A sensible censor for sharing medical records

July 24 2008

(PhysOrg.com) -- Newly developed MIT software will help to allay patients' fears about who has access to their confidential records, facilitating the use of that data for medical research.

In the July 24 issue of the journal *BMC Medical Informatics and Decision Making*, a team of MIT researchers describes a computer program capable of automatically deleting details from medical records that may identify patients, while leaving important medical information intact.

Patient records that are to be shared within the research community must have any identifying information removed, according to the U.S. Health Insurance Portability and Accountability Act (HIPAA). However, manual removal of identifying information is prohibitively expensive, time consuming and prone to error-constraints that have prompted considerable research toward developing automated techniques for "de-identifying" medical records.

The MIT team aimed to solve this problem. "We've developed a free and open-source software package to allow researchers to accurately de-identify text in medical records in a HIPAA-compliant manner," said Gari D. Clifford, a principal research scientist in the Harvard-MIT Division of Health Sciences and Technology (HST) who led the work with Principal Investigator Roger G. Mark, a professor in HST and MIT's Department of Electrical Engineering and Computer Science.

According to Dr. Zohara Cohen, program director at the National Institute of Biomedical Imaging and Bioengineering, sponsor of the

work, the information in patients' medical records is a "largely untapped treasure trove" that the biomedical research community could use to better understand diseases and their treatments.

"The automated de-identification software developed under the guidance of Dr. Mark is a big step forward in permitting the widespread sharing of patient information without the risk of compromised privacy and confidentiality," Cohen said.

Clifford, Mark and colleagues tested their censoring software on 1,836 nursing notes (a total of 296,400 words). Using multiple experts and additional algorithms, they replaced all personal information with "fake" data. In their BMC paper, they report that "the software successfully deleted more than 94 percent of the confidential information, while wrongly deleting only 0.2 percent of the useful content. This is significantly better than one expert working alone, at least as good as two trained medical professionals checking each other's work and many, many times faster than either."

The team is providing researchers access to the evaluation dataset together with the software to allow others to improve their systems, and to allow the software to be adapted to other data types that may exhibit different qualities.

Provided by MIT

provided for information purposes only.