Jose Arboleda/Muhammad Mamsa

11/4/24

CSC 10800

Questions

Q1) Is this data still considered private even though it does not include the names of individuals? Why or why not?

In our opinion, the data within the dataset is still private. Although PHI is being disclosed, PII isn't. There's no way to assign any of this PHI data to anyone as no names exist. The yes or no answers, BMI, cholesterol, weight, circumference numbers, etc. all seem arbitrary without having names for them to be assigned to. Since no one's health in specific is being put at risk for general knowledge, we would still consider this data to be private.

Q2) Which columns in the dataset might reveal identifiable information about individuals, potentially violating data privacy? List and briefly explain why.

Three columns might violate data privacy. The age, Height, and Weight columns could be problematic in terms of if someone has an outlier height or weight they can easily be identified.

Q3) This data is being used to help predict if an individual has type-2 diabetes or not. Is this an appropriate use of the data? Why or why not?

This is not an appropriate use of the data because nothing correlates to diabetes. It asks about blood pressure and cholesterol. It tracks the blood pressure and cholesterol of certain individuals.

Q4) Consider the following scenarios:

- Scenario A: A researcher wants to use the dataset to study the correlation between weight and hypertension in middle-aged adults without obtaining consent from the individuals.

Is this an appropriate use of the data? Why or why not?

> In our opinion, as long as this data isn't being used for any other reason other than research, it's fine. If PII and PHI isn't being disclosed to the general public and only being used to conduct research out of curiosity with no financial gain attached to it, it doesn't seem like much of a problem. Research is always being conducted and it's important to have knowledge of things available at all times, however privacy of

individuals involved has to be kept, especially if consent for the use of their data isn't being asked.

- Scenario B: A pharmaceutical company requests access to the dataset, including personal identifiers, to market diabetes medication directly to individuals.

Is this an appropriate use of the data? Why or why not?

In this scenario, the usage of the data would be incorrect. The company may have requested and granted the usage of this information, but the consent of those involved hasn't been granted. This is completely different from the first scenario as where the interest in the first scenario is purely research, this scenario presents financial gain through marketing. PII and PHI are both being put at risk all for financial gain which is very shady in the medical field.

Q5) Suppose a new column, "Full Name," is added to the dataset.

1. In what contexts might it be appropriate to share the full dataset, including Full Name, and with whom?
    - If the dataset is intended for internal use by an organization (e.g., for employee management, research, or team collaboration), sharing the full dataset among authorized personnel may be appropriate, provided that there are policies in place to protect personal information.
    - If individuals have given informed consent for their data to be used in a research study, sharing the dataset may be appropriate, especially if the research aims to improve services or products that benefit the participants.

2. In what contexts would it be inappropriate to share this information? Discuss the potential risks to individual privacy.
    - Sharing the full dataset in a public forum or repository without any restrictions can expose individuals to privacy risks, especially if the data can be linked back to individuals.
    - Using full names for marketing or commercial initiatives without explicit consent raises significant ethical concerns and may violate privacy regulations.

Q6) Using the ACM Digital Library, find a technical article related to privacy concerns in diabetes data or medical data analysis

2. Summarize the key privacy insights discussed in the article.
    - Subjective Awareness: People who utilize automated insulin delivery (AID) devices frequently have differing degrees of knowledge about the privacy concerns and the use of their data. Their willingness to disclose data and level of system confidence are impacted by this awareness.
    - Data Sensitivity: Very sensitive information can be found in medical datasets, particularly those pertaining to long-term illnesses like diabetes. The necessity for

strong privacy protections is highlighted by the possibility of misuse or unintentional exposure of sensitive data.

- The significance of informed permission in data sharing is emphasized in the article, which makes sure consumers are aware of the data being gathered and its intended purpose.

3. How do the findings relate to the dataset we are examining?
    - The results highlight the need of taking user privacy into account while handling personal data, which is relevant to the dataset we are looking at. The observations regarding subjective awareness and the necessity of informed consent are especially pertinent if the dataset contains sensitive data, such as complete identities or medical histories. If users don't believe their data will be managed appropriately, they could be less inclined to provide it.

4. What recommendations do the authors make regarding privacy preservation in medical datasets?
    - The authors suggest a number of tactics to protect patient privacy in medical datasets:
    - Enhanced Transparency: Gaining users' trust requires being transparent about data collection, use, and sharing policies.
    - Strong Consent Procedures: Ensuring that users have control over their personal data is achieved by putting in place procedures that make it simple for them to grant and withdraw consent.
    - Data minimization lowers the risk of data breaches and misuse by only gathering the information required for the intended usage.
    - User education may empower users and enhance their comprehension of the ramifications of data sharing by raising user awareness of privacy policies and how AI systems operate.