston, MA, USA (Mohammed et al., 2023).
- Clinicaltrials.gov clinical trials dataset which is a public repository of registered trials around the world.
- TREC synthetic medical dataset is a synthetic dataset used for testing in medical contexts (Roberts et al., 2022).

Evaluation Data

- Primarily MIMIC-IV dataset test data split
- Clinicaltrials.gov clinical trials dataset
- TREC synthetic data is an additional dataset that may be used

Ethical considerations

- The data uses sensitive health data, but this is either fully de-identified in the MIMIC-IV set or synthetic in the case of the TREC dataset. In deployment, considerations need to be made regarding de-identification of patients.
- The generated match from the model although not intended to provide a medical opinion does potentially lead to decisions about health.
- Risk mitigations include using deidentified data during training and evaluation. The team has also considered a LLM API provider who will minimise data storage.
- Risks during use in deployment include data insecurity risks and data dissemination risks. The team is planning to deploy modern cybersecurity measures to mitigate these risks in line with Fair Information Practice Principle (DHEW, 1973).

Caveats and Recommendations

- Training and evaluation data should consider addition of other datasets in training and evaluation of the model. MIMIC-IV underrepresents many groups including non-white races, patients younger than 65 and patients outside the Boston area (Mohammed et al., 2023). Furthermore, there should be consideration of using de-identified data with consent from patients, when the initial model is deployed. These ideas touch on idea of 'justice' from the Belmont Report which considers fair access to different populations and ensure distribution of benefits of research (DHEW, 1978).
- Clear notice of data collection and consent needs to be implemented in the final web interface. Secondary use consent for fine-tuning and updating models should also be considered. These are important FIPPs principles to implement.
- In continuation of the previous recommendation, a clear privacy policy and terms of service should be available on the web interface. This should be transparent about data collection and use – but also clearly talk about third party data sharing (e.g. OpenAI API) if applicable. This should also provide information on access and redress from end users. Finally, there should be some consideration of appropriate regulations such as GDPR and HIPAA in these policies.

References

DHEW. (1973). *Records, Computers and the Rights of Citizens*.

DHEW. (1978). *The Belmont Report*. United States Government Printing Office.

Hutchings, E., Loomes, M., Butow, P., & Boyle, F. M. (2021). A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: A focus on consent. *Systematic Reviews*, *10*(1), 132. https://doi.org/10.1186/s13643-021-01663-z

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x

McGraw, D., & Mandl, K. D. (2021). Privacy protections to encourage use of health-relevant digital data in a learning health system. *NPJ Digital Medicine*, *4*, 2. https://doi.org/10.1038/s41746-020-00362-8

Mohammed, S., Matos, J., Doutreligne, M., Celi, L. A., & Struja, T. (2023). Racial Disparities in Invasive ICU Treatments Among Septic Patients: High Resolution Electronic Health Records Analysis from MIMIC-IV. *The Yale Journal of Biology and Medicine*, *96*(3), 293–312. https://doi.org/10.59249/WDJI8829

Mulligan, D. K., Koopman, C., & Doty, N. (2016). Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160118. https://doi.org/10.1098/rsta.2016.0118

NIH. (n.d.). *ClinicalTrials.gov* [Dataset]. https://clinicaltrials.gov/

Nissenbaum, H. (2011). *A Contextual Approach to Privacy Online* (SSRN Scholarly Paper No. 2567042). Social Science Research Network. https://papers.ssrn.com/abstract=2567042

OpenAI. (n.d.). *Models*. https://platform.openai.com/docs/models/how-we-use-your-data

Pool, J., Akhlaghpour, S., Fatehi, F., & Burton-Jones, A. (2024). A systematic analysis of failures in protecting personal health data: A scoping review. *International Journal of Information Management*, *74*, 102719. https://doi.org/10.1016/j.ijinfomgt.2023.102719

Research Handbook on Privacy and Data Protection Law. (2022). https://www.elgaronline.com/edcollbook/edcoll/9781786438508/9781786438508.xml

Roberts, K., Demner-Fushman, D., Voorhees, E. M., Bedrick, S., & Hersh, W. R. (2022). *Overview of the TREC 2022 Clinical Trials Track.* TREC.

Solove, D. J. (2005). *A Taxonomy of Privacy* (SSRN Scholarly Paper No. 667622). Social Science Research Network. https://papers.ssrn.com/abstract=667622

Tovino, S. (2025). Artificial Intelligence and the HIPAA Privacy Rule: A Primer. *Houston Journal of Health Law & Policy*, *77.*