

A Natural Language Processing Model to Identify Confidential Content in Adolescent Clinical Notes

Naveed Rabbani¹ Michael Bedgood² Conner Brown³ Ethan Steinberg^{4,5} Rachel L. Goldstein⁶
Jennifer L. Carlson⁶ Natalie Pageler¹ Keith E. Morse¹

¹Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States

²California Department of Public Health, Richmond, California, United States

³Information Services Department, Lucile Packard Children's Hospital, Palo Alto, California, United States

⁴Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, United States

Address for correspondence Naveed Rabbani, MD, Department of Pediatrics, Stanford University School of Medicine, 453 Quarry Road, Stanford, CA 94304, United States (e-mail: nrabbani@stanford.edu).

⁵Department of Computer Science, Stanford University, Stanford, California, United States

⁶Division of Adolescent Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States

Appl Clin Inform 2023;14:400–407.

Abstract

Background The 21st Century Cures Act mandates the immediate, electronic release of health information to patients. However, in the case of adolescents, special consideration is required to ensure that confidentiality is maintained. The detection of confidential content in clinical notes may support operational efforts to preserve adolescent confidentiality while implementing information sharing.

Objectives This study aimed to determine if a natural language processing (NLP) algorithm can identify confidential content in adolescent clinical progress notes.

Methods A total of 1,200 outpatient adolescent progress notes written between 2016 and 2019 were manually annotated to identify confidential content. Labeled sentences from this corpus were featurized and used to train a two-part logistic regression model, which provides both sentence-level and note-level probability estimates that a given text contains confidential content. This model was prospectively validated on a set of 240 progress notes written in May 2022. It was subsequently deployed in a pilot intervention to augment an ongoing operational effort to identify confidential content in progress notes. Note-level probability estimates were used to triage notes for review and sentence-level probability estimates were used to highlight high-risk portions of those notes to aid the manual reviewer.

Results The prevalence of notes containing confidential content was 21% (255/1,200) and 22% (53/240) in the train/test and validation cohorts, respectively. The ensemble logistic regression model achieved an area under the receiver operating characteristic of 90 and 88% in the test and validation cohorts, respectively. Its use in a pilot intervention identified outlier documentation practices and demonstrated efficiency gains over completely manual note review.

Conclusion An NLP algorithm can identify confidential content in progress notes with high accuracy. Its human-in-the-loop deployment in clinical operations augmented an ongoing operational effort to identify confidential content in adolescent progress notes. These findings suggest NLP may be used to support efforts to preserve adolescent confidentiality in the wake of the information blocking mandate.

Keywords

- confidentiality
- patient portals
- natural language processing
- machine learning
- health information exchange

received

October 12, 2022

accepted after revision

March 1, 2023

accepted manuscript online

March 10, 2023

© 2023. Thieme. All rights reserved.

Georg Thieme Verlag KG,

Rüdigerstraße 14,

70469 Stuttgart, Germany

DOI <https://doi.org/10.1055/a-2051-9764>.

ISSN 1869-0327.

Background and Significance

The Information Blocking Final Rule of the 21st Century Cures Act mandates the timely, electronic release of a wide variety of health care data to patients.¹ This legislation represents a vital step in promoting health care information technology interoperability.² Furthermore, there is growing evidence that sharing health data with patients has several benefits including increased engagement, increased care plan adherence, and an improved patient experience.^{3–7}

While there are significant benefits to information sharing with patients, careful consideration is required to protect privacy in the case of adolescents.^{8–10} Providing confidential care for adolescents around sensitive topics such as substance use, sexual, and mental health is an important part of providing high-quality health care for this population and in many cases is mandated by state law.^{11–15} Maintaining this confidentiality is critical to promoting an environment in which adolescent patients will communicate openly with providers and access essential care.^{16–18}

As a result, information sharing as mandated by the 21st Century Cures Act must be implemented in a way that does not unintentionally disclose confidential information to adolescent patients' proxies without consent. This requires confidential information to be documented in the electronic health record (EHR) in a way that it can be segmented and withheld from release to parents or guardians.¹⁹ In the case of clinical progress notes, confidential information may be documented in a separate note type.²⁰ Such an approach would yield two types of notes: a regular progress note that is shared with patient and proxy and an adolescent confidential note that is either not shared or shared only with the patient.²¹ Like any change to workflow and documentation practices, provider education is an important part of the process. Additionally, health systems currently lack a scalable method for identifying and correcting inappropriate inclusion of confidential information in routine progress notes.

To address this issue, we sought to determine whether a natural language processing (NLP) algorithm can be developed to identify confidential content in adolescent progress notes in a way that is clinically relevant and useful for health care operations. In this manuscript, we demonstrate the development of that algorithm, its implementation into clinical operations, and results from a pilot intervention.

Methods

Dataset and Model Development

The following study was performed at a predominantly subspecialty outpatient pediatric network affiliated with a tertiary care academic children's hospital. To inform documentation changes in anticipation of the 21st Century Cures Act, a sample of outpatient progress notes from visits with adolescent patients were reviewed for confidential information. To perform this audit, 1,200 outpatient progress notes written between January 1, 2016 and December 31, 2019 for visits with patients aged 12 to 17 years were randomly

sampled and then equally divided among a team of five physician reviewers (N.R., M.B., R.L.G., J.L.C., and K.E.M.). Physician reviewers were trained in the California adolescent confidentiality laws and annotated portions of the assigned clinical progress notes that contained confidential information. Both positive and negative references were determined to be confidential. The proportion of clinical progress notes containing confidential information was calculated. Further discussion of the methodological details around this annotation process including labeling rubric, interrater reliability, and summary of confidential content identified are summarized in a related manuscript.²²

Since portions of the note were manually annotated, this process yielded two types of ground-truth labels for training and evaluation: a note-level label of whether the progress note as a whole contained any confidential information and sentence-level labels for the sentences that comprise the notes. Using an 80–20 training–test split, we evaluated two types of models: a note-based model and a sentence-based model.

The note-based model consists of a single logistic regression model that takes the note text and returns the probability that the note contains confidential information. For the note-based model, we featurized each note using unigram counts followed by a term frequency-inverse document frequency (TF-IDF) transformation.²³ We then fed those features into an L2 regularized logistic regression model to predict whether the note contains any confidential information.²⁴ Regularization strength was selected via cross validation on the training set.

The sentence-based model consists of two logistic regression models: one that takes a sentence and returns the probability that the sentence contains confidential information and another that takes all of the sentence-level probabilities from a note and generates a note-level probability estimate. To calculate sentence-level probabilities, we divided each note into “sentence” chunks of 10 tokens each. Each chunk was then featurized using unigram counts followed by a TF-IDF transformation. As before, those features were then used in an L2 regularized logistic regression model to predict whether that 10-token chunk contains any confidential information. As standard, we selected regularization strength using cross validation. This provided a logistic regression model that can predict whether a 10-token chunk contains confidential information. We then trained a second logistic regression model that transforms these sentence-level probabilities to an overall note probability estimate. To do this, we generated eleven features for each note by calculating the histogram of sentence-level probability estimates and calculating every 10th percentile (0th, 10th, 20th ..., 80th, 90th, 100th). Those percentiles were fed into a logistic regression model and trained to predict the note-level label of whether the note as a whole contained any confidential content. By using a second model, as opposed to simply taking the maximum sentence-level probability from the first step, the system can take advantage of high-signal cases where a note has multiple high-probability sentences and thus a higher chance of containing confidential content.

Prospective Model Validation

To ensure the model would continue to perform adequately on recent notes, a prospective validation of the NLP modeling was performed. For this validation, 240 clinical progress notes were randomly sampled from visits occurring between May 1, 2022 and May 31, 2022 for patients aged 12 to 17 years. These notes were again assigned to the same five reviewers and annotated according to the same rules. The models trained above were applied to this dataset and note-level performance metrics were calculated.

Implementation into Clinical Operations and Pilot Intervention

Our system currently employs two note types used for adolescents: a regular progress note, which is being prepared for sharing with patient and proxy, and a confidential note type, which is not shared. The sentence-based logistic regression model was deployed into clinical operations as part of an ongoing documentation optimization quality improvement project at our institution that involves a clinic-by-clinic rollout in which providers are educated about adolescent confidentiality and nonconfidential outpatient progress note templates are optimized to decrease note bloat and prepare them for sharing with adolescents and their proxies. Prior to each educational intervention, a manual audit of the clinic's recent progress notes is performed to identify unintentional disclosures of confidential information and other high-risk documentation practices. Findings from the audit are used to inform the intervention.

The language processing model proposed in this manuscript was used to augment the documentation auditing for one of our subspecialty clinics. As part of this pilot intervention, clinical progress notes from patients ages 12 to 17 years seen in this clinic during the month of July 2022 were queried from our EHR database. The language model was applied to these notes to generate note-level risk estimates and to highlight high-risk portions of these notes to aid the manual reviewer.

A note-level threshold of 0.5 was applied to flag notes for potential review. The total number of notes and the number of notes exceeding the 0.5 risk score threshold was calculated per provider. This distribution was visualized to identify outlying providers, for whom the top 20 highest risk notes were manually reviewed. Within reviewed notes, the top 5% highest scoring sentences as identified by the sentence-based model were highlighted to expedite manual review.

Results

Development and Validation of Natural Language Processing Models

Descriptive statistics of the corpus of notes used to train and evaluate the language models and the prospective validation cohort are shown in ▶ **Table 1**. In the initial corpus of 1,200 notes, 255 (21%) contained confidential information. In the validation cohort, the prevalence of notes containing confidential content was 53 of 240 (22%).

Table 1 Demographic information for both the initial cohort of notes used to train and test the model and the prospective validation cohort

	Initial cohort	Validation cohort
Total notes	1,200	240
Notes with confidential information	255 (21%)	53 (22%)
Patient sex		
Female	621 (52%)	112 (47%)
Male	577 (48%)	128 (53%)
Unknown	2 (<1%)	0
Patient age (y)		
12	178 (15%)	29 (12%)
13	209 (17%)	42 (18%)
14	215 (18%)	35 (16%)
15	202 (17%)	48 (20%)
16	193 (16%)	44 (18%)
17	203 (17%)	42 (18%)
Patient race		
Asian	174 (15%)	32 (13%)
Black or African American	21 (2%)	5 (2%)
Native Hawaiian or Other Pacific Islander	11 (1%)	1 (<1%)
White or Caucasian	494 (41%)	81 (34%)
Other/unknown/declines	500 (42%)	121 (50%)
Patient language		
English	1,007 (84%)	186 (78%)
Spanish	153 (13%)	47 (20%)
Other	40 (3%)	7 (3%)

Note: Components of certain categories may add up to more or less than 100% due to rounding.

The note-based logistic regression model achieved an area under the receiver operating characteristic (AUROC) curve of 0.88 on the initial corpus and 0.80 on the validation corpus and an F1 score of 0.72 and 0.55, respectively. The sentence-based logistic regression model achieved an AUROC of 0.90 and 0.88, respectively, and an F1 score of 0.76 and 0.67, respectively. These and other model performance metrics are summarized in ▶ **Table 2**. The sentence-based logistic regression receiver operating characteristic curve and precision-recall curve are shown in ▶ **Fig. 1**.

In addition to providing classification at the level of the entire note, the sentence-based logistic regression model also calculates a risk score for each 10-word sentence in the note. This output can be used to highlight high-risk areas of a note to expedite manual review. An example of this is illustrated in ▶ **Fig. 2**, which demonstrates how this feature might behave on synthetic note text data. The right pane shows high-risk excerpts highlighted by the algorithm from

Table 2 Summary of performance metrics of natural language processing models on the initial cohort test set and on the prospective validation set

	Initial test cohort						Validation cohort					
	AUROC	AUPRC	F1 score	Sensitivity ^a	Specificity ^a	PPV ^a	AUROC	AUPRC	F1 score	Sensitivity ^a	Specificity ^a	PPV ^a
Note-based logistic regression	0.88	0.71	0.72	0.39	0.97	0.77	0.80	0.48	0.55	0.21	0.96	0.61
Sentence-based logistic regression	0.90	0.77	0.76	0.59	0.95	0.81	0.88	0.66	0.67	0.53	0.91	0.60

Abbreviations: AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic; PPV, positive predictive value.

^aSensitivity, specificity, and PPV for logistic regression models are calculated at threshold of $\hat{y} = 0.50$.

notes in the validation cohort. In our pilot intervention, we elected to highlight the top 5% highest risk sentences.

Implementation and Pilot Intervention

In the pilot intervention, the model was used to audit notes from one of our subspecialty clinics from July 2022. During this time, 264 notes were written by 15 different providers. Of these, 60 notes were flagged as high risk. The number of total notes and high-risk notes per provider is shown in ▶Fig. 3. One provider, who had 29 flagged notes, was identified as an outlier. As a result, their highest risk 20 notes were selected for manual review. This review identified frequent use of an automated phrase populated with the tobacco-use history of the patient (pulled from a provider entry social history flowsheet). This documentation practice was noted and informed educational interventions.

Discussion

This study applies NLP to identify confidential content in adolescent progress notes and successfully demonstrates both computational feasibility and clinical utility in a pilot intervention. There are two notable takeaways from this study. First, this work demonstrates the feasibility of using NLP to optimize the technical implementation of information sharing with adolescent patients in a way that protects their confidentiality. This has far-reaching implications not just for adolescents, but for other sensitive topics in medicine as well such as reproductive and maternal–infant health.¹⁹ Furthermore, this study serves as an example of a successful implementation of machine learning into clinical operations, with a human-in-the-loop approach that demonstrates efficiency gains in an otherwise burdensome manual task.

Promoting Adolescent Confidentiality

This study extends the body of work around promoting adolescent confidentiality in the wake of the 21st Century Cures Act and the information sharing mandate. Prior studies by Xie et al and Ip et al have used NLP algorithms to identify inappropriate health portal access by parents and guardians.^{25,26} This occurs when a parent or guardian accesses the health portal with the patient's account instead of a proxy account. In a related study by Lee et al, keyword expansion was used to identify the prevalence of potentially confidential content in adolescent progress notes.²⁷ Similarly, Ni et al also demonstrated the feasibility of using a combination of language processing techniques to identify information around substance use among pediatric patients in their proof of concept study.²⁸ From an operational perspective, Murugan et al enumerated their experience from the “learning mode” deployment of a confidential adolescent note type. They found that the use of autopopulated note elements was a common source of unintentional confidential information disclosure.²⁹

Much of this recent work has focused on characterizing and measuring the issues around adolescent confidentiality. These prior studies lay the foundation for our work in which we develop and test a potential, scalable solution to support

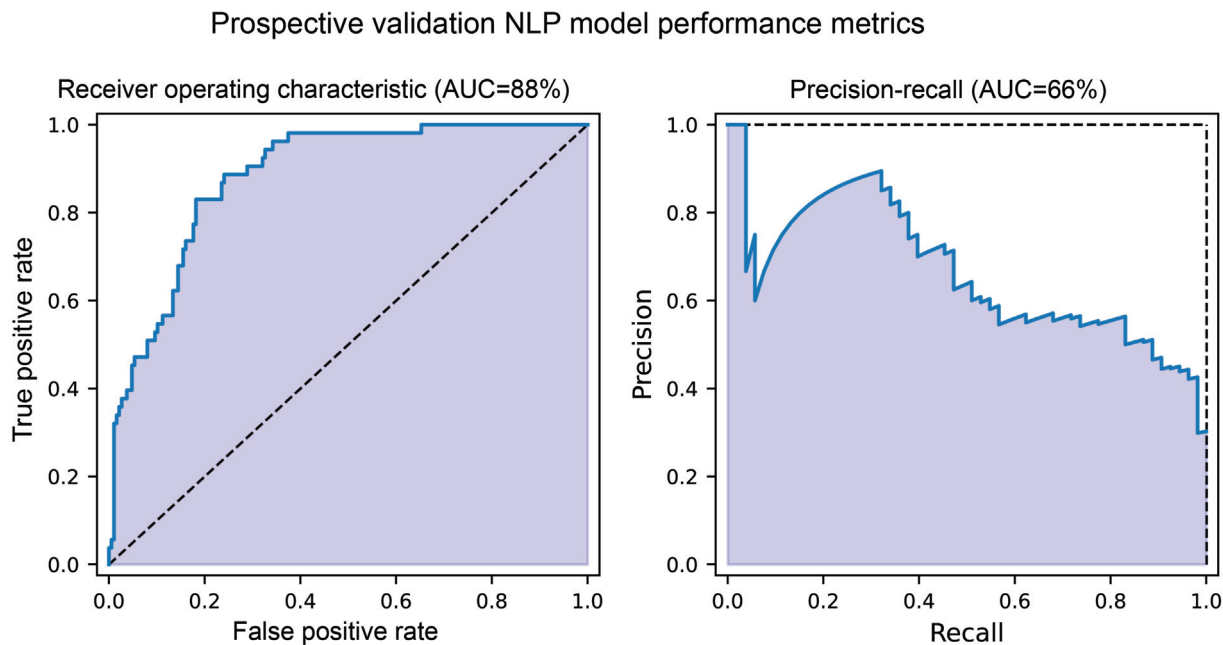


Fig. 1 Note-level performance metrics, including (A) receiver operator curve and (B) precision–recall curve for the sentence-based model selected for implementation into clinical operations based on performance on the prospective validation set. AUC, area under curve; NLP, natural language processing.

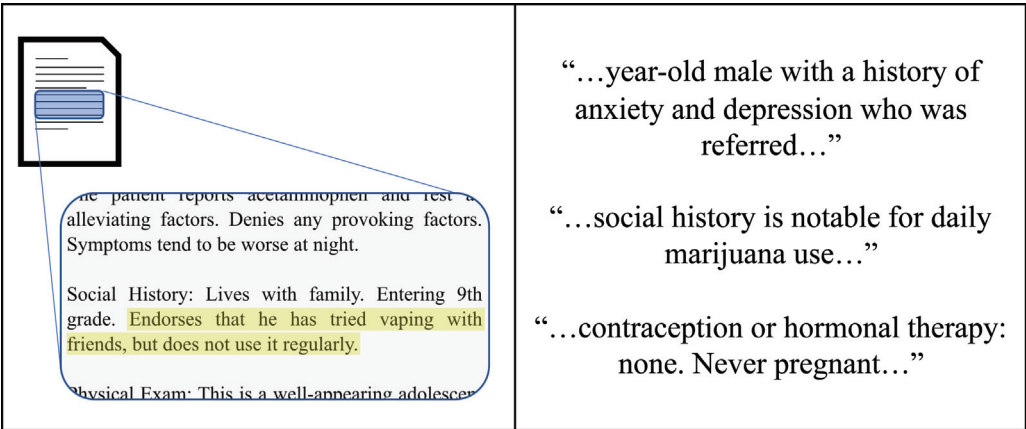


Fig. 2 (A) Example of synthetic text highlighted with the sentence-based model; (B) examples of sentences identified by the model on the validation set.

adolescent information sharing using NLP. In its current form, the algorithm is used to more efficiently audit adolescent progress notes for unintentional disclosures of confidential content. Findings from these audits inform targeted feedback and educational interventions. As the algorithm is improved, it may eventually be used autonomously to support such efforts. We envision advancements from this line of work may enable health systems that are currently not releasing adolescent notes due to the technical feasibility exception of the mandate to begin sharing this information to their patients in the near future.³⁰

Efficiency Gains from Algorithm Implementation

The human-in-the-loop implementation of our pilot intervention demonstrates the successful use of a machine learning

algorithm to augment the efficiency of a manual task. While our algorithm does not operate autonomously, this method of implementation still yields important operational gains.

Consider the context in which our algorithm was deployed. Previously, as part of an ongoing documentation optimization program, a random sampling of notes was periodically obtained and reviewed by a group of physicians on a specialty-by-specialty basis. The manual review of clinical progress notes for confidential content is resource-limited by provider availability. Our study demonstrated that approximately 20% of these notes had confidential content. Compare this to the proposed NLP algorithm, in which sampling notes with a risk score of 0.5 or higher yields a positive predictive value (PPV) of greater than 60%, already a three times gain in efficiency in sampling for manual review.

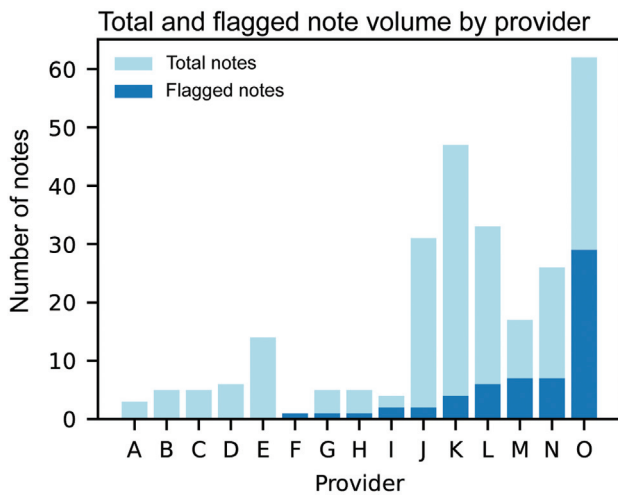


Fig. 3 Summary of results from pilot intervention depicting the number of notes at high risk for containing confidential information by provider (anonymized).

Consider further that the sentence-highlighting feature of our algorithm also expedites the manual review of those selected notes. For example, if a health system had the capacity to manually review 20 notes in a given time period, this would have yielded 4 notes with confidential content previously. However, suppose that with the use of the proposed sentence-based algorithm, which expedites manual review, this capacity would increase to reviewing 40 notes in the same amount of time. If these notes were chosen in a risk-stratified way with a PPV of at least 60%, this would yield 24 notes with confidential content, a total of six times increase in efficiency in identifying notes with confidential content.

Natural Language Processing for Supporting Clinical Operations

Additionally, this study is also an illustration of the successful deployment of an NLP model to improve clinical operations and builds upon a growing body of literature in this space. Other examples of language models deployed to support care delivery include medical coding, clinical trial recruitment, and chatbots for medical triage and education.^{31–33} Outside of the scientific literature there are also third-party software that employ NLP models to review records for purposes of billing, quality metrics, and clinical decision support.^{34,35}

While the focus of our use case is adolescent confidentiality, there are lessons learned that are generalizable to other issues in clinical operations. For example, we observed that our language model, which was trained on featurized sentences (the “sentence-based model”), outperformed the “note-based model,” which was trained using the entire featurized note text. This suggests that the identification of confidential information is a relatively localized task, as in, it generally requires only looking at specific parts of a note.

Furthermore, the use of similar methods may be applied to other pressing issues in health information confidentiality, including protecting health information in the context of maternal–infant health and reproductive health, the latter being particularly pertinent in the wake of *Dobbs v. Jackson*.^{19,36,37}

Limitations and Future Work

There are important limitations to this work. First, while the implementation of our algorithm promises efficiency gains in documentation optimization efforts, it is not yet accurate enough to operate autonomously and requires human input in its current form, which limits its scalability. Additionally, the model is limited by its sensitivity, 59 and 53% in our testing and prospective validation experiments, respectively (at a threshold of $\hat{y} = 0.5$). As a result, it cannot yet be used to identify all notes with confidential content. Additionally, because the algorithm was trained on data from a single site, it may be learning institutional patterns (e.g., relying on specific note templates or author writing styles) that are not generalizable to other settings. As such, the model may require retraining or fine-tuning at external sites. Similarly, because the training data were annotated using the California minor consent laws, this limits the algorithm’s generalizability to states with differing rules.

Future work will focus on developing a data pipeline that will allow for the ongoing monitoring of clinical notes. We envision this may be achieved through the creation of a dashboard that visualizes risk scores from the model. Additionally, there is work ongoing to establish an operational workflow that will allow for the continued refinement of the model to combat data drift and improve its accuracy, including a process to identify and review misclassifications.

Conclusion

This study illustrates the development and successful implementation of an NLP algorithm to identify confidential content in adolescent progress notes. Our human-in-the-loop deployment into clinical operations demonstrates significant efficiency gains in the manual task of clinical note review. The proposed system shows promise as an operational solution to support the health information sharing with adolescents in a way that maintains patient confidentiality.

Clinical Relevance Statement

This manuscript describes the development and implementation of an NLP algorithm that has been directly deployed into health care operations to support information sharing and adolescent patient confidentiality. Our study includes elements of health information exchange, NLP, application/implementation of machine learning, and health IT regulation/policy with direct clinical relevance.

Multiple-Choice Questions

1. In California State, adolescents may consent to care relating to the following domains
 - a. Sexual/reproductive health
 - b. Mental health
 - c. Substance use
 - d. All of the above

Correct Answer: The correct answer is option d. Minor consent laws allow adolescents to consent to certain medical services without parent or guardian involvement. These laws vary state-by-state. In California, patients aged 12 and older may generally consent to care around reproductive/sexual health, mental health, and substance use without parental or guardian involvement.

2. Which of the following are an effective way of using machine learning to estimate whether a clinical note contains confidential content?
 - a. Training a logistic regression model to directly estimate whether a note contains confidential content
 - b. Training a logistic regression model to estimate whether a sentence contains confidential content and then taking the maximum probability for each sentence in the note
 - c. Training a logistic regression model to estimate whether a sentence contains confidential content and then feeding the probabilities into another logistic regression model
 - d. Training a deep neural network to estimate whether a note contains confidential content

Correct Answer: The correct answer is option c. As we have limited data, logistic regression models are generally the best option. And in that setting, sentence-level models tend to outperform note-level models. Finally, using a second logistic regression model to aggregate the probabilities from the sentence model yields better estimates of the note-level probability.

3. In what ways could information sharing through a health portal result in breach of adolescent confidentiality?
 - a. Parent/guardian may receive an explanation of benefits from insurance regarding confidential medical care
 - b. Confidential information may be inadvertently released to a proxy health portal account
 - c. Parents or guardians may overhear a confidential part of the patient visit from outside the room
 - d. Confidential information may be accidentally relayed to a parent/guardian by a phone encounter with clinic staff

Correct Answer: The correct answer is option b. Information sharing through the patient portal as mandated by the 21st Century Cures Act may cause unintentional breaches of adolescent confidentiality. As minors, adolescent patients cannot consent to all types of medical care. As a result, health portals for this population typically have two types of accounts: an account for the adolescent patient and a proxy account for the parent/guardian. Confidentiality may be breached if information is accidentally released to both the patient and proxy accounts. While choices c and d also explain potential ways that adolescent confidentiality may be breached, they are not related to information sharing through the electronic health portal.

Protection of Human and Animal Subjects

The presented work was performed as part of a quality improvement effort at our institution and does not qualify as human subjects research.

Funding

None.

Conflict of Interest

None declared.

References

- 1 Office of the National Coordinator for Health Information Technology 21st Century Cures Act: interoperability, information blocking, and the ONC health IT certification program [Internet]. 2020. Accessed July 22, 2022 at: <https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification>
- 2 Holmgren AJ, Patel V, Charles D, Adler-Milstein J. US hospital engagement in core domains of interoperability. *Am J Manag Care* 2016;22(12):e395–e402
- 3 Nazi KM, Turvey CL, Klein DM, Hogan TP, Woods SSVA. VA OpenNotes: exploring the experiences of early patient adopters with access to clinical notes. *J Am Med Inform Assoc* 2015;22(02):380–389
- 4 Mishra VK, Hoyt RE, Wolver SE, Yoshihashi A, Banas C. Qualitative and quantitative analysis of patients' perceptions of the patient portal experience with OpenNotes. *Appl Clin Inform* 2019;10(01):10–18
- 5 Delbanco T, Walker J, Bell SK, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Ann Intern Med* 2012;157(07):461–470
- 6 Walker J, Leveille S, Bell S, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. *J Med Internet Res* 2019;21(05):e13876
- 7 Wright E, Darer J, Tang X, et al. Sharing physician notes through an electronic portal is associated with improved medication adherence: quasi-experimental study. *J Med Internet Res* 2015;17(10):e226
- 8 Pageler NM, Webber EC, Lund DP. Implications of the 21st Century Cures Act in pediatrics. *Pediatrics* 2021;147(03):e2020034199
- 9 Carlson J, Goldstein R, Hoover K, Tyson N. NASPAG/SAHM Statement: the 21st Century Cures Act and adolescent confidentiality. *J Adolesc Health* 2021;68(02):426–428
- 10 Schapiro NA, Mihaly LK. The 21st Century Cures Act and challenges to adolescent confidentiality. *J Pediatr Health Care* 2021;35(04):439–442
- 11 Reddy DM, Fleming R, Swain C. Effect of mandatory parental notification on adolescent girls' use of sexual health care services. *JAMA* 2002;288(06):710–714
- 12 Vukadinovich DM. Minors' rights to consent to treatment: navigating the complexity of State laws. *J Health Law* 2004;37(04):667–691
- 13 Pathak PR, Chou A. Confidential care for adolescents in the U.S. health care system. *J Patient Cent Res Rev* 2019;6(01):46–50
- 14 Pampati S, Liddon N, Dittus PJ, Adkins SH, Steiner RJ. Confidentiality matters but how do we improve implementation in adolescent sexual and reproductive health care? *J Adolesc Health* 2019;65(03):315–322
- 15 Sharko M, Jameson R, Ancker JS, Krams L, Webber EC, Rosenbloom ST. State-by-state variability in adolescent privacy laws. *Pediatrics* 2022;149(06):e2021053458

- 16 Ginsburg KR, Slap GB, Cnaan A, Forke CM, Balsley CM, Rouselle DM. Adolescents' perceptions of factors affecting their decisions to seek health care. *JAMA* 1995;273(24):1913–1918
- 17 Ford CA, Millstein SG, Halpern-Felsher BL, Irwin CE Jr. Influence of physician confidentiality assurances on adolescents' willingness to disclose information and seek future health care. A randomized controlled trial. *JAMA* 1997;278(12):1029–1034
- 18 Lothen-Kline C, Howard DE, Hamburger EK, Worrell KD, Boekeloo BO. Truth and consequences: ethics, confidentiality, and disclosure in adolescent longitudinal prevention research. *J Adolesc Health* 2003;33(05):385–394
- 19 Arvisais-Anhalt S, Lau M, Lehmann CU, et al. The 21st Century Cures Act and multiuser electronic health record access: potential pitfalls of information release. *J Med Internet Res* 2022;24(02):e34085
- 20 Parsons CR, Hron JD, Bourgeois FC. Preserving privacy for pediatric patients and families: use of confidential note types in pediatric ambulatory care. *J Am Med Inform Assoc* 2020;27(11):1705–1710
- 21 Bedgood M, Kuelbs CL, Jones VG, Pageler N. Organizational perspectives on technical capabilities and barriers related to pediatric data sharing and confidentiality. *JAMA Netw Open* 2022;5(07):e2219692
- 22 Bedgood M, Rabbani N, Brown C, et al. The prevalence of confidential content in adolescent progress notes prior to the 21st Century Cures Act information blocking mandate. *Appl Clin Inform* 2023
- 23 Ramos J Using TF-IDF to determine word relevance in document queries. In: *Proceedings of the First Instructional Conference on Machine Learning*. International Conference on Machine Learning; 2003
- 24 Ng AY Feature selection, L1 vs. L2 regularization, and rotational invariance. In: *Proceedings of the Twenty-First International Conference on Machine Learning*; 2004;78
- 25 Ip W, Yang S, Parker J, et al. Assessment of Prevalence of Adolescent Patient Portal Account Access by Guardians. *JAMA Netw Open* 2021;4(09):e2124733
- 26 Xie J, McPherson T, Powell A, et al. Ensuring adolescent patient portal confidentiality in the Age of the Cures Act final rule. *J Adolesc Health* 2021;69(06):933–939
- 27 Lee J, Yang S, Holland-Hall C, et al. Prevalence of sensitive terms in clinical notes using natural language processing techniques: observational study. *JMIR Med Inform* 2022;10(06):e38482
- 28 Ni Y, Bachtel A, Nause K, Beal S. Automated detection of substance use information from electronic health records for a pediatric population. *J Am Med Inform Assoc* 2021;28(10):2116–2127
- 29 Murugan A, Gooding H, Greenbaum J, et al. Lessons learned from OpenNotes learning mode and subsequent implementation across a pediatric health system. *Appl Clin Inform* 2022;13(01):113–122
- 30 Office of the National Coordinator for Health Information Technology Cures Act final rule: information blocking exceptions [Internet]. 2022. Accessed September 12, 2022 at: <https://www.healthit.gov/sites/default/files/2022-07/InformationBlockingExceptions.pdf>
- 31 Campbell S, Giadresco K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *HIM J* 2020;49(01):5–18
- 32 Woo M. An AI boost for clinical trials. *Nature* 2019;573(7775):S100–S102
- 33 Maguire H, Elana M, Roy R, Vivian L, Washington V, Volpp KG Asked and answered: building a chatbot to address Covid-19-related concerns. *Catalyst non issue content* 2020. Accessed April 14, 2023 at: <https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0230/>
- 34 3M CodeAssist System [Internet]. Accessed September 12, 2022 at: https://www.3m.com/3M/en_US/health-information-systems-us/improve-revenue-cycle/coding/professional/code-assist/
- 35 Health Language NLP for unstructured data [Internet]. Accessed September 12, 2022 at: <https://www.wolterskluwer.com/en/solutions/health-language/clinical-natural-language-processing>
- 36 Scibilia JP. How to protect maternal health information in newborn's medical record. *AAP News* 2014;35(12):4–4
- 37 Spector-Bagdady K, Mello MM. Protecting the privacy of reproductive health information after the fall of Roe v Wade. *JAMA Health Forum* 2022;3(06):e222656–e222656