SenRev: Measurement of Personal Information Disclosure in Online Health Communities

Faysal Hossain Shezan
University of Virginia
Charlottesville, USA
fs5ve@virginia.edu

Minjun Long
University of Virginia
Charlottesville, USA
ml6vq@virginia.edu

David Hasani
University of Virginia
Charlottesville, USA
dh8rsv@virginia.edu

Gang Wang
University of Illinois at Urbana-Champaign
Urbana-Champaign, USA
gangw@illinois.edu

Yuan Tian
University of California Los Angeles
Los Angeles, USA
yuant@ucla.edu

ABSTRACT

With life style shifting during the pandemic, online health communities start to attract more users (including healthcare workers and patients) to discuss health-related questions. While such online platforms provide convenience to users, with health-related information shared broadly over text and images (*e.g.,* X-Ray scans, photocopies of documents), they also raise questions regarding privacy. In this paper, we propose SenRev to systematically measure the leakages of sensitive information in those publicly available discussions. We use SenRev to analyze 1,894,900 multi-modal and multi-lingual data elements from four different online health communities. We find that sensitive data leakages are common; overall 1,324,064 (69.88%) pieces of evidence of data leakages are detected, with 23,587 (1.78%) of them involving identifiers and 1,300,477 (98.22%) involving quasi-identifiers. Surprisingly, leakages through medical images occur more frequently in the community of healthcare professionals compared with other communities. Finally, based on our results, we discuss the potential directions for countermeasures.

KEYWORDS

Privacy, Personal Information Leakages, Online Health Communities

1 INTRODUCTION

During the recent pandemic, online health communities (OHC) have become increasingly popular amongst patients and healthcare professionals [8, 10, 77, 96], as people tend to avoid going outside, including– visiting healthcare facilities [51, 67]. OHC such as Sermo, Doximity, DailyRounds, and PatientInfo are attracting users from all over the world [17, 88]. These sites connect patients with caregivers, physicians, clinicians, nurses, and even other individuals (non-healthcare-workers) to share their medical problems and seek experts' suggestions [41]. To describe their symptoms, people often share information about their family history, medical history, and images of relevant testing results on these websites [47].

The benefits of OHC are not only limited to patients—healthcare professionals also use OHC to get suggestions from other experts regarding their patients' cases [36, 92]. Despite the benefits, such discussions can also increase the risks of leaking patient information. When sharing a patient's case, healthcare professionals usually include available medical images and detailed information about the patient [31, 85]. While such contextual information is helpful to explain the patient's conditions [72], oversharing can also lead to real-world consequences. For example, using publicly shared information, adversaries may be able to link a medical condition (e.g., cancer) to the patient's real-world identity. Such re-identification [33, 36, 42, 95] can further lead to personal attacks such as blackmail, discrimination, harassment, or bullying [5, 84].

To prevent sensitive information leakages, most OHC websites only warn the users not to reveal sensitive personal information in the *terms and conditions* during the registration time. More active protection mechanisms are currently missing. Prior works have investigated information sharing behaviors in online social networks [44, 64, 65, 70, 82] and visual lifelogging [54, 69]. However, there is a lack of systematic understanding of the data leakages in OHC.

Our Goal. In this paper, our goal is to empirically measure the prevalence and patterns of sensitive data leakages in OHC (both from mobile and web platforms). We consider the threat model of a re-identification attack where adversaries use OHC data to link a medical condition/disease to a user's real-world identity. We focus on analyzing leakages of two types of information: (1) identifiers (name, email, phone number, face, national ID number) that can be used to re-identify people, and (2) quasi-identifiers (date of birth, age, sex, location, medical history) which may not directly reveal identity but can be used to link the user records across (external) datasets [36, 42].

Challenges. There are several key challenges for analyzing sensitive information in OHC. First, the heterogeneity of health-related data makes it difficult to accurately identify different sensitive information fields. Information extraction technique in text differs from that of images. Moreover, there are a variety of images shared on OHC, such as photocopies of medical documents, X-Ray, and even images of faces. These images are of different sizes and resolutions. Finally, the communication languages vary across different platforms. As a result, we need to analyze multi-language text (both structured and unstructured) from those medical images and text.

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–19

https://doi.org/XXXXXXX.XXXXXXX

Our Approaches. We develop a system called SenRev to analyze the data shared on public OHC sites. Considering the potential privacy implications of such data analysis, we focus on publicly available data and have obtained approvals from our local IRB-Institutional Review Board (see the detailed ethics discussion in Section 7). Our dataset contains two types of data: (a) text descriptions and (b) medical images. People describe in detail their medical questions through text descriptions. Whereas, to provide a clear view of their current medical state, they attach the medical images with the post. We found three types of medical images: (1) images of medical documents, (2) X-Ray images, and (3) face images. We propose a novel design, SenRev to analyze this multi-modality of data to identify sensitive data leakages. We first design classifiers to detect images of medical documents, X-Ray images, and face images as these images might include sensitive data leakage. Then for images of medical documents, and X-Ray images that might contain textual information, we use OCR (Optical Character Recognition) to extract the text. Text descriptions in the posts and X-Ray images contain unstructured textual information, while images of medical documents contain structured textual information. As a result, we develop tools to extract sensitive data fields that support both structured and unstructured textual data (Section 4). As for face images, we detect if the face is obfuscated or not. SenRev incorporates multi-languages to extract sensitive information from the multi-modality of OHC data.

Evaluation results show that SenRev can identify and extract multi-languages with an F1-Score of 99% (Section 5). Overall, the ground-truth evaluation shows that SenRev achieves 93.72% true positive with a low false negative (2.55%) in terms of detecting sensitive information with a negligible computation overhead (Appendix A.6).

We investigate 11 OHC sites and find that five of them restrict their data access to verified healthcare professionals. We manually inspect the rest and locate four OHCs with some signs of information leakages. We then apply SenRev to analyze sensitive data leakages on these four OHCs. We investigate a total of 200,683 posts and 1,694,217 comments (1,894,900 data samples in total). Using SenRev, we observe 1,324,064 (69.88%) sensitive data leakages in those four websites. We find that these leakages happen not only in posts but also in comments (on an average of 34.82% across four websites). Out of the 1,324,064 leakages of sensitive information, 23,587 (1.78%) involve identifiers and 1,300,477 (98.22%) involve quasi-identifiers. While identifier leakage has a lower percentage, the absolute number of leakages is non-trivial (23K). Compared to posts made by regular users, we find that posts from healthcare professionals contain more sensitive information. While we occasionally observe some efforts to redact and obfuscate sensitive information fields, such efforts are ad-hoc and insufficient. We find that people put less or no effort into obfuscating face images (identifier). They tend to obfuscate names, UIDs, phone numbers, and DOB more often compared to other sensitive information. Finally, based on our results, we provide several suggested countermeasures to prevent sensitive data leakages in OHC.

Contributions. We have the following contributions.

- **New Tool.** Our tool systematically identifies sensitive data leakage in online health communities both in image and text. We open-source our tool at https://github.com/faysalhossain2007/SenRev for facilitating future research.
- **New Measurement Results.** We apply our tool to analyze four online health communities to explore the sensitive data flow pattern and the leakage caused by different interactions. Our findings are alarming.

We organize the rest of the paper as follows. Section 2 discuss how people use OHC, share their information, and compare the works related to ours, Section 3 demonstrates the detailed process of the threat model and our data collection, Section 4 discuss design architecture and the implementation details of SenRev, Section 5 evaluates our tool's performance including measurement analysis and case study, we propose our countermeasure in Section 6, we include our ethical discussion in Section 7, we illustrate some of the key insights based on our analysis and limitations in Section 8, lastly, we conclude our paper in Section 9.

2 BACKGROUND AND RELATED WORK

We first introduce how user shares their information in OHC. Then, we discuss works related to ours.

2.1 How Does OHC Work?

In OHC, users post questions seeking suggestions about medical diagnoses and treatments, which often requires sharing medical information. The discussions contain two types of data: text descriptions, and images. We further divide images into three categories: (1) images of the medical documents, (2) X-Ray images, and (3) face images. These three categories of images are widely used to facilitate OHC discussions and might leak sensitive information. **First**, the image of a medical document often contains sensitive information about a patient's identification number, medical history, and family history. They may also contain patients' progress notes, prescriptions, operative notes, physician certification, physician orders, therapy notes, and emergency room records. **Second**, X-Rays images are often shared in OHC. These images not only show tissues and structures inside the human body, but also contain the patient's sensitive information such as name, age, and sex [47, 85]. **Third**, users may attach an image of the body part (*e.g.*, faces, hands) with the post to show the current condition of infections or diseases.

2.2 Type of OHC Users

There are two types of users in the current OHC. We call them "healthcare professionals" and "anyone". Note that OHCs often require verifying the background of healthcare professionals[1]. The healthcare professionals broadly include medical student, medical doctor, psychiatrist, psychotherapist, pharmacist, medical assistant, nurse assistant, physician assistant, nurse, optometrist, paramedic, and podiatrist. The rest of the users are considered as "anyone". In the rest of the paper, we refer to "Healthcare Professional" as "H", and "Anyone" as "A".

[1] To get certified as a healthcare professional in OHC, one needs to submit proof documents that contain national provider identifier number, medical education number, medical regulatory authority, and/or medical license number. Moderators will manually verify the submitted documents before certifying the user's identity and granting the corresponding privilege.

We find three main types of OHCs based on the interaction of healthcare professionals (H) and anyone (A) else: (1) A-H (where *anyone* can seek help and verified healthcare professional can respond), (2) H-H (only verified healthcare professional can post and respond), and (3) A-A (anyone can post or comment). In this work, we select popular OHCs that represent different OHC types for our analysis. More specifically, we focus on DailyRounds (H-H), IIYI (A-H), DoctorsLounge (A-H), and PatientInfo (A-A). Note that PatientInfo does not have any verification system to certify healthcare professionals, and thus we consider all the users as "anyone".

2.3 Information Leakage through OSN

Prior works have characterized various privacy problems in online social networks (OSNs) [30, 56, 58, 64] and explored solutions to resolve these problems [29, 46, 48, 82, 87]. For example, researchers find that people are unaware of the data that they are sharing [25, 49]; users may compromise the privacy of others through annotations [60], leak their location information through shared images [26], leak sensitive attributes via their social connections. A number of studies are focused on re-identification attacks by establishing links between popular social networks [42], or local graph structures [57, 59, 91], or via matching with public datasets [33, 55, 95] to re-identify people from data. Related to online communities, researchers have studied lifeloggers' groups and their privacy issues during image sharing [69]. Compared to these works, we investigate the multi-modality of medical data and seek to identify privacy violations in OHC.

2.4 Medical Image and Text Analysis

Researchers have performed analysis on clinical data for various applications such as characterizing the importance of contextual medical information [72], predicting diseases [66], and summarizing patient's health issues [40]. First, prior works have developed diagnosis models based on medical images. Examples include models to detect lymph nodes [75] and pancreas segmentation [74] in CT (Computed Tomography) images, neuronal membrane segmentation in electron microscopy images [35], and knee cartilage segmentation in MRI (magnetic resonance imaging) scans [68]. Second, prior works have performed text analysis to support medical applications. For example, Zhang *et al.* have implemented a tool to respond to clinical questions [94]. Several works have investigated the benefits of contextual information in clinical domains [61, 80]. Alsentzer *et al.* propose clinical BERT embedding to further improve the performance of DNN models for analyzing the medical text data [27]. Finally, researchers also use the online social network to predict different diseases [79]. Unlike those works, we focus on the *text* and *image* data shared in OHC to detect privacy leakages.

3 THREAT MODEL & DATA COLLECTION

Threat Model. In this paper, we seek to empirically measure the prevalence and patterns of sensitive data leakages in OHC.

We define the following information as "sensitive" which needs to be protected from unauthorized parties [37, 50]. They are– name, email, DOB, age, phone number, sex, patient's face image, UID, location (geographic subdivisions smaller than a state), medical history (chronic diseases which affect an individual for extended periods). All this information is identified as protected health information by HIPAA [1], DISHA [21], PIPL [18]. We categorize social security number (SSN) [15], citizen ID (SSN-equivalent for Chinese websites), and aadhaar number (Indian national ID) as UID.

The threat model is primarily focused on re-identification risks of linking a user's medical condition to their real-world identity using OHC data [33, 36, 42, 95]. Such re-identification may further lead to personal attacks such as blackmail, discrimination (e.g., during job search or dating), harassment, or bullying [5, 84]. Under this threat model, we focus on two types of leakages:

- **Identifier:** Information such as name, email, phone number, face, and UID can be used to directly re-identify people.
- **Quasi-identifier:** Information such as DOB, age, sex, location, and medical history may not directly reveal identity but can be used to *link user records* across (external) datasets to re-identify users [36, 42].

In this paper, we focus on the *behavior* of information leakage rather than the *intention* of leakage since it is difficult to judge whether the information is leaked intentionally or due to oversight.

Datasets. To answer our research questions, we constructed a dataset by collecting publicly available posts and comments from OHCs. First, we make an initial list of 11 different OHCs based on their popularity according to Alexa ranking (Alexa ranking is primarily based on traffic volume) [13]. Alexa determines website popularity by estimating the number of daily unique visitors and pageviews in the preceding three months [20]. The idea is to focus on OHCs that are used/visited by more people. Then, we follow two criteria to filter out 8 websites: (1) *publicly accessible:* we only consider OHCs where the data is publicly available. Following this criterion, we remove 5 websites that require verification of medical background before users can access the website data. (2) *sign of leakages:* we then perform an in-depth analysis on the 7 remaining websites by visiting those sites and manually analyzing 300 posts from each site. We find that two websites (*e.g.,* "iCliniq", "Figure1") do not show any evidence of sensitive data leakages. This is likely due to the active presence of moderators who would manually remove any sensitive information they observe. Note that the sizes of these websites are fairly small (*e.g.,* each site has only 400 posts on average)—which may have made it possible for manual reviews by moderators. After filtering out these websites, we have 4 popular OHCs (DailyRounds, IIYI, DoctorsLounge, PatientInfo) for our study.

We capture all the user-to-user interactions by visiting the "Forum" pages of those websites. From that page, we collect the list of categories and sub-categories. Then, we visit each of the categories and all the listed sub-categories to extract post content. From each post, we fetch text descriptions and times of the posts, comment(s), author information, category of the post, and attached medical images. For collecting data from these websites, we build a selenium-based python crawler. For the mobile-based platform (i.e., DailyRounds), we collect the API endpoints of the posts using proxy [2]. Then, from those endpoints, we ran our crawler to collect all the post-related information as described above. In this way, we create our full dataset (Table 1).

As shown in Table 1, the four sites are "DailyRounds" (popular in India, with 28,710 posts/comments), "IIYI" (popular in China,

Platform	Acro.	Type	#Post	#Cmnt.	#Total
DailyRounds	DR	H-H	5,005	23,705	28,710
IIYI	IIYI	A-H	45,173	246,711	291,884
DoctorsLounge	DL	A-H	26,464	47,188	73,652
PatientInfo	PI	A-A	124,041	1,376,613	1,500,654

Table 1: Summary of Our Datasets.

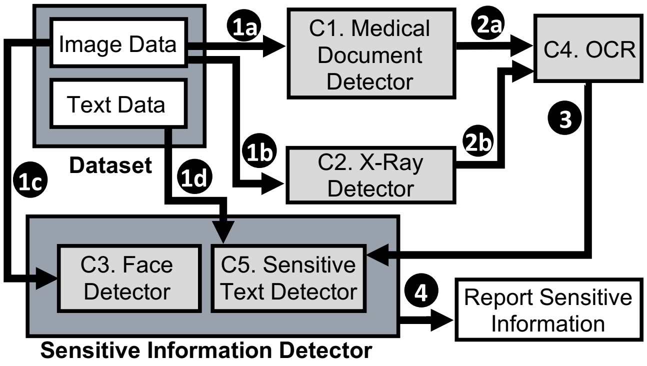


Image Data
Text Data
Dataset
1a
C1. Medical Document Detector
2a
C4. OCR
3
1b
C2. X-Ray Detector
2b
1c
1d
C3. Face Detector
C5. Sensitive Text Detector
4
Report Sensitive Information
Sensitive Information Detector

Figure 1: System architecture of SenRev.

with 291,884 posts/comments), “DoctorsLounge” (popular in the USA, with 73,652 posts/comments), and “PatientInfo” (popular in the USA, with 1,500,654 posts/comments). Only DailyRounds is a mobile application, and the others are web-based platforms. As shown in Table 1, the dataset contains a total of 1,894,900 posts and comments. The ethical justifications for data collection are discussed in Section 7.

4 SYSTEM DESIGN & IMPLEMENTATION

We face three major challenges while building SenRev. First, we have a multi-modality of data containing medical images and text descriptions. Second, there are different types of medical images, where the information extraction and processing techniques are different. Third, we have to analyze multi-lingual data. As is shown in Figure 1, we aim to detect privacy leakage in both images and text. For text, we build a sensitive text detector (C5) to identify sensitive medical information. For the image data, we observe that the privacy leakage is different based on the types of images. As a result, we design C1, C2, and C3 to detect the three types of medical images, and then extract the text from the medical images using C4. Finally, we integrate privacy leakage detection from both text and image to report the overall data leakage. In the following, we explain the design of each component.

C1. Medical document detector. In the medical document, we can find the type of information by searching the representative form field. That’s why we want to separate images of medical documents from the other two types of medical images using C1 (1a in Figure 1). Motivated by the performance of existing works on image analysis, we start building this component by selecting four different DNN models (Xception [34], VGG [81], Inception [83], Resnet [52]) and calculating the performance of each of those models on our data. However, we have limited annotated data to train a DNN model. Labeling a large amount of data is very expensive. So, we start with initializing the weight vector of the models with “ImageNet” [38]. We fine-tune this pre-trained model using our limited labeled data (93 images of the medical document and 541 images of the non-medical document). Three computer science students label those data and resolve the confusion by taking the majority vote. Using transfer learning, we overcome the cost of labeling large amounts of data. We use the grid search approach to select the proper combination of hyperparameters (Appendix A.1). We set the hyperparameters (learning rate = 0.001, epoch = 40, batch size = 32, dropout = 0.2, optimizer = adam, activation = softmax, loss = categorical_crossentropy) based on the performance of the 159 validation data (28 images of medical document and 131 images of non-medical document). Finally, to find the best model, we compared the performance of each of those models with the same set of test data.

C2. X-Ray detector. Unlike medical document images, X-Ray images contain sensitive information without any representative form field (*i.e.* unstructured). To identify the X-Ray images 1b in Figure 1, we calculate the histogram of RGB values of images. However, this process (HRGB) results in a high misprediction rate due to the variation of contrast values in X-Ray images. Fascinated by the performance of different DNN models (Xception [34], VGG [81], Inception [83], Resnet [52]) on the existing work and also on detecting the images of the medical document, we evaluate their performance on detecting X-Ray images as well. Similar to the previous approach, we solve the problem of limited labeled data by initializing the weight vector of the models with “ImageNet” [38]. We used 634 labeled images (272 X-Ray images and 362 non-X-Ray images) to fine-tune the model. Similar to C1, we use grid search to identify the proper set of hyperparameters (Appendix A.1. We configure the best combination of hyperparameters (learning rate = 0.001, epoch = 50, batch size = 32, dropout = 0.2, optimizer = adam, activation = softmax, loss = categorical_crossentropy) by checking the performance on 159 validation data (65 X-Ray images and 94 non-X-Ray images).

C3. Face detector. Apart from medical documents and X-Rays, there is another type of medical image that can leak sensitive information, *i.e.*, face images. According to our threat model, we mark the human face as an identifier. We develop our face detector component (1c in Figure 1) to detect human faces in images. First, we build our baseline using haar cascades [86]. But the performance of this model is poor due to a high false-positive rate. Then, we use dual shot face detector [39, 62] to identify the human faces. That model also performs worse as it identifies images without revealing any identifiable information (e.g., images with only a partial face or blur-out faces) as leakage. We experience the same problem with existing facial recognition datasets as they also contain partial face images and blur images. Models trained on these datasets will treat partial or blur-out faces as leakage. Similar to the other two components (Medical document and X-Ray detector), we leverage the notion of transfer learning [53, 71, 76, 78, 93] to detect face images. But this time, we are not able to achieve a good performance due to a very small amount of face data in our dataset (22 in total). We experience this imbalanced data issue due to the nature of those websites. To this end, we perform data augmentation by randomly rotating each of the face images, and then use active learning to train the model with augmented data. We have discussed the detailed process in Appendix A.2. After this process, we have 2,504 labeled images (1,630 faces and 604 non-face). Again, we follow a grid search (illustrated in Appendix A.1) to select the hyperparameters. We find the best combination of hyperparameters (learning rate = 0.001, epoch = 40, batch size = 32, dropout = 0.1, optimizer = adam, activation = softmax, loss = categorical_crossentropy) by validating the performance on 232 validation data (30 face and 202 non-face images).

C4. OCR. We extract the embedded text in medical images using our OCR component (2a and 2b in Figure 1). First, we use Google's OCR tool, 'Tesseract' [16] to pull out the text from the medical images. We find that this tool worked well on the images of the medical document with 97% F1-Score, but fail to accurately extract text from X-Ray images (68% F1-Score). Then, we investigate the performance of Amazon's "Rekognition" tool[14]. This tool performs well for both types of medical images. However, some of our data (medical images from the IIYI website) contains Chinese text that the Amazon Rekognition tool fails to extract[2]. To solve multiple language problems, we use the OCR tool developed by Baidu [7]. It can identify both English and Chinese text at any place in an image. To keep the false-negative rate low, we construct our OCR component by taking the combined output from these two OCR tools.

C5. Sensitive text detector. We observe the patterns of sensitive text embedded inside the data which we can capture using a rule-based approach. We incorporate those rules into our sensitive text detector component (1d and 3 in Figure 1) to detect sensitive text (both from text descriptions and images of medical documents and X-Ray). We build two rule-based approaches.

First, it is relatively easier to detect sensitive information from screenshots of medical documents. Unlike X-Ray images, medical records/documents are highly structured (confirmed with manual examination). As such, we can directly search for the desired data fields and map out their values. More specifically, once we locate the desired data field, we consider the next extracted text as the corresponding value of that field. For example, to identify *name* information, we first look for the "name" field in the extracted text, then mark the next extracted text as the value of the "name".

Second, X-Ray images usually contain unstructured data fields. To extract the needed information, we build rules using the method described in Table 2. To handle different languages, we have separated the rules for Chinese websites presented in Appendix A.3. Some of the data fields (*e.g.,* name) listed in Sec 3 are not trivial to detect in our dataset. We need to incorporate that knowledge to detect sensitive text in our data. In the following, we describe the working procedure of the sensitive text detector component.

C5.1. Identifier. We next explain how we analyze unstructured text (descriptive text and extracted text from X-Ray images) and structured text (extracted text from medical document images) to detect *identifiers* including name, email, phone number, and UID.

Type	Rules
Name	1,000,000 US names [45] ⋃ 1,000,000 Indian names [45]
Email	[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+[a-zA-Z0-9-.]+

Phn No.	/\(([0-9]3)?|[0-9]3-)[0-9]3-[0-9]4;
	/\(?:(?:+|0{0,2})91(\s*[-]\s*)?|[0]?)?[789]\d{9}

UID	[2-9]1[0-9]3\s*[0-9]4\s*[0-9]4, XXX-XX-XXXX, XXXXXXXXX, XXX XX XXXX; X=[0-9]
DOB	(?:(?<!:)(?<!:\d)[0-3]?\d(?:st|nd|rd|th)? \s+(?:of\s+)?(month)\s+(?<!:)(?<!:\d)[0-3]?\d(?:st|nd|rd|th)?)(?:,)?\s*(?:\d{4})?|[0-3]?\d[-./][0-3]?\d[-./]\d{2,4}
Age	\d\syears, \d\syrs, \d\syr, \d\sy/o, \d\sy, \d\sm, \d\sf
Sex	male, female, \d\sm, \d\sf
LOC	10,000 US cities [32] ⋃ 2,000 Indian cities [90]
MH	67 chronic conditions [89] in English

Table 2: Rule-based approach to detect sensitive information in non-structured data in English websites. Here, MH = Medical History.

C5.1.1. Name. We start with extracting names from unstructured text. In natural language processing (NLP), recognizing name entities using NER tagger has been extensively studied [43, 73]. As such, we start by using a NER tagger for this task. To evaluate the performance, we randomly sample 50 names predicted by this tool. Among them, we find nine false positives (FP). NER tagger does not perform well in our case because — (1) people don't maintain proper grammar while writing a post, and (2) it detects the name of the hospital as a person's name. For these reasons, we instead take the approach of matching the exact names in the text.

For our English websites, we build two name datasets (US and Indian names), by collecting 1,000 most popular first names and 1,000 most popular last names (for both US and India) [45]. Furthermore, based on the manual investigation, we find that if we only focus on first name or last name, it will result in false positives (e.g., Apple, Auburn as a first name, whereas Brown, Green as a last name). To address this problem, we exhaustively create each combination of full names so that every single first name is paired with every single last name to create a full name (*i.e.* 1,000,000 full names). We search each name within all of the posts and comments of Doctors Lounge, Patient Info, and Daily Rounds.

To extract Chinese names, we face extraction challenges due to the overlapped words commonly used for *Chinese surnames* and *Chinese traditional medicine names*. More specifically, we start with using a popular Chinese word segmentation tool [9] for Chinese name recognition. Because we need to perform word segmentation for Chinese words (which is not the case for English words) before applying NER tagger [63]. After testing it on 100 samples, it produces 30 false positives. Some of the false positives are invalid surnames. To eliminate invalid surnames, we filter the result with a list of 1000 Chinese family names [6]. This improves the performance to 80% precision and 88% F1-Score. As mentioned above, some *Chinese traditional medicine names* (or medical jargon) in the Chinese OHC posts are incorrectly flagged as surnames. To addres‘s this problem, we extract all the unique names (5,098) and rank their frequency of appearance in the posts. We find that most of the Chinese names only appeared once in all the posts, while some common medicine jargon (*e.g.,* 黄染 (translation-‘stained yellow'),

[2] We also try to crop out the image containing the embedded text. But that doesn't help to improve the performance of the OCR.

皮疹 (translation-'rash')) appear frequently. We manually compile a jargon list based on the high-frequency names. Then we use this jargon list to filter our results by removing non-personal names. Note that we do not have such jargon problems on English websites.

After handling unstructured text, next we describe how SenRev extracts names from *structured* text. In English websites, while analyzing the structured text, we match the extracted text with the field name. But for the Chinese website, it is not that straightforward. For 'name' in Chinese character form, we first use rules to extract all Chinese characters from images. In X-Ray images, simply filtering the results with a list of Chinese family names [6] give us 100% accuracy. Things are trickier when it comes to names in *Pinyin* [23] form. We manually translate (using Google Translator) the list of Chinese family names into Pinyin form and filter the OCR results. After manually checking 8,114 data, we find 3,869 FP which occurred due to the conflict of hospital names with person names. To avoid FP, we remove the entries with the keyword 'hosp' from this detected name set. For example, 'Nan' is a valid Chinese surname, but it is also the first character of 'Nan Ning No.1 Hospital'.

C5.1.2. Email. Email leakages represent a great risk to an individual, making them susceptible to phishing and malware attacks in particular– with each of those carrying the possibility of inflicting severe financial harm. Sometimes, email addresses also leak users' first name and last name. We capture email address leakages in unstructured data by checking the rule listed in Table 2. However, to identify email information in the structured data, we need to check the representative field in the text (for Chinese see Appendix A.3).

C5.1.3. Phone Number. Similar to an email address, a phone number is also considered an identifier of a user. All the phone numbers in the US, China, and India are either 10 or 11-digit. As such, we seek to extract 10 and 11-digit phone numbers from the unstructured text on English and Chinese websites respectively. We have illustrated the rules in Table 2 to detect phone numbers from unstructured text. For structured text, we search the field with keywords '电话' (translation-'phone') and 'phone'. Note that, we use 'phone' as the keyword because it also detects keywords such as 'telephone', and 'phone number' from the structured text.

C5.1.4 UID. We consider SSN (USA), Aadhar number (Indian national ID), and Chinese citizen ID as UID-type information. SSN contains a 9-digit number, Aadhar number contains 12 digit number with or without blank space in a group of 4, and Chinese citizen ID contains an 18-digit number (where the sequence of numbers represents a 6-digit area code, an 8-digit date of birth, 3-digit serial numbers, and a 1-digit verification code). We identify UID by matching the following rules – "XXX-XX-XXXX, XXXXXXXXX, XXX XX XXXX", where X = [0-9] (for SSN), [2-9]1[0-9]3\s*[0-9]4 (for Indian national ID) and "[1-9]\d{5}(18|19|20)\d{2}(0[1-9]|1[0-2])([0-2][1-9]|[12]0|3[01])\d{3}(\d|X)", where X represents alphabet (for Chinese website). The last digit of a UID can be either a digit (*i.e.* 0-9) or an alphabet. For structured text, we find the UID in the medical document by searching the following field name– 'ssn', 'social security', 'ss#', 'ss #', and '身份证(translation-'citizen ID')'.

C5.2. Quasi-identifier. We consider a range of quasi-identifiers including DOB, age, sex, location, and medical history. While quasi-identifiers may not directly re-identify people, they can be used to link the user records with other datasets where identifiable information is present (e.g., voting records) for re-identification [36, 42]. Also, the risk of re-identification becomes higher if multiple quasi-identifiers are leaked in the same post (compared to only leaking one data field). In the following, we explain how we extract quasi-identifiers from unstructured text (descriptive text, and extracted text from X-Ray images) and structured text (extracted text from medical document images).

C5.2.1. DOB. An attacker can use the date of birth information to launch a re-identification attack. We devise a rule set to detect the DOB from the unstructured data as stated in Table 2. For example, 'My birthdate is on 4th January, 2000', where we use our rule set to detect the date of birth. For the text extracted from the images of the medical document, we searched for the representative name.

C5.2.2. Age. We identify age in unstructured text by using the following rules: "\d\syears, \d\syrs, \d\syr, \d\sy/o, \d\sy, \d\sm, \d\sf, \d\s岁(translation-'years old')". We develop this rule set (Table 2) after investigating our data. For medical documents, we check whether the representative field contains any of the following: '年龄' (translation-'age') or 'age' to detect the entry of age.

C5.2.3. Sex. For unstructured text, leakage of sex is detected by searching for '男'(translation-'male'), '女'(translation-'female'), 'male' and 'female'. For Pinyin and English forms, we find that sometimes they include the patient's age and sex together in the X-Ray. For example, '19F' means the patient is a 19-year-old female. That's why we also analyze a digit followed by the character 'm' or 'f' to identify sex. For medical documents, we search fields sex with '性别' (translation-'sex'), 'sex', and 'gender'.

C5.2.4. Location. To detect location information, we first try to use location tagger [9], since it is widely used in NLP. However, in our case, this approach has not worked well due to excessive false positives (36% FP by manually verifying 200 random samples). Instead, we take the approach of searching for an exact match with a location dataset from the text. We build our location dataset by collecting three large city name datasets (two for English and one for Chinese), containing names of 10,000 largest US cities [32], 2,000 Indian cities [90], and 3,272 Chinese cities [4]. We search those city names in all the unstructured text. For extracting location information from structured text, we search the following field names: '住址'(translation-'address'), '医院'(translation-'hospital'), 'address', or 'hospital'.

C5.2.5. Medical history. We consider the detection of chronic diseases as medical history leakages because short-term diseases are temporary– that is, they do not last long. Hence, they are far less likely to be able to identify a specific individual. By contrast, a chronic disease persists and is likely to affect the individual for extended periods. From our manual investigation, we find that medical documents are structured and well-formatted. So, we can simply identify sensitive information leakage by searching the corresponding fields in the record form. The only exception is to identify if there is a chronic disease mentioned in the form. For both structured and unstructured data, we search the OCR result with a data set of 67 common chronic conditions [89]. The list of 67 chronic conditions represents the most common diseases. We create the Chinese version of this data set with the help of *Google Translator* to detect medical history leakages. We manually investigate the

Model	C1		C2		C3	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**
Resnet	78%	62%	57%	36%	86%	54%
Inception	90%	83%	**95%**	**95%**	86%	54%
VGG	95%	92%	93%	93%	**96%**	**90%**
Xception	**97%**	**94%**	93%	93%	85%	53%
HRGB	-	-	76%	75%	-	-
DSFD [39, 62]	-	-	-	-	92%	76%
Cascade [86]	-	-	-	-	84%	53%

Table 3: Performance evaluation of C1, C2, and C3.

translation results of these 67 listed chronic diseases (by a native Chinese speaker). The accuracy is 100%.

5 EVALUATION

In this section, we run a series of experiments to evaluate each of the components of SenRev to detect sensitive information. Then we evaluate the end-to-end performance of SenRev (Section 5.2). We run the experiments on a Desktop PC with 32 GB of RAM, 8 GB of graphics card (NVidia GTX-1070), and 3.1 GHz Intel Core i5 processor, running Ubuntu 18.04.

5.1 Evaluation of SenRev's Components

E1. Medical document detector (C1). To calculate the performance of the medical document detector, we randomly sample 250 images (80 medical documents and 170 non-medical documents) as test images. We have included the detailed performance of those four models in Table 3. We can observe that the Xception model performs best compared to the other models, with 94% F1-Score. We use this model to detect medical documents.

E2. X-Ray detector (C2). To evaluate the performance of the X-Ray detector, we randomly select 200 images (85 X-Ray and 115 non-X-Ray). We compare the performance of four different DNN models and HRGB (Section 4) on this test data. Here, the Inception model outperforms the others, with a 95% F1-Score (95% Accuracy, 96% Precision, and 94% Recall). So, we use the Inception model to build our C2 component.

E3. Face detector (C3). To select the best model for face detection, we sample 810 images (150 face images and 660 no face images) for evaluation. Our baseline approach [86] achieve 53% F1-Score (84% Accuracy). From Table 3, we can observe that Inception, Xception, and Resnet models perform almost similar to the baseline. However, both DSFD (76% F1-Score) and VGG (90% F1-Score) achieve better performance than the baseline. As the VGG model outperforms other models, we select this for detecting face images.

E4. OCR (C4). To evaluate the performance of the OCR tool, we randomly select 30 medical images which contain 2,920 English and 1,232 Chinese characters. We achieve a 26% F1-Score and can extract 620 characters (out of 2,920) successfully using the Tesseract (baseline) [16]. Using Rekognition tool [14], we extract 2,600 characters out of 2,920 English characters successfully but fail to extract any Chinese characters (0 out of 1,232). We use Baidu's OCR tool [7] to extract Chinese characters, and it extracts 3,748 characters (out of 4,152) successfully with an F1-Score of 94% (Accuracy 95%, Precision 96%, Recall 93%). We find that several English characters for which Baidu's OCR fails to detect, the Rekognition tool can detect. Therefore, we merge Rekognition and Baidu's OCR tool which help us to extract 3,808 characters (out of 4,125) successfully with an F1-Score of 97% (97% Accuracy, 96% Precision, 97% Recall). This combined approach outperforms all others listed in Table 4.

Tool	Acc	Prec	Rec	F1
Tesseract (baseline) [16]	57%	21%	34%	26%
Rekognition [14]	81%	64%	96%	77%
Baidu's OCR [7]	95%	96%	93%	94%
Rekognition+Baidu's OCR	**97%**	**96%**	**97%**	**97%**

Table 4: Performance evaluation of C4.

Type		English Website		Chinese Website	
		Acc	**F1**	**Acc**	**F1**
I	Name	98%	97%	99%	99%
	Email	100%	100%	100%	100%
	Phone Number	98%	98%	100%	100%
	UID	100%	N/A	96%	91%
Q	DOB	93%	94%	98%	98%
	Age	93%	94%	98%	98%
	Sex	100%	100%	100%	100%
	LOC	93%	93%	99%	99%
	Medical History	92%	92%	100%	100%

Table 5: Performance evaluation of C5. Note that, we don't find any UID leakages in English websites using SenRev which is consistent with our manual analysis. "I" represents Identifier, and "Q" represents Quasi-identifier.

E5. Sensitive text detector (C5). Next, we evaluate the performance of our sensitive text detector component. We randomly sample 200 data (100 from English and 100 from Chinese) each time for evaluating the performance of SenRev.

E5.1. Identifier. In the following, we evaluate SenRev in terms of detecting all the *identifiers* (except face) in descriptive text.

E5.1.1. Name. From Table 5, we can observe that SenRev achieve 97% F1-Score (98% Accuracy, 94% Precision, 100% Recall) for English OHC. Similarly, it performs well on Chinese OHC with a performance of 99% F1-Score (99% Accuracy, 99% Precision, and 99% Recall). Later, we investigate failed cases of our detector. We summarize two main reasons for that– (1) lack of coverage: one main drawback of the rule-based approach is the lack of coverage of rules. From our initial investigation, we find several common typical suffix, including– "<surname>xx", "<surname>*" and "<surname>某(translation-'someone')". But we miss a few untypical suffixes like– '<surname>...' (*e.g.,* "*<surname>... 18 years old, female.*"), for ethical reasons we remove the actual data. (2) conflict: infrequently used words or even made-up words often match with a valid surname (especially, a Chinese surname) as the first character.

E5.1.2. Email. Next, we evaluate our sensitive text detector component for correctly identifying email information. In this case, our tool achieves good performance with 100% F1-Score (100% Accuracy, 100% Precision, and 100% Recall) in both English and Chinese websites. One such email leakage example is: "*Please send responses to <user email address>*". As this public information contains an email address, this is marked as sensitive information leakage.

E5.1.3. Phone Numbers. The ground-truth evaluation shows 10 cases of phone number leakages (9 in Chinese and 1 in English) out of 200 posts. SenRev achieves 98% F1-Score (98% Accuracy, 96% Precision, 100% Recall) in English website and 100% F1-Score (100% Accuracy, 100% Precision, 100% Recall) on Chinese website. SenRev detects phone number leakages in the following text– "*contact further on <phone number>*".

E5.1.4. UID. Our tool is not able to find any UID on English websites. We verify it manually. Indeed, it is consistent with our tool. However, we find 20 UIDs on the Chinese website. These are of the same types, as our tool detects '身份证' (translation-'citizen id') field in medical document images. Our evaluation shows that SenRev achieves 91% F1-Score (96% Accuracy, 83% Precision, 100% Recall) on Chinese website (Table 5).

E5.2. Quasi-identifier. Next, we evaluate SenRev in detecting all the five *quasi-identifiers*.

E5.2.1. DOB. We find that SenRev achieve 94% F1-Score (Accuracy 93%, Precision 96%, Recall 92%) in detecting DOB in English websites (Table 5). The performance of SenRev in Chinese websites is also consistent with that of English websites. The ground-truth evaluation shows that SenRev achieves 98% F1-Score (Accuracy 98%, Precision 100%, Recall 96%) in detecting DOB in Chinese websites. An example of DOB leakage is– "*the last two borns were twins born on <month date year>.*". This post is marked as leakage given a date of birth is listed.

E5.2.2. Age. We investigate the performance of SenRev in detecting age information. We find 83 cases of age leakages (49 in Chinese and 34 in English) out of 200 posts. In summary, SenRev achieve 94% F1-Score (Accuracy 93%, Precision 96%, Recall 92%) and 98% F1-Score (Accuracy 98%, Precision 100%, Recall 96%) in English and Chinese websites, respectively (Table 5). Examples of manually verified age leaks includes– "*<age> elderly <sex> came in...*". We find that our tool fails to detect where people mention the age by providing contextual information along with the month name (*e.g.*, "*5月 (translation-'5 months')*"). This is the only exception where people don't use phrases like "years old".

E5.2.3. Sex. Our tool achieves 100% F1-Score (100% Accuracy, 100% Precision, and 100% Recall) on detecting sex information in both English and Chinese websites (Table 5). One such example is– "*<age> <sex> complaining of <rest of the text>*". Evident by the example, it is common to detect more than simply sex being leaked in this type of post, which is marked as leakage given the presence of sex.

E5.2.4. Location. We find 78 location leakages (78 in Chinese and 0 in English) out of 200 posts. Overall, SenRev achieves 93% F1-Score (Accuracy 93%, Precision 91%, and Recall 95%) in English websites (Table 5). It performs slightly better on Chinese website (Table 5) with an F1-Score of 99% (Accuracy 99%, Precision 100%, and Recall 98%). We examine that our tool misses those cases which don't have any explicit city name. For example, "*The Third Xiangya Hospital of Central South University*" actually reveals a city-level location, as there is only one Central South University in China, which is located in Changsha. Although we search the text with all the areas in China, we will likely ignore cases that do not directly contain a location name but can be inferred from the context.

Type		FP Analysis		FN Analysis	
		TP	**FP**	**TN**	**FN**
I	Name	88	12	197	3
	Email	95	5	200	0
	Phone Number	93	7	200	0
	UID	20	4	200	0
	Face	93	7	191	9
Q	DOB	89	11	199	1
	Age	99	1	195	5
	Sex	100	0	194	6
	LOC	93	7	199	1
	Medical History	96	4	198	2
	Total	**866** **(93.72%)**	**58** **(6.28%)**	**1,973** **(98.65%)**	**27** **(2.55%)**

Table 6: End-to-end evaluation of SenRev. "I" represents Identifier, and "Q" represents Quasi-identifier.

E5.2.5. Medical History. We find 29 cases of medical history leakages (2 in English and 27 in Chinese) out of 200 posts. SenRev achieves 92% F1-Score (92% Accuracy, 88% Precision, and 96% Recall) in English website (Table 5). However, we get even better performance on the Chinese website, with 100% F1-Score (100% Accuracy, 100% Precision, and 100% Recall). One such leakage example is: "*<age> <sex> with history of <condition>*". This post is marked as leakages given the presence of chronic illnesses used to characterize the patient.

5.2 End-to-End Validation of SenRev

Next, we investigate the end-to-end performance of SenRev. We perform two separate experiments – FP (False Positive) analysis and FN (False Negative) analysis.

FP Analysis. We randomly select 100 leakages (50 text descriptions and 50 medical images) from each category detected by our tool (except for the UID, as we only have 24 total UID leakages from medical images). In Table 6, we show the performance of SenRev. SenRev achieves 100% TP (True Positive) for sex, and the lowest performance is in detecting UID (with 83.33% TP). SenRev achieves 16.67% FP in detecting the UID. One failure case which our tool mark as leakage in the medical document is: "*the patient need to bring a copy of her* ***citizen id*** *to retrieve this diagnose report and lab results.*". Here, the healthcare professional took note where she recommend bringing citizen id, not the actual value of citizen ID. On average, SenRev achieves 93.72% TP, combining all the ten different types of sensitive data leakages.

FN Analysis. We may miss some rules which limit our tool's ability in finding rare leakages. To evaluate SenRev in failing to identify that information, we randomly select 50 posts from each of the websites which don't contain any sensitive information (according to our tool). In total, we examine 2,000 posts (Table 6), 200 for each type of sensitive information. Then, we measure the FN and TN. In our context, it is very important not to miss any leakages. That's why in this analysis, we consider 1,000 more posts for FP analysis. For name, we investigate 200 posts and find 197 posts not containing any name information. Our tool fails to detect 3 (out of 200) names. One such failure case is: "*I went with Dr <name> at <LOC>.*". Here, we have an FN where SenRev detects the name of the healthcare professional as name leakage. Overall, we find that SenRev have a very low FN (around 2.55%).

5.3 Measurement Results

We use SenRev to run an analysis on our full dataset. In total, we identify a total of 1,324,064 (69.88%) pieces of leaked sensitive information. From Table 7 and Table 8, we can observe that 98.78% (1,307,917 out of 1,324,064) leakages happened through text descriptions. The rest of the 16,147 leakages (1.2%) happen through the attached medical images (*i.e.* medical document, X-Ray, and faces).

Comparing different OHC sites, data leakages through medical images are higher in IIYI than others (79.66%, 12,863 leakages). 4,566 name (35.5%) information gets frequently leaked in medical images, whereas the less leaked information is a UID (24 out of 12,863 total leakages). In DailyRounds, 319 (out of 1,352) name leakages happen through medical images. For example, the highest number of face images (25.25%, 50 out of 198 face image leaks) gets leaked in the "Pediatric Rounds" category of DailyRounds. Medical image leakages in DoctorsLounge and DailyRounds are low compared to PatientInfo. However, compared with PatientInfo, we observe higher leakages via medical images in IIYI (12,863, covering 79.68% of the total leakages via medical images). We consider PatientInfo as baseline because of its type (A-A type websites, see Table 9).

Findings #1: In general, information leaked through descriptive text is more sensitive than that of medical images.

We identify 23,587 (1.78%) identifier leakages and 1,300,477 (98.22%) quasi-identifier leakages (1,324,064 leakages in total). Out of the 23,587 identifiers, 16,437 (69.69%) are leaked from descriptive text, and 7,150 (30.31%) are leaked through medical images. For quasi-identifier leakages, 1,291,480 (99.3%) happen through descriptive text and 8,997 (0.7%) happen in medical images. For medical images, the number of leaked identifiers and quasi-identifier is similar. For text description, there are 78 times more quasi-identifier leakages compared to the number of identifier leakages.

Findings #2: Information leakage through medical images occur frequently in healthcare professionals.

In Table 8, we can observe different types of leakages through medical images, and from Table 9, we can observe the total user distribution in different websites. Now, we normalize the leakage by the number of users − *normalized = #image_leakage / #total_users*. As a result, the normalized ratios for healthcare professionals become 1.08 (1,314/1217) and 2.59 (2,821/1,089) for DailyRounds and IIYI, respectively. The normalized ratios for anyone become 3.45 (38/11) and 0.57 (10,042/17,468) for those two websites, respectively. Note that, we are not considering DailyLounge (the normalized ratio becomes 0 for both healthcare professionals and anyone) and PatientInfo (there is no user ranking available) website for this analysis. As there is a big gap between the number of healthcare professionals and anyone in DailyRounds, we are only considering IIYI for observing the frequency of information leakages via medical images from different users. And we can notice that the normalized ratio of healthcare professionals is much higher than that of anyone.

Findings #3: Among the identifier, people pay more attention to obfuscating names, UID, and phone numbers, while putting less effort into obfuscating face images. Whereas, they tend to obfuscate DOB the most among all the quasi-identifier.

We check three types of medical images for investigating the obfuscation effort done by the user manually. Given that obfuscations are not applicable to the text of the post, We start our analysis with structured data extracted from the images of medical documents. We identify obfuscation information by checking the corresponding form field. For example, we extract the form field "name" and the contents followed by this. If there is no actual content that's recognizable after the name field, it indicates an obfuscation effort for the name. We randomly sample 100 name leakages detected by SenRev in structured data to check for obfuscation. Out of these 100 samples, 62 are obfuscated. Similarly, we calculate the number of obfuscation for age, DOB, sex, phone number, LOC, and UID as illustrated in Table 10. Interestingly, we find only one case where people forget to remove UID before posting. In summary, people are more likely to obfuscate their name, phone number, UID, and DOB in medical documents.

We then analyze the obfuscation of unstructured text data in X-Ray images. We randomly selected 100 X-Ray images from our full dataset using SenRev to identify the obfuscation in extracted unstructured text data. Out of these 100 samples, there are 60 reduced images (cut or capture in a way either accidentally or intentionally that there is no text on those images), 24 images contain at least one type of sensitive information, and 16 images show obvious obfuscation effort (*e.g.,* blackout name field). We also identify the type of sensitive information leakage among those 100 images as illustrated in Table 11. In summary, we find evidence of obfuscating the name, sex, age, and location in X-ray images.

We finally examine the obfuscation effort in face images. We randomly sample 100 face images from our full dataset. We find that people put no effort into obfuscating 79 face images. For the other 21 images, people put a white/black bar only on the eyes of the full face. The effort is insufficient to truly anonymize the face image.

Findings #4: A significant portion of the information leakage happens through comments.

In the comments, people may ask for more sensitive information. Sensitive information that is not presented in the post may be leaked through the comments. We have illustrated the detailed result in Table 12. In total, 34.82% of leakages happen through the comments.

Findings #5: The leakage ratio is low in H-H website (DailyRounds) and high in A-A websites (PatientInfo).

From Table 7 & Table 8, we can find that data leakage through text description and medical images are more common in PatientInfo and IIYI website, respectively. In terms of overall leakage, the PatientInfo website has the highest number of data leakage. The second most leakages happened in IIYI, with 82,609 leakages. Using SenRev, we find 28,556 leakages of sensitive information in DoctorsLounge. And DailyRounds have the lowest number of data leakage with a total of 5,068 instances of data leakage. The total number of data also varies from one website to another (1,500,654

	Type	DR		IIYI		DL		PI		Total
		H	A	H	A	H	A	H	A	
I	Name	0	0	538	2,559	1,170	103	-	9,106	13,476
	Email	0	0	0	0	21	218	-	2,114	2,353
	Phone	0	0	50	204	0	3	-	351	608
	UID	0	0	0	0	0	0	-	0	0
Q	DOB	12	0	6	72	8	70	-	1	169
	Age	948	34	3,564	23,929	541	7,618	-	281,526	318,160
	Sex	1,507	32	3,632	23,717	766	7,410	-	126,002	163,066
	LOC	916	0	206	2,027	248	1,202	-	230,780	235,379
	MH	256	11	1,348	7,894	4,465	4,678	-	556,054	574,706
	Total	**3,639**	**77**	**9,344**	**60,402**	**7,219**	**21,302**	-	**1,205,934**	**1,307,917**

Table 7: We applied SenRev to detect sensitive information leakages in text. MH = Medical History. "I" represents Identifier, and "Q" represents Quasi-identifier.

Plat.	H	A	Total
DR	1,217 (99.10%)	11 (0.90%)	1,228
IIYI	1,089 (5.87%)	17,468 (94.13%)	18,557
DL	1,081 (4.1%)	25,066 (95.9%)	26,147
PI	N/A	212,094 (100%)	212,094

Table 9: User distribution of different sites. For PatientInfo, we referred all the user to *Anyone* as there was no user-ranking.

	Type	#Data	#Obfuscated
I	Name	100	62
	Phone Number	75	63
	UID	20	19
Q	DOB	76	67
	Age	100	15
	Sex	100	2
	LOC	100	2

Table 10: Obfuscation ratio before posting medical documents online. We did not detect any email in medical document images. For medical history, we used a keyword search instead of field detection—if we detect a positive case, that means it was not obfuscated.

	Name(I)	Sex(Q)	Age(Q)	LOC(Q)	DOB(Q)	Total
#Contain	21	19	18	18	5	**24**

Table 11: Obfuscation ratio of 100 X-Ray images before posting online. Note that, we don't find any email, phone number, and UID in these images.

	Type	DR		IIYI		DL		PI		Total
		H	A	H	A	H	A	H	A	
I	Name	312	7	1,314	2,931	0	2	-	0	4,566
	Email	0	0	0	0	0	0	-	0	0
	Phone	8	0	27	48	0	0	-	0	83
	UID	0	0	10	14	0	0	-	0	24
	Face	197	1	52	414	0	5	-	1,808	2,477
Q	DOB	19	0	12	170	0	3	-	1	205
	Age	257	5	330	1,632	0	3	-	31	2,258
	Sex	194	7	526	1,896	0	3	-	33	2,659
	LOC	257	10	490	2,808	0	3	-	13	3,581
	MH	70	8	60	129	0	16	-	11	294
	Total	**1,314**	**38**	**2,821**	**10,042**	**0**	**35**	-	**1,897**	**16,147**

Table 8: Sensitive information leakage in medical images. MH = Medical History. "I" represents Identifier, and "Q" represents Quasi-identifier.

		Post	Cmnt.	H	A	Total
DR	Text	2,296 (61.8%)	1,420 (38.2%)	3,661 (98.5%)	55 (1.5%)	3,716
	Image	376 (31.5%)	816 (68.5%)	1,176 (98.6%)	16 (1.4%)	1,192
IIYI	Text	22,169 (72.7%)	8,342 (27.3%)	4,037 (13.2%)	26,474 (86.8%)	30,511
	Image	3,240 (100%)	0 (0%)	468 (14.5%)	2,772 (85.5%)	3,240
DL	Text	17,793 (65.3%)	9,464 (34.7%)	316 (1.2%)	26,941 (98.8%)	27,257
	Image	34 (97.1%)	1 (2.9%)	0 (0%)	35 (100%)	35
PI	Text	286,530 (23.8%)	919,404 (76.2%)	N/A	1,205,934 (100%)	1,205,934
	Image	1,324 (69.8%)	573 (30.2%)	N/A	1,897 (100%)	1,897

Table 12: Different ways of sensitive information leakage across four different websites. Note that, for PatientInfo, there was no user-ranking system available.

in PatientInfo whereas 28,710 in DailyRounds). So, we normalize the leakage ratio by following: *normalize = #leakage / #total_data * 100.* The normalization ratio of DailyRounds, IIYI, DoctorsLounge, and PatientInfo are 17.65, 28.3, 38.77, and 80.4, respectively. Even though IIYI has more leakages than DoctorsLounge, comparing the normalization ratio, DoctorsLounge has more leakages than IIYI. Both of these are A-H websites. Based on the normalization ratio, we can conclude that the leakage rate is high in A-A website, medium in A-H websites and low in H-H website. We find the same magnitude of quasi-identifier leakages among those websites as well. It is worth noting that, users posting their information is self-disclosure and they are aware of it. On the other hand, if healthcare professionals don't take consent from the patient, then they (patients) won't be able to know about such leakages. We analyze the data to find any evidence of healthcare professionals taking consent from the patient before disclosing their data online. But we don't find any such evidence on those websites.

> **Findings #6:** Information leaked by healthcare professionals is more sensitive.

While comparing the leakages from a healthcare professional and anyone, we cannot consider PatientInfo and DailyRounds because– (1) PatientInfo has no user-ranking system, and (2) the number of anyone users (in DailyRounds) is very small compared to healthcare professionals (99% users are healthcare professional). We illustrate the total user-distribution in Table 9. For the other two websites, we compare the sensitive data leakages from healthcare professionals and anyone. We list sensitive data leakages from healthcare professional and anyone across different categories through text in Table 7 and medical images in Table 8. We calculate the average number of data leakages by healthcare professionals in the following way: *avg = #(sensitive data leakages by a healthcare professional) / #(total healthcare professional).* In IIYI and DoctorsLounge, the top four most common types of sensitive data leakages are– age, sex, location and medical history, which are very sensitive [1]. Again, we normalize the ratio of identifier leakages among these four websites. We notice that the identifier leakage rate is high on H-H website, medium on A-H websites and low on H-H website. Based on our observations, we summarize a few reasons why a single healthcare professional's post leaks a lot of sensitive information on these websites. First, healthcare professionals have access to the medical documents and history of patients. Usually, they share information in detail compared to anyone. In addition, as healthcare professionals are domain experts, they tend to ask for more medical documents to get a better sense of the patient's medical condition. From our analysis (Appendix A.4), we can observe that the top active 4 (out of 5) users in both IIYI and DailyRounds (who leak a lot of information) are health care professionals, while all the top 5 users in DoctorsLounge are healthcare professionals. As PatientInfo has no user-ranking all the top 5 leakages happen from anyone. Failure to ensure the privacy of those data may bring a threat to the patient.

Findings #7: Both healthcare professionals and other users on the Chinese website tend to have more combined information leakages (i.e., leaking multiple types of information together) than English websites.

We use SenRev to search for identifier and quasi-identifier leakages by *healthcare professionals*. In Table 13, we list the top five most common information leakages for each site by considering both *single leakages* and *combined leakages*. The combined leakages refer to multiple types of information being leaked within the same post. Note that, PatientInfo does not have healthcare professionals, and thus it is not listed in this table. From Table 13, we observe that the Chinese site (IIYI) has more combined leakages compared to English sites (DoctorsLounge and DailyRounds).

Next, we perform a similar analysis for finding identifier and quasi-identifier leakages from *non-healthcare professionals* (i.e., anyone). We listed the top five combined leakages in Table 14. First, on the Chinese website IIYI, *non-healthcare professionals* also have many combined leakages just like *healthcare professionals*. Interestingly, the type of leakages from both healthcare professionals and non-healthcare professionals are almost similar. For example, in IIYI, a healthcare professional asks for treatment advice for the patient by posting “患者为<age><sex>，因<symptom>入院” (Translation - The patient is a <age><sex>, and was hospitalized due to <symptom>). We find same type of query from the non-healthcare professionals in IIYI website. Second, single leakages are more common on the three English sites. For PatientInfo, we can observe some combined leakages (especially combined leakages with location information).

In Appendix A.7, we have further discussed the nature of leakages by only considering the combined leakages.

Findings #8: Excessive sharing of unnecessary information.

Our investigation has uncovered proof that both healthcare professionals and non-healthcare professionals share unnecessary sensitive information, resulting in a trade-off between privacy and utility that is less than optimal. We mark name, email address, phone number, UID, and location as unnecessary information as they do not contribute to the identification of the disease. Nevertheless, we have observed that 19.6% (260,070 out of 1,324,064) unnecessary leakages are widespread across all four websites.

DR		IIYI		DL	
Comb.	#Leak	Comb.	#Leak	Comb.	#Leak
l	653	a+s	1,809	m	2,408
a	139	a+s+m	622	n	1,052
s+l	137	n+a+s	317	s	512
s	127	s	230	a	403
a+l	103	a	201	l	367

Table 13: Top five most common information leakages for each site by considering both *single leakages* and *combined leakages* by *healthcare professionals*. Here, a=Age, s=Sex, l=Location, m=Medical History, n=Name. Note that, PI does not have healthcare professionals.

DR		IIYI		DL		PI	
Comb.	#Leak	Comb.	#Leak	Comb.	#Leak	Comb.	#Leak
s	34	a+s	12,602	s	1,450	l	58,868
a	27	a+s+m	3,683	a	4,602	s+l	14,052
m	11	a	2,215	m	4,043	a+l	8,896
a+m	8	s	2,056	a+s	1,361	a+s+l	3,715
l	5	n+a+s	1,199	n	1,112	l+m	2,704

Table 14: Top five most common information leakages for each site by considering both *single leakages* and *combined leakages* by *Anyone*. Here, a=Age, s=Sex, l=Location, m=Medical History, n=Name.

Findings #9: The leakage rates of medical history, age, and location are high.

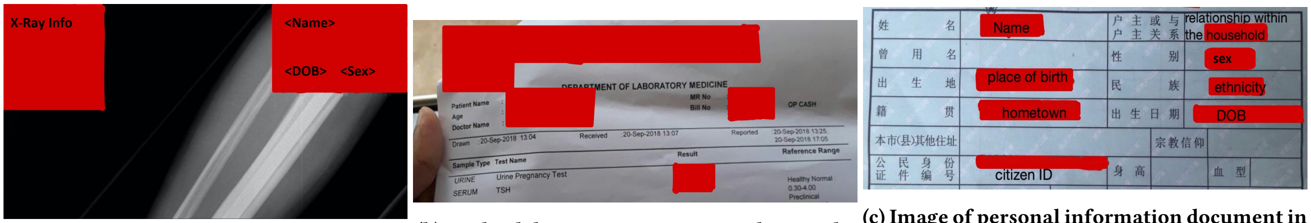
From Table 7 and 8, we find that the top three types of sensitive information leakages—medical history (575,000), age (320,418), and location (238,960)—are all quasi-identifiers. As mentioned in the threat model, quasi-identifiers can be used to link the patient's medical condition with external datasets (e.g., voter information) to launch re-identification attacks.

5.4 Case Study

SenRev identifies 1,324,064 sensitive information leakages from four different websites. In the following, we show a few interesting examples of sensitive data leakages.

DoctorsLounge. One interesting case detected on DoctorsLounge involves a patient seeking a review of his X-Ray after suffering from broken bones. The patient uploaded the picture of his X-Ray image (Figure 2a, we blur out the sensitive information) to the forum. This document contains significant amounts of sensitive information (including– name, DOB, and sex) that are completely visible.

DailyRounds. We find many posts on DailyRounds where healthcare professionals share their patient's information along with medical images. For example, a healthcare professional said, “*investigations are below. does she require any treatment?*” and attaching an image of the patient's laboratory results (Figure 2b). These results



(a) X-Ray image on DoctorsLounge.

(b) Medical document image on DailyRounds.

(c) Image of personal information document in IIYI.

Figure 2: Sensitive information leakage via different medical images in English and Chinese websites. Note that, we have obfuscated the sensitive information by adding red bar in the image before reporting in the paper.

feature the patient's name, age, sex, findings, and location of the laboratory.

IIYI. An interesting case detected on IIYI involves a healthcare professional asking for treatment suggestions for a girl who is admitted to the hospital (Figure 2c). The post reveals her name, age, location, family medical history, symptoms, and phone number. We have added additional case study in Appendix A.8.

6 COUNTERMEASURES

Based on our analysis, we discuss several countermeasures to reduce the sensitive data leakages in OHC.

First, from our measurement analysis (Findings #1, #2 and #4), we found evidence of many leakages via medical images and text. Therefore, OHC platforms should consider deploying automated systems to detect sensitive information leakage in the uploaded images and text. As shown in Appendix A.6, SenRev is highly efficient and could be deployed to perform detection in real time. OHC platforms could use/customize SenRev to detect information leakages and pop-up warning messages to users before they post the image/text to the OHC. So the deployed models need to be trained in a way that achieves a good performance in automatically detecting that information successfully.

Second, according to Findings#3 and #8, people put more effort into obfuscating certain sensitive information, while other information gets exposed. To prevent those, OHC platforms may provide built-in tools to help their users automatically redact information in images before uploading. Such tools could be helpful to lay users who are not able to redact images themselves.

Third, our investigation indicates that information leakages by a healthcare professional are more sensitive (Findings #6). OHC may consider making these attachments visible only to verified healthcare professionals. Because only they are able to interpret such enriched medical information. OHC should restrict displaying those to non-healthcare professionals.

Fourth, OHCs may learn from the successful experience of other social media platforms (*e.g.,* Reddit) to form a moderation team by recruiting members from their communities. For example, a group of active members (*e.g.,* verified healthcare professionals) can be recruited as moderators to confirm unnecessary information leakages and remind the posters to redact such information. Such a moderation team, with the help of automated detection tools, could potentially scale well for a large OHC. In Findings #5, we can observe that the leakage ratio is high in A-A websites. So keeping humans in the loop will surely help in this case.

7 ETHICS

We are aware of the privacy implications of our study and have taken active steps to ensure research ethics. First, we worked closely with our IRB to refine the research plan and received their approval. Second, our study involves collecting *public* data from online services. In general, web scraping for (non-commercial) research purposes is acceptable. This is recently backed up by court rulings and CFAA justifications [3, 11, 12, 24]. The rationale is to allow researchers to independently audit online services (e.g., in terms of algorithm fairness, and data protection effectiveness). In our study, we only collected public data to investigate the oversharing problem in online health communities, with the goal of raising awareness and improving the current practice. Third, given the sensitive nature of the data, we only focus on aggregated statistics. We do not plan to share the data with any parties. Fourth, after the analysis, we have fully anonymized our dataset by removing all personal data (e.g., 'Adam Smith' to '[Name]', 'ada@gmail.com' to '[Email]'). For the images, we only store the URL of the images. In case we need to analyze an image during the project period, we can temporally retrieve the file via the URL, perform the analysis, and delete the file from the local storage. After the project period, we also plan to *delete all the data* from our storage. Fifth, in our study, we have used cloud APIs for OCR analysis (*e.g.,* Amazon), which involves sending query data to the cloud. We want to clarify that we have opted out of having our query data stored by the Amazon server and requested to delete all data associated with our Amazon account [14]. Regarding Google Tesseract, we did not use their cloud service and only performed the analysis on our machine with a locally compiled Tesseract [16]. Regarding Baidu, their OCR APIs do not store any queried data [19], and thus have no risk. Finally, we have disclosed our research, data collection, and findings to all the websites' operators, and offered suggestions for mitigation.

8 DISCUSSION

Next, we discuss the key implications, countermeasures on avoiding leakages and limitations of SenRev.

Extra information leakage from metadata. Additional information can be extracted from the metadata of images, (*e.g.,* image properties, thumbnail, EXIF, GPS, interoperability, and makernote). Among these, GPS information reveals the location where the photo is taken. After processing all 54,677 images from IIYI using a *exifread* tool [22], we found 6,561 images containing metadata, and 2,063 (3.77%) of them contained GPS information (Latitude, LatitudeRef, Longitude, LongitudeRef, Altitude, and AltitudeRef). Such information leakage should have been prevented. The IIYI developer should have removed all the metadata before publishing the images on the website. We applied the same analysis to images of the other three websites (DailyRounds, DoctorsLounge, and PatientInfo), but we didn't find any such leakages.

Importance of consent. According to the health data privacy and protection act [1, 21], healthcare professionals should get patients' consent before sharing their data. From our analysis, we cannot tell whether there has been a consent process between healthcare professionals and patients behind the scenes. Such consent can be given in various ways (*e.g.,* verbal consent, written consent). To look for signs of consent in the posts, we manually analyzed 200 random posts from healthcare professionals and did not find any mention of such a consent process.

Privacy-Utility trade-off. OHC sites face the tension between privacy and the need for information sharing to support their core functionality which is to facilitate help/advice-seeking for patients and healthcare professionals. However, we argue that current OHC sites still have too much *unnecessary* information sharing which leads to a sub-optimal privacy-utility trade-off. More specifically, most of the *identifiers* and certain *quasi-identifiers* are unnecessary for the purpose of medical diagnosis and case discussion. These information include name, email, phone number, full face image, UID, and location. They should be properly redacted from the descriptive text or raw medical images/documents. From Table 7 and Table 8, we can observe many unnecessary sensitive information leakages across four websites. For example, we find evidence of name leakages in PatientInfo (9,106), IIYI (7,342), DoctorsLounge (1,275), and DailyRounds (319),

Certain quasi-identifiers/identifiers such as age, gender, and partial face images can be useful for the case discussion. In these cases, proper pre-processing of the information can be helpful to improve privacy (e.g., sharing age instead of the detailed year and date of birth, redacting parts of the face that are not related to the disease).

Potential impact of different privacy regulations. We consider HIPAA [1], DISHA [21], PIPL [18] privacy regulations when defining sensitive information (Section 3). These regulations are effective in the US (1996), India (2018), and China (2021). There might be a correlation between the maturity of the privacy regulation and the privacy protections of OHC sites in different countries. For example, HIPAA [1] was established in the US in 1996, which is the earliest among the three countries. During our initial search of OHC sites, we observed it is primarily US-based sites that commonly had moderators and identity verification for healthcare professionals to ensure data privacy (e.g., Doximity, MedShr). Both India and China are relatively new to implementing privacy regulations. Privacy protections for OHC users are still less common in these areas. To draw a reliable conclusion, we need to perform in-depth interviews or surveys with OHC developers/stakeholders to understand their perceptions of privacy regulations and privacy-protection mechanisms. We leave it for future work.

Identity verification of healthcare professionals. OHCs such as iCliniq and Sermo restrict the access of certain data exclusively to verified healthcare professionals. To verify identity, healthcare professionals need to submit required proofs (*e.g.,* a photocopy of medical license), which will be reviewed by moderators. While we did not attempt to study the robustness of the verification process, we accidentally discovered a bug in the OHC platform A, which would allow any users to bypass the identity verification to register as healthcare professionals. We immediately reported this issue to the corresponding OHC. They acknowledged our findings and fixed the bug.

Limitations. This work has a few limitations. First, the rule-based approach suffers from a lack of coverage. This might lead to false negatives due to insufficient rules. However, based on our analysis, we found that the false negative rate was very low, only 2.55%. So, we can claim that we have a sufficient amount of rules incorporated into our tool. Second, due to medical verification, we were not able to investigate many H-H websites (*e.g.,* Sermo). However, among the four websites that we analyzed, one of them was H-H (DailyRounds). And it represents the scenario of data leakages in H-H community. From our analysis, we also found a consistent result there. Third, we divided all the leakages caused by either a healthcare professional or anyone. From our investigation, we found that relatives sometimes share the patient's data. Extracting family information from the text description is still an open research area [28]. We leave it for future directions. Fourth, we have not tested our tools against adversarial text input. An attacker may trick our tool by providing sensitive information in a non-traditional way. We can ask the crowdsourcer periodically to identify the output of the tool to mitigate this risk.

9 CONCLUSION

In this paper, we propose a tool SenRev (using machine learning models and rule-based approaches), to systematically identify the sensitive data leakages in OHC. We applied SenRev to analyze 1,894,900 total data from four different OHC. Our tool detected 69.88% of sensitive data leakages across those four websites. In particular, our research shows there is a lack of carefulness of individuals while seeking help in those OHC. Often, they share medical images without any obfuscation. Specially, healthcare professionals need to be more aware of this fact, as they tend to share more medical images (4.7%) compared to the other types of OHC.

ACKNOWLEDGMENTS

We extend our sincere gratitude to the anonymous reviewers for their invaluable feedback and comments. Our work received support from National Science Foundation (NSF) grants, including NSF grant 2030521, NSF CNS 2114074, NSF OAC 2002985, NSF CNS 1943100, NSF OAC 1920462, Meta Research Award.

REFERENCES

[1] 1996. Health Insurance Portability and Accountability Act. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.
[2] 2018. Mitm Proxy. https://mitmproxy.org/.
[3] 2019. SANDVIG V. BARR – CHALLENGE TO CFAA PROHIBITION ON UNCOVERING RACIAL DISCRIMINATION ONLINE. https://www.aclu.org/cases/sandvig-v-barr-challenge-cfaa-prohibition\-uncovering-racial-discrimination-online.

[4] 2020. 2020 China's five-level administrative divisions (province, city, county, town, village). https://github.com/adyliu/china_area.

[5] 2020. CCTV is concerned about the online violence in Chengdu's new crown cases: Is privacy leakage a longer-term epidemic? https://finance.sina.com.cn/tech/2020-12-13/doc-iiznctke6212571.shtml.

[6] 2020. Chinese Names Corpus (Chinese-Names-Corpus). https://github.com/wainshine/Chinese-Names-Corpus.

[7] 2020. General text recognition (high-precision version). https://cloud.baidu.com/doc/OCR/s/1k3h7y3db.

[8] 2020. The Growth of Telehealth During COVID-19 and Its Future After: Dr Patricia Salber Interviews Dr Joseph Kvedar. https://www.ajmc.com/view/the-growth-of-telehealth-during-\covid-19-and-its-future-after-dr-patricia\-salber-interviews-dr.

[9] 2020. Pkuseg: a multi-domain Chinese word segmentation toolkit. https://github.com/lancopku/pkuseg-python.

[10] 2020. Vast majority of specialists increased use of telehealth tech during COVID-19 pandemic. https://www.healthcareitnews.com/news/vast-majority-specialists-increased-use-\telehealth-tech-during-covid-19-pandemic.

[11] 2020. Web scraping is legal (for UK researchers). https://aballatore.space/2020/04/01/web-scraping-is-legal/.

[12] 2020. Web scraping is now legal. https://medium.com/@tjwaterman99/web-scraping-is-now-legal-6bf0e5730a78.

[13] 2021. Alexa Ranking. https://www.alexa.com/siteinfo.

[14] 2021. Amazon Rekognition Documentation. https://docs.aws.amazon.com/rekognition/index.html.

[15] 2021. Social Security Number and Card. https://www.ssa.gov/ssnumber/.

[16] 2021. Tesseract OCR. https://github.com/tesseract-ocr/tesseract.

[17] 2021. Top Social Networking Sites for Medical Professionals. https://wsoms.org/news/techsense/top-20-social-networking-\sites-for-medical-professionals/.

[18] 2021. Translation: Personal Information Protection Law of the People's Republic of China – Effective Nov. 1, 2021. https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/.

[19] 2021. User business data and public information. https://cloud.baidu.com/doc/Agreements/s/Ek72myukw.

[20] 2022. Alexa Rank: Everything You Need to Know About It. https://kinsta.com/blog/alexa-rank/.

[21] 2022. DISHA and HIPAA, How Do They Compare? https://compliancy-group.com/disha-and-hipaa-how-do-they-compare/.

[22] 2022. ExifRead 2.3.2. https://pypi.org/project/ExifRead/.

[23] 2022. Pinyin. https://en.wikipedia.org/wiki/Pinyin.

[24] 2022. SANDVIG V. BARR — CHALLENGE TO CFAA PROHIBITION ON UNCOVERING RACIAL DISCRIMINATION ONLINE. https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2016cv1368-67.

[25] Alessandro Acquisti and Ralph Gross. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *International workshop on Privacy Enhancing Technologies*. Springer, 36–58.

[26] Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Over-exposed? Privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 357–366.

[27] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[28] Mahmoud Azab, Stephane Dadian, Vivi Nastase, Larry An, and Rada Mihalcea. 2019. Towards extracting medical family history from natural language interactions: A new dataset and baselines. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1255–1260.

[29] Randy Baden, Adam Bender, Neil Spring, Bobby Bhattacharjee, and Daniel Starin. 2009. Persona: an online social network with user-defined privacy. In *Proceedings of the ACM SIGCOMM conference on Data communication*. 135–146.

[30] Natã M Barbosa, Gang Wang, Blase Ur, and Yang Wang. 2021. Who Am I? A Design Probe Exploring Real-Time Transparency about Online and Offline User Profiling Underlying Targeted Ads. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.

[31] Anhui Dolphin Broadcast. 2020. A 29-year-old woman had a CT scan and found 2 needles on her head! Has existed for many years. https://www.myzaker.com/article/5f67729e8e9f093c9869c53e/.

[32] United States Census Bureau. 2020. Biggest US Cities By Population. https://www.biggestuscities.com/.

[33] Abdelberi Chaabane, Gergely Acs, Mohamed Ali Kaafar, et al. 2012. You are what you like! information leakage through users' interests. In *Network and Distributed System Security Symposium (NDSS)*. Citeseer.

[34] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 1251–1258.

[35] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. 2012. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in Neural Information Processing Systems* 25 (2012), 2843–2851.

[36] Yuanyuan Dang, Shanshan Guo, Xitong Guo, Doug Vogel, et al. 2020. Privacy Protection in Online Health Communities: Natural Experimental Empirical Study. *Journal of medical Internet research* 22, 5 (2020), e16246.

[37] Tobias Dehling, Fangjian Gao, Stephan Schneider, and Ali Sunyaev. 2015. Exploring the far side of mobile health: information security and privacy of mobile health apps on iOS and Android. *JMIR mHealth and uHealth* 3, 1 (2015), e8.

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 248–255.

[39] Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. In *arxiv*.

[40] Murthy V Devarakonda and Ching-Huei Tsou. 2015. Automated Problem List Generation from Electronic Medical Records in IBM Watson.. In *AAAI*. 3942–3947.

[41] Jackie Drees. 2019. Google receives more than 1 billion health questions every day. https://www.beckershospitalreview.com/healthcare-information-technology/google-\receives-more-than-1-billion-health-\questions-every-day.html.

[42] Ahmed Al Faresi, Ahmed Alazzawe, and Anis Alazzawe. 2014. Privacy leakage in health social networks. *Computational Intelligence* 30, 3 (2014), 514–534.

[43] Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Association for Computational Linguistics (ACL)*. 363–370.

[44] Michael Fire, Roy Goldschmidt, and Yuval Elovici. 2014. Online social networks: threats and solutions. *IEEE Communications Surveys & Tutorials* 16, 4 (2014), 2019–2036.

[45] Forebears. 2022. Search millions of names & places. https://forebears.io.

[46] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*. 17–32.

[47] Wei Gao, Huiling Wang, and Ning Jiang. 2022. The Impact of Data Vulnerability in Online Health Communities: An Institutional Assurance Perspective. *Frontiers in psychology* 13 (2022).

[48] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. 2012. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the ACM SIGCOMM conference on Internet Measurement*. 131–144.

[49] Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. In *Proceedings of the ACM workshop on Privacy in the Electronic Society*. 71–80.

[50] Quinn Grundy, Kellia Chiu, Fabian Held, Andrea Continella, Lisa Bero, and Ralph Holz. 2019. Data sharing practices of medicines related apps and the mobile ecosystem: traffic, content, and network analysis. *bmj* 364 (2019).

[51] Zhenyu Han, Haohao Fu, Fengli Xu, Zhen Tu, Yang Yu, Pan Hui, and Yong Li. 2021. Who Will Survive and Revive Undergoing the Epidemic: Analyses about POI Visit Behavior in Wuhan via Check-in Records. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–20.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[53] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[54] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 571–582.

[55] Shouling Ji, Qinchen Gu, Haiqin Weng, Qianjun Liu, Pan Zhou, Jing Chen, Zhao Li, Raheem Beyah, and Ting Wang. 2020. De-Health: all your online health information are belong to us. In *Proceedings of the International Conference on Data Engineering (ICDE)*. IEEE, 1609–1620.

[56] Thivya Kandappu, Archan Misra, Shih-Fen Cheng, Randy Tandriansyah, and Hoong Chuin Lau. 2018. Obfuscation at-source: Privacy in context-aware mobile crowd-sourcing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–24.

[57] Aleksandra Korolova, Rajeev Motwani, Shubha U Nabar, and Ying Xu. 2008. Link privacy in social networks. In *Proceedings of the ACM conference on Information and Knowledge Management*. 289–298.

[58] Balachander Krishnamurthy and Craig E Wills. 2008. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online Social Networks*. 37–42.

[59] Haewoon Kwak, Yoonchan Choi, Young-Ho Eom, Hawoong Jeong, and Sue Moon. 2009. Mining communities in networks: a solution for consistency and its evaluation. In *Proceedings of the ACM SIGCOMM conference on Internet Measurement*.

301–314.

[60] Ieng-Fat Lam, Kuan-Ta Chen, and Ling-Jyh Chen. 2008. Involuntary information leakage in social network services. In *International Workshop on Security*. Springer, 167–183.

[61] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[62] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: Dual Shot Face Detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR).*

[63] Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *arXiv preprint arXiv:1906.11455* (2019).

[64] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*. 1–12.

[65] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *Proceedings of the IEEE symposium on security and privacy*. IEEE, 173–187.

[66] Serguei Pakhomov, Steven J Jacobsen, Christopher G Chute, and Veronique L Roger. 2008. Agreement between patient-reported symptoms and their documentation in the medical record. *The American journal of managed care* 14, 8 (2008), 530.

[67] Chunjong Park, Hung Ngo, Libby Rose Lavitt, Vincent Karuri, Shiven Bhatt, Peter Lubell-Doughtie, Anuraj H Shankar, Leonard Ndwiga, Victor Osoti, Juliana K Wambua, et al. 2021. The design and evaluation of a mobile system for rapid diagnostic test interpretation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–26.

[68] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on Medical image Computing and Computer-assisted Intervention*. Springer, 246–253.

[69] Blaine A Price, Avelie Stuart, Gul Calikli, Ciaran Mccormick, Vikram Mehta, Luke Hutton, Arosha K Bandara, Mark Levine, and Bashar Nuseibeh. 2017. Logging you, logging me: A replicable study of privacy and sharing behaviour in groups of visual lifeloggers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–18.

[70] Jianwei Qian, Xiang-Yang Li, Chunhong Zhang, and Linlin Chen. 2016. De-anonymizing social networks and inferring private attributes using knowledge graphs. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. IEEE, 1–9.

[71] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*. 759–766.

[72] Shriti Raj, Joyce M Lee, Ashley Garrity, and Mark W Newman. 2019. Clinical data in context: towards Sensemaking tools for interpreting personal health data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–20.

[73] Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*. 147–155.

[74] Holger R Roth, Amal Farag, Le Lu, Evrim B Turkbey, and Ronald M Summers. 2015. Deep convolutional networks for pancreas segmentation in CT imaging. In *Medical Imaging 2015: Image Processing*, Vol. 9413. International Society for Optics and Photonics, 94131G.

[75] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M Summers. 2014. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. In *International conference on Medical image Computing and Computer-assisted Intervention*. Springer, 520–527.

[76] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research* 304 (2021), 114135.

[77] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the International Conference on World Wide Web*. 194–205.

[78] Faysal Hossain Shezan, Kaiming Cheng, Zhen Zhang, Yinzhi Cao, and Yuan Tian. 2020. TKPERM: Cross-platform Permission Knowledge Transfer to Detect Overprivileged Third-party Applications. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society.

[79] Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip S Yu, and Ming-Syan Chen. 2016. Mining online social data for detecting social network mental disorders. In *Proceedings of the International Conference on World Wide Web*. 275–285.

[80] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association* 26, 11 (2019), 1297–1304.

[81] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[82] Agrima Srivastava and G Geethakumari. 2013. Measuring privacy leaks in online social networks. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2095–2100.

[83] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9.

[84] Absolute Team. 2009. Medical Students Leak Patient Information on the Internet. https://www.absolute.com/blog/medical-students-\leak-patient-information-on-the-internet/.

[85] William Tsing. 2017. Please stop posting your X-rays to social media. https://blog.malwarebytes.com/cybercrime/2017/06/please-stop-posting-your-x-rays-to-social-media/. (June 09, 2017).

[86] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, I–I.

[87] Zhiyuan Wan, Lingfeng Bao, Debin Gao, Eran Toch, Xin Xia, Tamir Mendel, and David Lo. 2019. Appmod: Helping older adults manage mobile security with online social help. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–22.

[88] Leilani Wertens. 2020. Top 20 Social Networks for Doctors and Healthcare Professionals. https://upcity.com/blog/top-20-social-networks-for-doctors/.

[89] Wikipedia. 2021. Chronic condition. https://en.wikipedia.org/wiki/Chronic_condition.

[90] Wikipedia. 2021. List of towns in India by population. https://en.wikipedia.org/wiki/List_of_towns_in_India_by_population.

[91] Gilbert Wondracek, Thorsten Holz, Engin Kirda, and Christopher Kruegel. 2010. A practical attack to de-anonymize social network users. In *Proceedings of the IEEE symposium on security and privacy*. IEEE, 223–238.

[92] Wang Yuchao, Zhou Ying, and Zangyi Liao. 2021. Health privacy information self-disclosure in online health community. *Frontiers in public health* 8 (2021), 602792.

[93] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3712–3722.

[94] Yuteng Zhang, Wenpeng Lu, Weihua Ou, Guoqiang Zhang, Xu Zhang, Jinyong Cheng, and Weiyu Zhang. 2019. Chinese medical question answer selection via hybrid models based on CNN and GRU. *Multimedia Tools and Applications* (2019), 1–26.

[95] Elena Zheleva and Lise Getoor. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the international conference on World Wide Web*. 531–540.

[96] Yushan Zhu, Xing Tong, and Xi Wang. 2019. Identifying privacy leakage from user-generated content in an online health community-a deep learning approach. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–2.

A APPENDIX

Now, we discuss the detailed process of our hyperparameter selection technique, data augmentation approach, sensitive data detection on the Chinese website, user activity in OHC, category-wise leakages, the computation overhead of our tool, the combined leakages, and additional case study.

A.1 Grid search for selecting hyperparameters

We perform a grid search to select the best combination of hyperparameters for our machine learning model. To that end, we build the following sets of hyperparameters because these are normally used by the machine learning community: batch size $\mathcal{B}$ = {8, 16, 32, 64}, dropouts $\mathcal{D}$ = {0, 0.1, 0.2, 0.5}, optimizers $\mathcal{O}$ = {adam, adamax}, learning rate α = {0.1, 0.01, 0.001, 0.0001}, *epoch* = 50, activation function = softmax, loss = categorical_crossentropy. In each step of our grid search algorithm, we take a single value for setting each of the hyperparameters listed above and compute the model's performance on the validation data. After training the model with all the combinations, we select the best model based on the performance on the validation data.

A.2 Data augmentation for face detector

In the beginning, we had 613 trains (9 face images and 604 non-face images), 205 validation (3 face images and 202 non-face images), and 706 test data (6 face images and 697 non-face images) from IIYI website.

<table>
<tr><th>Type</th><th>Rules</th></tr>
<tr><td>Name</td><td>1,000 Chinese surnames [6]</td></tr>
<tr><td>Email</td><td>[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+.[a-zA-Z0-9-.]+</td></tr>
<tr><td>DOB</td><td>‘[0-9]*[\u4e00-\u9fff]*[0-9]*[\u4e00-\u9fff]*[0-9]+[\u4e00-\u9fff]*出生(translation-‘was born’)’, ‘出生于[0-9]*[\u4e00-\u9fff]*[0-9]*[\u4e00-\u9fff]*[0-9]+[\u4e00-\u9fff]*(translation-‘born on’)’ (where month indicates month name)</td></tr>
<tr><td>Age</td><td>‘\d\s岁(translation-‘years old’) ’</td></tr>
<tr><td>Phone No</td><td>\D13[0-9]\d{8}\D|\D14[5|7]\d{8}\D|\D15[0-9]\d{8}\D|\D18[0-9]\d{8}\D</td></tr>
<tr><td>Sex</td><td>‘\d\sf’, ‘\d\sm’, ‘男(translation-‘male’)’, ‘女(translation-‘female’)’</td></tr>
<tr><td>UID</td><td>[1-9]\d{5}(18|19|20)\d{2}(0[1-9]|1[0-2])([0-2][1-9]|[12]0|3[01])\d{3}(\d|X), X represents alphabet</td></tr>
<tr><td>LOC</td><td>3,272 city names in China [4]</td></tr>
<tr><td>MH</td><td>67 chronic conditions [89] in English and Chinese</td></tr>
</table>

Table 15: Rule-based approach to detect sensitive information in X-Ray images and text description in Chinese website. We have provided the English translation of Chinese word inside the parenthesis right next to the Chinese word.

Then, we trained four DNN models (similar to medical document detector and X-Ray detector) with those data and evaluated the performance of the test data. But, the best performance that we achieved was 50% F1-Score in terms of identifying images with faces. This happened due to the highly imbalanced dataset. So, we augmented positive (image with faces) data by randomly rotating the original face images. In this way, we created nine different positive images from each original positive image. After that, our train data contained 90 (=9*9+9), validation data contained 30 (=3*9+3), and test data contained 100 (=10*9+10) face images. In total, we had 694 train data, 232 validation data, and 760 test data. Even with these increased data, our best model achieved 83% F1-Score in distinguishing face images. But we observed an improvement in F1-Score. Labeling data is very expensive. So, we used this trained model to collect more face images using an iterative approach combined with a data augmentation technique. The key idea was to reduce the human effort in finding the positive data from the unlabeled dataset (containing 22,531 images). We can divide our active learning approach into the following steps:

Step 1. We used our trained model to collect face images from an unlabeled dataset. **Step 2.** We verified the predicted face images with the help of a human annotator. **Step 3.** Then, we augmented the positive data using the data augmentation approach (random rotation). We replicated nine different positive images by randomly rotating each positive image. **Step 4.** Once we had the newly labeled data, we again trained the model by adding those new data to our train corpus. **Step 5.** In each round of the iterative approach, we evaluated our model's performance. Note that, we kept the test and validation data the same for all the rounds. We decided to stop the iterative approach when the model was less effective in having true predictions. We wanted to make our face detector component be able to identify the most number of true face images. That's why we selected this criterion to stop the iterative approach. In the third round, we got 2,504 training data and our trained model achieved 90% F1-Score on the test data. This model correctly identified 86 positive images out of 196 total predicted face images. In rounds 4,5 and 6 we noticed 92%, 89%, and 89% F1-Score, respectively. And in those rounds, the model predicted 16, 2, 0 positive images correctly out of 90, 33, and 13 predicted images, respectively. It seemed that in round 3 the model worked effectively in terms of identifying positive images. From our model selection criteria, it seemed that in round 3 the model was able to achieve a high true positive rate and from round 4 the true prediction rate was degrading. On the other hand, some may argue that from round 4 the number of face images reduced in the unlabeled dataset, that's why the model was not able to get a good true positive rate. To validate the model's (from the third round) performance, we compared these models' performance on the DailyRounds image dataset. This website contained 7,097 images in total. So, we used these three models (round 3, round 4, and round 6) on these full unlabeled datasets. For those three models, we got 77, 16, and 24 correctly predicted face images out of 130, 21, and 33 face images, respectively. That's why we decided to use the model and labeled dataset that we found in round 3 for building our face detector component.

Type	Representative Form Field
Name	‘姓名’ (‘name’)
Email	‘邮箱’ (‘email’)
DOB	‘出生日期’ (‘birth date’)
Age	‘年龄’ (‘age’)
Phone	‘电话’ (‘telephone’)
Sex	‘性别’ (‘sex’)
UID	‘身份证’ (‘citizen ID’)
LOC	‘住址’ (address), ‘XX医院’ (‘hospital’), (where XX represent the name of the hospital)
MH	67 chronic conditions [89] in English and Chinese

Table 16: Rule-based approach to detect sensitive information in medical document of Chinese website. Note that, we followed the same rule-based approach for identifying Medical History in both structured and unstructured text.

A.3 Investigation on Chinese Website

We used the rules listed in Table 15 & Table 16 to identify sensitive information leakages in structured and unstructured text found in Chinese website.

A.4 User activity in OHC

We analyzed different user activities in OHC. We calculated the full distribution of posted data by healthcare professional and anyone in Table 18.

DR	IIYI	DL	PI
Usr#1 (88)	Usr#1 (661)	Usr#1 (3,841)	Usr#1 (10,808)
Usr#2 (87)	Usr#2 (463)	Usr#2 (3,283)	Usr#2 (4,345)
Usr#3 (40)	Usr#3 (395)	Usr#3 (2,729)	Usr#3 (4,258)
Usr#4 (38)	Usr#4 (348)	Usr#4 (2,359)	Usr#4 (4,009)
Usr#5 (29)	Usr#5 (217)	Usr#5 (2,063)	Usr#5 (3,549)

Table 17: Top five user leaking sensitive information across four different websites.

Later, we investigated whether certain users were more likely to post sensitive data compared to others (i.e., they were less concerned about privacy issues and posted sensitive data to get an informed answer). We listed the top five users (categorizing the leakages using username) across four websites who leaked more sensitive information in Table 17.

		H	A	Total
DR	Text	27,989 (99.13%)	247 (0.87%)	28,236
	Image	7,033 (99.1%)	64 (0.90%)	7,097
IIYI	Text	5,921 (13%)	39,252 (87%)	45,173
	Image	11,278 (21%)	43,399 (79%)	54,677
DL	Text	30,365 (41%)	43,452 (59%)	73,650
	Image	0 (0%)	48 (100%)	48
PI	Text	N/A	1,500,633	1,500,633
	Image	N/A	95,316	95,316

Table 18: Total number of text and image data posted by both healthcare professionals and anyone on four different websites.

A.5 Category wise leakages

We also investigated the post categories where the most leakages happened. In Table 20, we have reported the top five categories based on the leakage rate from each website. Among those, 'Cardiology' is the common one in all the websites, where 3,635 (out of 27,287) in DoctorsLounge, 4,009 (out of 97,316) in IIYI, 481 (out of 5,068) in DailyRounds, and 67,849 (out of 1,207,831) in PatientInfo. Usually, posts from this category had lots of medical images (*e.g.*, ECG, MRI, etc.) containing highly sensitive information.

A.6 Computation overhead of SenRev

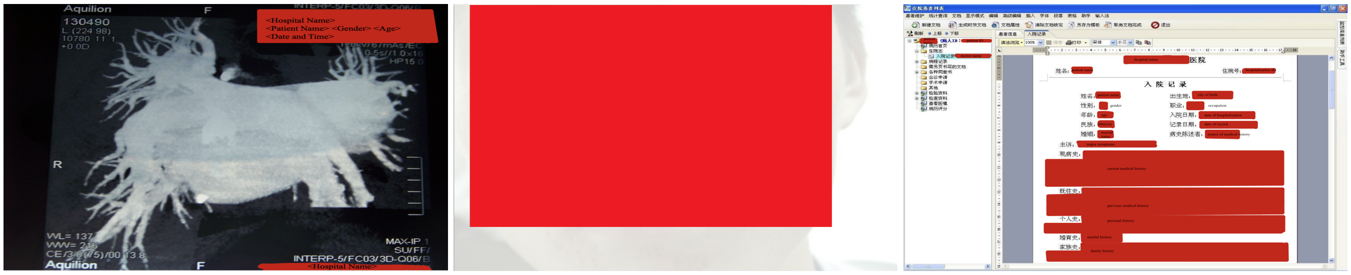
We examined the computation overhead of five different components of our tool. Particularly, we checked the feasibility of our tool in integrating with the existing OHC. We calculated the train time as the required time to train a model (for a machine learning model) or to pre-compute a keyword list. Note that, train time was a one-time cost. Similarly, we marked the test time as the time required to evaluate each data. To test on a new instance, we can simply load the pre-trained model to perform a further evaluation of the new data. We have reported the detailed results in Table 19. First, we measured the overhead time of detecting images of the medical document. We used Xception (DNN) model in building a medical document detector. We calculated the time taken for training this model which was 1,088s. Later, we loaded this pre-trained model to detect medical documents from 200 randomly sampled medical images. It took almost 338.3s (1.7s on average) to complete the full testing. Overall, it required 1,089.7s to finish evaluating a single image. Similarly, we followed the same approach for measuring the total overhead of the other two components (*i.e.* X-Ray detector and Face detector). The total overhead time of those two components was 1,107.8s and 1,785.3s, respectively. However, for OCR we had no training cost as we directly used the pre-trained models. So, we only calculated the test time for OCR. We applied our OCR component on 6,365 medical images. It took 2,546s for Baidu's OCR tool and 3,190s for the Amazon Rekognition tool to extract all the text. On average, it took 0.9s to extract text from one medical image using both tools. In the sensitive text detector component, we generated a keyword list (*e.g.* name) to find their matching in the text. Due to its large size, our 1,000,000-entry name data set required the most time (0.38s) to be formed compared to others. So, we considered this as the upper bound for calculating the train time of the sensitive text detector component. Once the training phase was done, we used the pre-computed list to identify leakages in 200 data. We found that on average, it took 1.1s to

Component	Train(s)	Test(s)	Total(s)
C1. Medical document detector	1,088	1.7	1,089.7
C2. X-Ray detector	1,106	1.8	1,107.8
C3. Face detector	1,783	2.3	1,785.3
C4. OCR	N/A	0.9	0.9
C5. Sensitive text detector	0.376	1.1	1.48
Total	**3,977.38**	**7.8**	**3,985.18**

Table 19: Computation overhead of SenRev. Note that, we used pre-trained OCR tool developed by Amazon and Baidu. That's why we didn't have any training overhead for this component.

Plat	Category	#Leakage
DR	Anaesthesia Rounds	673 (13.28%)
	Pulmonology Rounds	525 (10.36%)
	DailyRounds Primary	484 (9.55%)
	Cardiology Rounds	481 (9.49%)
	OB & GYN Rounds	477 (9.41%)
IIYI	General Medicine	4,416 (4.54%)
	Cardiology	4,009 (4.12%)
	Neurology	3,156 (3.24%)
	Pediatrics	2,752 (2.83%)
	Skin and sexual transmission	2,043 (2.10%)
DL	Cardiology	3,635 (13.32%)
	Obstetrics & Gynecology	2,783 (10.20%)
	Oncology	2,533 (9.28%)
	Gastroenterology	2,217 (8.12%)
	Neurology	2,198 (8.06%)
PI	Gut, bowel, and stomach	234,982 (19.45%)
	Mental health	175,047 (14.49%)
	Contra. and sexual health	103,037 (8.53%)
	Bones, joints and muscles	98,997 (8.20%)
	Cardiology (Heart health)	67,849 (5.62%)

Table 20: Top five categories based on the total number of information leakage. Total number of categories in DR, IIYI, DL, PI are 11, 33, 22, 29, respectively.



(a) image of X-ray. (b) face image. (c) image of medical record.

Figure 3: Healthcare professional leaking patient's sensitive information found in IIYI. We modified opacity and added boxes to ensure the obfuscation of private information in the paper.

evaluate a single piece of data. The total computation overhead for the sensitive text detector component was 1.48s. From Table 19, we can conclude that our tool is feasible for large-scale analysis with a low computation overhead (in total 3,985.2s).

A.7 Combined leakages

We use SenRev to search for identifier and quasi-identifier leakages by *healthcare professionals*. In Table 21, we list the top five most common information leakages for each site by considering *combined leakages*. Sometimes, the quasi-identifier leaks by itself whereas other times it gets leaked with other identifiers. For example, name, age, sex, and medical history leaked together 112 times in IIYI. We can notice that the leakages of age and sex are common among the three websites.

DR		IIYI		DL	
Comb.	#Leak	Comb.	#Leak	Comb.	#Leak
s+l	137	a+s	1,809	s+m	132
a+l	103	a+s+m	622	a+m	102
a+s	101	n+a+s	317	l+m	83
m+l	98	n+a+s+m	112	m+n	73
a+s+l	93	a+s+l	60	a+s	51

Table 21: Combination of sensitive information leakage per post by *healthcare professionals*. Here, a=Age, s=Sex, l=Location, m=Medical History, n=Name. Note that, PI does not have healthcare professionals.

Next, in Table 22, we list the top five combined leakages by *anyone*. We can observe many combined leakages for PatientInfo and IIYI website which is different from the other two English websites. Note that, DailyRounds doesn't have many leakages, that's why it only has three types of combined leakages.

A.8 Additional case study

Sharing patients' information by healthcare professionals is common on IIYI. We present four more examples below.

1. X-ray. As shown in Figure 3a, a healthcare professional asks for treatment advice by posting the patient's X-ray image. The post reveals the patient's name, age, and gender. The associated post

DR		IIYI		DL		PI	
Comb.	#Leak	Comb.	#Leak	Comb.	#Leak	Comb.	#Leak
a+m	8	a+s	12,602	a+s	1,310	s+l	14,052
a+l+s	4	a+s+m	3,683	s+m	663	a+l	8,896
l+s	1	n+a+s	1,199	a+m	514	a+s+l	3,715
-	-	a+s+l	649	a+s+m	306	m+l	2,704
-	-	n+a+s+m	638	d+s	159	a+m+l	2,701

Table 22: Top five combinations of sensitive information leakage per post by *Anyone*. Here, a=Age, s=Sex, l=Location, m=Medical History, n=Name, d = Date of Birth.

also reveals a detailed medical history. For example, the first line of the post is "*<name> <gender> <age>* 因 '*<symptoms>*于*<date>*入院*(translation - 'hospitalized on <date> due to <symptoms>')*". Moreover, by reading the full paragraph, we find that it also reveals the patient's family medical history, the age of her parents and siblings, and her parents' jobs.

2. Medical records. Figure 3c shows an example where the leakage happens through the image of medical records. A healthcare professional shares a patient's medical history, physical examination, laboratory examination, and imaging examination in both image and text form.

3. Face. Figure 3b is presented in a help-seeking post where a healthcare professional asks for diagnosis advice by sharing the patient's face without any obfuscation effort (Figure 3b). The poster also shares the patient's gender, age, and family history of skin disease.

4. Text. We have list an example where a healthcare professional asking for treatment advice for his patient (child). Original post (in Chinese) – "患儿, <gender>，<age>,缘于约<month>前无明显诱因（<a family-related accident>）出现<symptom>. 随后在<hospital name>, <a series of test names and results>. 既往史: <medical history> 体格检查: <results of physical exams> 诊断: <four possible diagnosis>" Translated version of the post is – "The child, <gender>, <age>, had <symptom> on <month> due to no obvious reason (<a family-related accident>). Then conducted <a series of test names and results> at <hospital name>. Medical History: <medical history> Physical examination: <results of physical exams> Diagnosis: <four possible diagnosis>". From the post, we can observe that throughout this help-seeking post, the poster reveals the patient’s age, gender, city-level information, as well as medical history, which includes the current treatment plan and associated symptoms.