Analysing Privacy Issues of Android Mobile Health and Medical Applications

Journal:	*BMJ*
Manuscript ID	BMJ-2020-063318
Article Type:	Research
BMJ Journal:	BMJ
Date Submitted by the Author:	23-Nov-2020
Complete List of Authors:	Tangari, Gioacchino; Macquarie University, Ikram, Muhammad; Macquarie University Ijaz, Kiran; Macquarie University Kaafar, Dali; Macquarie University Berkovsky, Shlomo; Macquarie University
Keywords:	mobile health, information privacy, mobile apps, trackers analytics

Analysing Privacy Issues of Android Mobile Health and Medical Applications

Gioacchino Tangari, Muhammad Ikram, Kiran Ijaz, Mohamed Ali Kaafar, Shlomo Berkovsky

Macquarie University, Australia

November 24, 2020

Abstract

OBJECTIVES: To investigate whether and what user data is collected by health-related mobile applications (mHealth apps), to characterise the privacy conduct of all the available on Google Play mHealth apps, and to gauge the associated privacy risks.

DESIGN: Automated data crawling, analysis of the static source code and apps' files/codes as well as run-time network traffic, and analysis of public app reviews.

SETTING: All health-related apps developed for the Android mobile platform, available on the Google Play app store in Australia, and belonging to the Medical and Health & Fitness app categories.

PARTICIPANTS: 20,991 mHealth apps found on the Google Play app store. Out of these, in-depth analysis of 15,838 free apps (requiring neither download nor subscription costs): 5439 Medical and 9648 Health & Fitness.

INTERVENTIONS: Laboratory-based, cross-sectional assessment of each app, including the inspection of the app resources (codes/files), and interception and analysis of app-generated network traffic. Data collection/sharing practices and privacy leaks found through pattern-matching search in the app code and automatic classification of the traffic flow. Analysis of the app public app policy and user reviews.

MAIN OUTCOME MEASURES: Identification and characterisation of the user data collected and shared by mHealth apps. Analysis of the primary recipients for each type of collected user data. Presence of ads and trackers in app traffic. Audit of the app privacy policy and compliance of app privacy conduct with the policy. Analysis of privacy perceptions and complaints.

RESULTS: 88% of mHealth apps collect/share user data. 16.8% of the detected data-collection practices, are towards the app developers (first-party), while 83.2% are towards external services providers (third-parties). A small number of third-parties (including popular services, like Google and Facebook) received 67.8% of the collected data. 23% of privacy leaks occur on insecure communications protocols. 20% of apps provide no privacy policies, while only 47% of data leaks are compliant with the practices disclosed by in the privacy policy. Less then 2% of users reviews raised privacy concerns.

CONCLUSIONS: Our large-scale analysis of mHealth apps surfaced serious privacy issues, with limited awareness of app users. It is important for clinicians to articulate these to patients, in order to be able to accurately weigh the benefits and risks of the apps.

1 Introduction

With the high growth in population having access to smartphone‘ devices, we have witnessed an explosive growth of mobile applications (in short apps) available through a variety of market places. As of 2017 there were approximately 2.7 million[1] apps available just on the Google Play store alone. Breaking down these apps by category, two of the most popular types of apps are *Medical* and *Health & Fitness*. These apps, referred to

[1]Number of Android Apps on Google Play https://www.appbrain.com/stats/number-of-android-apps.

as mobile health or *mHealth* apps, encompass a wide range of functions, from health condition management and symptom checkers to step/calorie counters and period trackers [15]. Today, mobile health is a booming market targeted at both patients and clinicians. Following the explosion of mobile health, recent guidelines [8] from the U.S. Food and Drug Administration (FDA) have formalized the use of mHealth apps for healthcare and recommended those providing aid to patients or clinicians to be considered as standard medical devices.

While the potential of mobile health to improve access to real-time monitoring and health care resources is well established [21] [11], mHealth apps may also pose serious risks to users. Many mHealth apps offer no validation measures of effectiveness from a medical standpoint [20] and a range of potential safety issues has been identified [4]. On top of these, concerns about data privacy in mHealth apps are particularly topical due to the sensitive types of information mHealth apps access, their business model centered on selling subscriptions or selling user data [14], and the enforcement of privacy standards around the world (e.g., the GDPR[2] in Europe). The limited quality and safety improvements observed over time for top mHealth apps [16], as well as their inadequate privacy disclosures [6,11], make the case for auditing this segment of mobile applications, evaluating their data collection practices and their privacy risks, and investigating the perceptions of mHealth users.

In this study, we embarked on a large-scale privacy analysis of mHealth apps. We deployed a suite of app collection and analysis tools, to perform a privacy audit of more than 20,000 mHealth apps available on the Google Play store. The scale of the analysis is orders of magnitude larger than previously reported analyses [6,9–11] and virtually covers all the Google Play store mHealth apps accessible from Australia. Our study aims to provide a comprehensive view on mHealth privacy risks by spanning the data collection practices performed by mHealth apps, the recipients of mHealth users' information, the security and transparency in user data transmission, and the users' perceptions around mHealth apps and the associated privacy conduct.

The key findings of our analysis are as follows:

- The majority of mHealth apps automatically retrieve and transmit personal user information. The acquired data includes persistent device identifiers and sensitive user information that allow tracking individuals over time and across different services, or can directly be used to profile individuals and their preferences. This way, the apps actively contribute to the creation of user profiles for advertisers.
- mHealth apps collect most of user data on behalf of external, third-party services, which generally incorporate analytics functionalities and advertisements in the apps. Furthermore, our results depict a concentration of user data transmission towards services owned by a small number of commercial entities, often owned by Google.
- While being routine in mHealth apps, data collection practices are far from being transparent and secure. Alarmingly, less than half of the detected user information transmissions comply with the disclosures made in the app privacy policies. In addition, sensitive user data is often shared on insecure channels, directly exposing users to data interception and surveillance risks.
- Despite potential privacy risks, mHealth users have a very limited awareness of the actual conduct of mHealth apps, as revealed by the analysis of public app reviews, where user's complaints on privacy appear only for a tiny portion of apps.

Collectively, these findings paint a worrisome picture, where numerous privacy and security breaches exist, whereas their awareness is minimal. Hence, it is important to bring our findings to the attention of clinicians, in order articulate the privacy risks to patients and be able to diligently weigh the benefits and risks of mHealth apps.

2 Methods

We begin by describing our dataset, data collection method, and the analysis methodology. First, we discover and collect at scale mHealth apps from the Google Play store. Next, we discuss the code (*static*) and run-time

[2]EU General Data Protection Regulation (GDPR), 2018. https://gdpr-info.eu/

behaviour (*dynamic*) analysis methods, deployed to characterise the collected apps.

2.1 mHealth App Dataset

Google Play neither provides a complete list of mHealth apps nor its search functionality yields all the available apps. To overcome this and detect as many mHealth apps as possible, we developed a crawler[3] that interacted directly with the app store's interface. Starting from the top-100 Medical and Health & Fitness apps on Google Play, the crawler systematically searched through other apps considered 'similar' by Google Play. For each app, the crawler collected the following metadata: app category and price, locations where the app is available, app description, number of installs, developer information, user reviews and app rating. In total, the crawler searched through more than 1.7 million apps over a 6-week period in 2019, from October, 1 to November, 15.

We then selected apps belonging to the Medical and Health & fitness categories on Google Play. Overall, we discovered 20,991 mHealth apps, of which 15,893 (75.7%) are free, 3,228 (15.4%) are paid, and 1,872 (8.9%) are geoblocked (cannot be downloaded from Australia). In addition, we used the crawler to sample a random set of popular non-mHealth apps to be used as comparative baseline. This set contains 8,468 apps from the Tools, Communication, Personality, and Productivity categories. Table 1 summarises our dataset.

2.2 Analysis Methodology

Figure 1 depicts our analysis methodology, which consists of three components: we analyse mHealth app files and source code; we investigate the network traffic generated during the app execution; and, we analyse users' perceptions of the apps expressed in the app reviews.

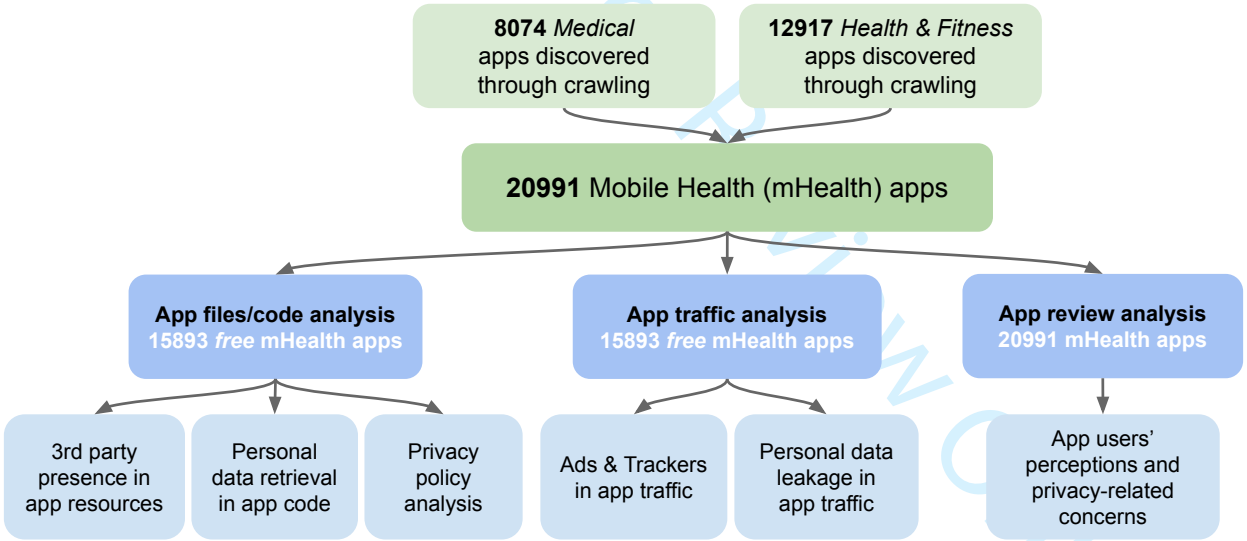


Figure 1. Overview of the mHealth app privacy analysis

[3]https://en.wikipedia.org/wiki/Web_crawler

Table 1. Summary of the 20,991 mHealth apps (broken down into Free, Paid, and Geoblocked) and 8,468 baseline (non-mHealth) apps, collected from the Google Play store.

Characteristics	No. (%)
mHealth Category	20,991 (100%)
Medical	8,074 (38.46%)
Health & Fitness	12,917 (61.54%)
Fee Required to Download	
Yes (Paid mHealth Apps)	3,228 (15.38%)
No (Free mHealth Apps)	15,893 (75.71%)
No (Geoblocked mHealth Apps)	1,872 (8.92%)
# of Downloads	
500+	7,481 (35.87%)
1,000+	4,009 (19.22%)
5,000+	1,683 (8.07%)
10,000+	3,582 (17.18%)
50,000+	1,253 (6.01%)
100,000+	1,882 (9.02%)
500,000+	375 (1.81%)
1,000,000+	462 (2.22%)
5,000,000+	127 (0.61%)
Avg. Rating	
0.0 – 1.0	6,146 (29.27%)
1.0 – 2.0	240 (1.14%)
2.0 – 3.0	1,350 (6.43%)
3.0 – 4.0	4,856 (23.14%)
4.0 – 5.0	8,396 (40.01%)
Contains Ads and Includes Tracking and Analytics Services	
mHealth Category (yes/no)	13,166 (62.72%) / 7,825 (37.28%)
Medical (yes/no)	4,516 (53.93%) / 3,558 (44.07%)
Health & Fitness (yes/no)	8,547 (66.17%) / 4,370 (33.83%)
Includes Privacy Privacy Link on Google Play's Webpage	
mHealth Category (yes/no)	15,088 (71.88%) / 5,902 (28.12%)
Medical (yes/no)	5,439 (67.36%) / 2,635 (32.64%)
Health & Fitness (yes/no)	9,648 (74.69%) / 3,269 (25.31%)
Users Perception Determined by $100 \times \frac{\#\ of\ Positive\ Reviews}{\#\ of\ All\ Reviews}$	
0.0% – 20%	3,428 (49.4%)
20% – 40%	1,374 (19.8%)
40% – 60%	880 (12.6%)
60% – 80%	487 (7.01%)
80% – 100%	769 (11.08%)

App files/code analysis: We analysed the app resources (app files and source code), to understand what personal user information the apps can potentially retrieve and quantify the presence of external services (such as analytics, advertisement and social networks, also called *3rd parties*) that may receive user information from the apps. Out of the initial set of 20,991 apps, we downloaded all the 15,893 free apps, while paid and geoblocked apps were excluded. To access the app resources, we processed the downloaded app packages using `apktool`[4], a tool for reverse engineering Android apps that allows decoding the compiled apps to their nearly original form. In addition, for all the 20,991 mHealth apps, we extracted the app's publicly-available *privacy policy*, disclosing the collection and use of personal data and describing the app's privacy practices.

[4]A tool for reverse engineering Android apps. https://ibotpeaches.github.io/Apktool/.

Typically, the link to the privacy policy is included in the app page on Google Play. If the link is broken or points to a page with no text, we tag the app as having no privacy policy.

Below, we briefly explain how we analyse the extracted app resources:

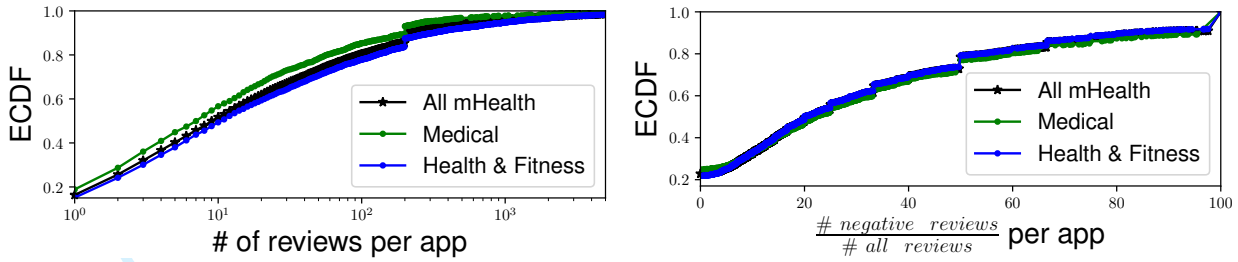
- *Third-party presence in app resources:* Given the folder containing the decoded app files and embedded libraries, we perform a dictionary-based search to retrieve and classify all third-party libraries included in the app. To this end, we employ a comprehensive dictionary of third-party libraries from [12], which comprises 338 third-parties, including ads (e.g., `GoogleAds`), analytics (e.g., `GoogleAnalytics`), utilities (e.g., `Github`) and other social, banking and gaming services (e.g., `Facebook` or `PayPal`).
- *Access to personal data in the app code:* We detect those operations in the app code involving user's personal data collection. To identify these, we use two resources. The *first* is the set of the Anroid operating system functions associated with access to personal data. For example, the occurrence of function `android.telephony.TelephonyManager.getLine1Number` in the app indicates the retrieval of the user's contact phone number. Given the name of the file where the function occurs, we can determine whether the personal data is accessed by the app's first-party (app developer) or collected on behalf of a third-party service. The *second* resource is the set of permissions requested by the app to access operating system components such as contact list or GPS location. Using the requested permissions, we check if each data-retrieval function in the app code has all the required authorisations for execution.
- *Privacy policy analysis:* Manually reviewing and annotating the app privacy policies is unfeasible due to the scale of the dataset. To overcome this, we perform an automatic privacy policy analysis using the state-of-the-art approach [23], which employs machine-learning to predict the disclosure of personal data in the privacy policy text. We train the machine-learning using a large public dataset of annotated privacy policies, APP-350[5]. The validation accuracy of the automated privacy policy analysis is presented in Appendix B.

Traffic analysis: We intercepted and analysed all the network traffic generated by the apps during the execution of automated app testing [13]. By doing so, we can assess the data transmitted by mHealth apps at *run-time*, as well as the recipients of this data. For the testing, we build a dedicated testbed composed of a smartphone that connects to the Internet via a computer configured as a WiFi access point, which runs a tool [6] intercepting all the traffic transmitted to the Internet. We individually test each of the 15,893 downloaded free apps [7]: for each app, we execute on average 35 different activities (e.g., open the app, open menu, click on button, etc.) in a 180-second test session. To reduce traffic contamination, we minimize as much as possible all the background processes of the smartphone (e.g., notifications of other apps).

In the following, we briefly describe how we analyse the intercepted app traffic:

- *Ads and trackers in app traffic:* In addition to the detection of third-party libraries performed in the app files/code analysis, we inspect the intercepted traffic to quantify interaction with external advertisement and tracking services – most likely third-party recipients of personal data [18]. Specifically, we match the resources requested in the app traffic against two comprehensive lists: *EasyList* [8]–an advertisement block list and *EasyPrivacy* [9]–a supplementary block list for tracking. This allows to isolate the traffic associated with ads and trackers.
- *Personal information leakage in app traffic:* By analysing the app files and code, we could identify the potential data retrieval/sharing practices of the apps. We complement this by extracting the personal information that actually leaks in the app traffic at *run-time*. To automatically detect

[5]Set of 350 privacy policies of popular mobile apps annotated by legal experts (https://usableprivacy.org/data).
[6]Mitmproxy - an interactive HTTPS proxy. https://mitmproxy.org.
[7]Paid and geoblocked apps were excluded again.
[8]https://easylist.to/easylist/easylist.txt
[9]https://easylist.to/easylist/easyprivacy.txt



(a) ECDF of # of reviews per app

(b) ECDF of fraction of negative reviews per app

Figure 2. Distribution of users reviews (overall and negatives) per mHealth app.

the information leaks, we deploy a machine-learning method [19] to find in the app traffic personally identifiable information, considered to be either the (*i*) specific device identifier (e.g., Android ID), (*ii*) user identifier (e.g., name or e-mail), (*iii*) credentials (e.g., password), or (*iv*) location. The machine learning was trained on a large public dataset of annotated mobile-app traffic flows[10].

App review analysis: The analysis of mHealth app reviews allows to shed light on the users' perceptions of the apps and their privacy-related concerns. We obtain the complete list of reviews for each app by downloading the content of the app's page on Google Play store. Upon excluding those reviews with no text, we obtain a dataset of 2,130,684 reviews for 6,938 mHealth apps, of which 1,764,486 (82.81%) refer to Health & Fitness apps and 366,198 (17.19%) to Medical apps. We categorize these reviews as *positive* (4 or 5 stars), *negative* (1 or 2 stars), or *neutral* (3 stars), obtaining 1,788,463 (83.94%) positive and 235,210 (11.04%) negative reviews.

Figure 2 summarises the number of reviews per app, as well as the fraction of negative reviews. for the different categories of mHealth applications. Figure 2a reports the cumulative distribution of the number of reviews per app. We observe that the number of reviews per app ranges from 0 to 5000, with 96% of apps receiving at most 1000 reviews. Figure 2b shows, for all app categories, a predominance of positive reviews, as the median fraction of negative reviews is 23%. However, a non-negligible portion of mHealth apps received mainly negative reviews. In Section 3.4, we describe the main complaints in the negative users' reviews and investigate the negative relationship with mHealth apps' privacy conducts.

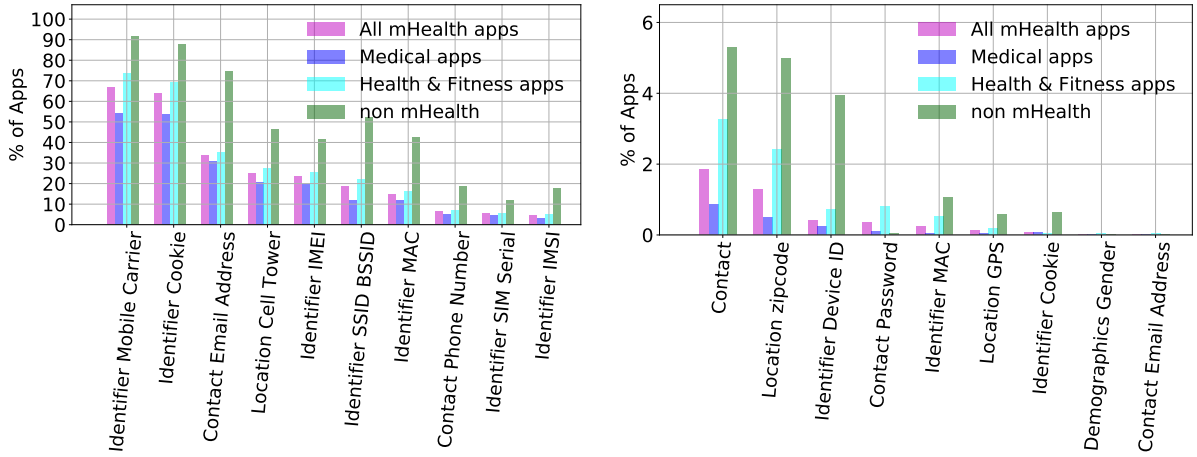
2.3 Public and Patient Involvement

We undertook this research from the perspective of apps available on Google Play in Australia. The data collection and analysis were carried on an automated testing platform designed by the authors, and with no involvement of mHealth app (physical) users nor developers.

3 Results

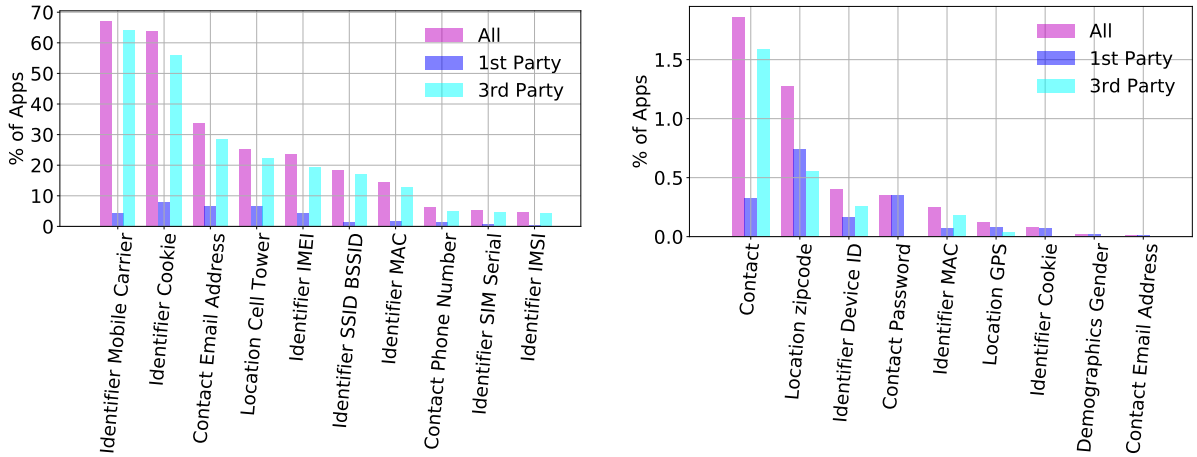
We explore how mHealth apps treat user's privacy by analysing app code/files, network traffic and user reviews. First, we investigate the data collection practices of mHealth apps, focusing on the personal (user's and user device's) information that is collected by the app's own servers (*first party*) or external services (*third parties*). Then, we analyse the main third parties interacting with the apps and retrieving user information. Lastly, we highlight key mis-behaviours in the apps' privacy conduct and the key user concerns expressed in the app reviews.

[10] https://recon.meddle.mobi/codeanddata.html



(a) Data collection practices in app code **(b)** Personal data leakage in app traffic

Figure 3. Personal data collection practices found in mHealth apps files/code and in mHealth apps traffic.



(a) First-party and Third-party practices in app code **(b)** First-party and Third-party data leaks in app traffic

Figure 4. Personal data recipients in mHealth apps files/code and in mHealth apps traffic.

3.1 Personal Data Collection Practices

The analysis of app files/code yielded a total of 65,068 total operations involving personal data collection, on average 4 per app, while from the analysis of the app traffic we identified 3,148 leaks of personal information across the traffic of 616 apps. For simplicity, we refer to both as "data collection practices". The main types of the collected data include personal and device information, user location, contact details, and more. Table 9 (Appendix A), describes the user data collected by mHealth apps, considered in our analysis.

Personal information collected by mHealth apps: Figures 3a and 3b depict the data collection practices found in mHealth app code and traffic, respectively. These are broken down according to the app category (All mHealth, Medical, and Health & Fitness). As shown Figure 3a, the majority of mHealth apps include code for collecting the Mobile Carrier Identifier[11] (67% of apps) and the app Cookies[12] (64% of apps). Other frequently collected data includes user's email address and current cell-tower location, found in 33%

[11] Allows to determine telephony services and states, and access some subscriber information.
[12] Small text files used for customising web browsing and app experience, but also for generating online user profiles.

and 25% of apps, respectively. Considering personal data leakage in the intercepted app traffic, we find leaks in 4% of mHealth apps, mostly in Health & Fitness apps (74% of total leaks). Based on Figure 3b, most frequent leaks are observed for Contact (user first or full name), and for Location (zipcode). Other leaks such as GPS Location and Cookies are not as evident in the app traffic, occurring only in 0.1% of apps, while others information leaks are rare. Comparison with baseline (non-mHealth) apps in Figures 3a and 3b shows that mHealth apps, especially Medical ones, are less prone to collect and transmit user and device data. For example, leaks of the DeviceID identifier[13] occur in 4% non-mHealth apps and only in 1% of mHealth ones. Similarly, operations in the app code that retrieve the user's email address are much more frequent in non-mHealth apps (found in 74.4% apps) than in mHealth ones (found in 33.6% apps).

Upon understanding what information is being collected, we study what entities collects this information. An overview of personal data recipients is shown in Figure 4, which summarises the data collection practices by the apps' first party servers, and those on behalf of third-party services (e.g, ads and trackers). As depicted in Figure 4, the majority of data collection is triggered by third parties. In particular, 54,155 out of 61,920 data collection operations in the app code (87%, Figure 4a) are conducted by third parties, *i.e.,* they originate from third-party libraries embedded in the apps. At the same time, 1,756 out of 3,148 personal information leaks (56%, Figure 4b) are towards third-party servers.

3.2 Third-Party Data Recipients

The above results indicated a prevalence of third parties in the user data collection by mHealth apps. We further analyse the penetration of third party services and we present the main third-party entities receiving user data. Overall, we identified 665 unique entities participating in the data collection practices found in the apps code/files and traffic. Out of the 665, a small list of prominent third parties is responsible for the majority of the data collection; namely, top-50 third parties receive 67.8% of the collected data.

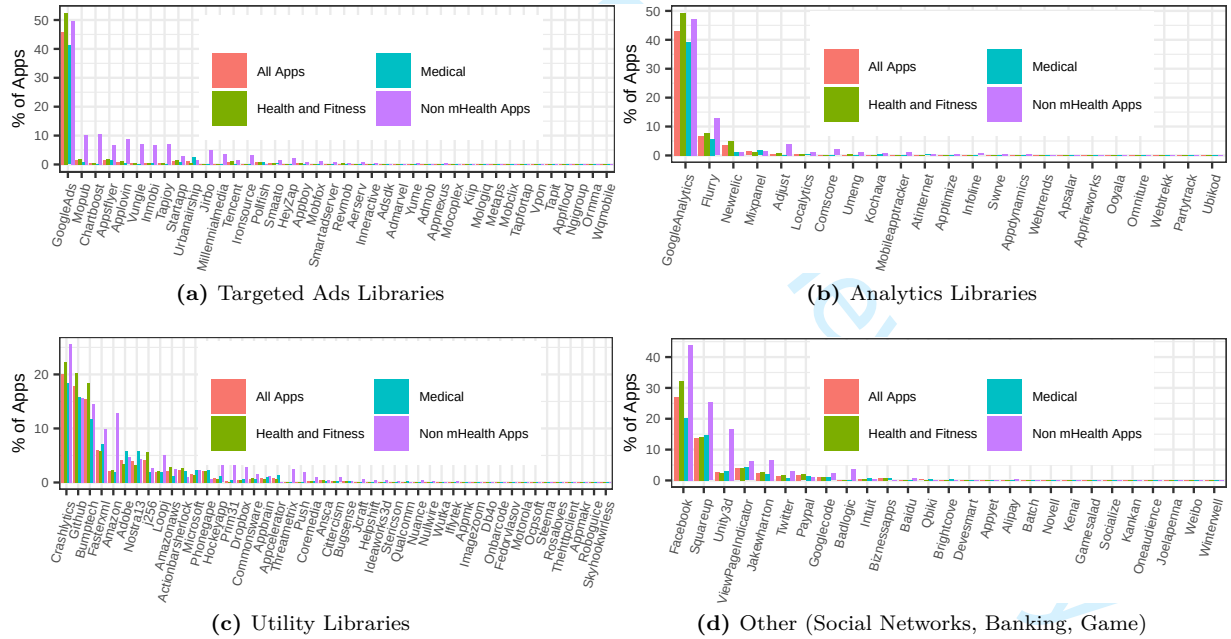
Third-party presence: We quantified the presence of third party services in mHealth apps and compared it with baseline, non-mHealth apps. In general, a strong integration (in code/files) and interaction (in traffic) with third parties indicates an increased collection of user data. As part of their engagement with app developers, third party services typically reserve the right to collect user data, and often also to share it with commercial partners or transfer it as a business asset.

The top row of Table 2 reports the number of third party libraries found across the different categories of apps. We observe that although 62.72% of mHealth apps embed at least one third party service, this penetration of third parties is substantially lower than in non-mHealth apps. In particular, only 5.71% of mHealth apps include 6 or more third party libraries, whereas for non-mHealth apps this ratio stands at 43.32%. We also note that although Medical and Health & Fitness categories exhibit similar trends, the latter integrate slightly more third parties in their code: 44.07% of Medical apps include no such libraries compared to 33.83% of Health & Fitness apps. The lower presence of third party code in Medical apps is likely to explain why data collection is less frequent in Medical apps (Figure 3a).

[13]Unique customer identifier, used also by advertisers.

Table 2. Overall third parties presence in mHealth apps: number of third-party libraries found in the app code, and % network traffic related to ad and tracker services.

		% Apps		
	mHealth(all)	Medical	Health & Fitness	non-mHealth
Number of embedded third party libraries				
0	37.28	44.07	33.83	6.18
1	21.87	23.35	20.88	5.38
2	12.99	12.44	13.19	9.62
3	9.05	7.33	9.99	13.29
4	6.65	4.71	7.71	11.68
5	6.46	3.06	8.34	10.52
≥6	5.71	5.05	6.07	43.32
Ads in network traffic (% requests)				
0.0%	94.74	95.32	93.57	81.97
0.0% – 2%	0.87	1.43	0.77	5.09
2% – 5%	0.86	0.71	1.10	4.50
5% – 10%	0.9	0.53	1.46	1.37
10% – 20%	0.78	0.29	1.10	3.91
> 20%	1.75	1.67	1.97	3.13
Trackers in network traffic (% requests)				
0.0%	90.87	91.60	89.30	79.82
0.0% – 2%	0.77	0.71	0.86	4.01
2% – 5%	2.03	1.43	2.51	4.70
5% – 10%	1.81	1.31	2.03	4.40
10% – 20%	1.91	2.03	2.12	2.74
> 20%	2.59	2.87	3.16	4.30



(a) Targeted Ads Libraries

(b) Analytics Libraries

(c) Utility Libraries

(d) Other (Social Networks, Banking, Game)

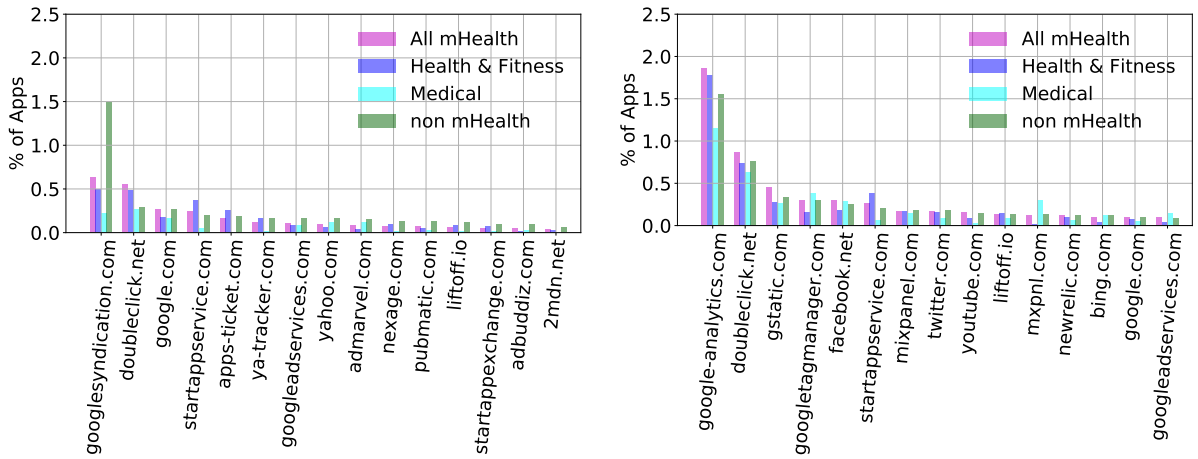
Figure 5. Third-party libraries found in various mHealth app categories and non-mHealth apps.

The bottom rows of Table 2 detail the presence of third party services in the app traffic, focusing on ad and tracker services[14]. We observe that mHealth apps tend to have less interaction with ads and tracking services than non-mHealth apps. In particular, only for 5% of mHealth apps we observe ads-related traffic, compared to 18% of non-mHealth apps. Similar applies to tracking, found in the traffic of over 20% of non-health

[14]Other third party services (e.g., social, widgets) have negligible presence in the intercepted traffic.

apps and only in 9% of mHealth apps. Focusing on the mis-behaving apps, we identify a small number of mHealth apps with significant presence of ads, which includes popular Health & Fitness apps (see Table 11 in Appendix C).

Most frequent third parties: We explore the most frequently present third party services in mHealth apps code and traffic. Figure 5 reports the third party libraries detected in mHealth apps code/files. The main ones are `GoogleAds` (advertisements) and `GoogleAnalytics` (analytics), included in almost 50% of Health & Fitness apps and 45% of Medical apps. While the results are mainly consistent across the two mHealth app categories, mHealth apps incorporate less `Facebook` widgets and they are integrated with `SquareApp` payment service and `Amazon` services weaker than non-mHealth apps.



(a) Ads **(b)** Trackers

Figure 6. Top-15 Ad and Tracker domains in mHealth and non-mHealth apps

Considering the interaction with third parties in the app traffic, Figure 6 reports the most prominent ads and tracker services contacted by the apps. We observe that the most frequent third parties are Google ad and tracking services, `googlesyndication.com` and `doubleclick.net`[15], while the top tracking domain is `google-analytics.com`.

Third party data recipients: We present the third-party entities receiving user data from mHealth apps. These are reported in Table 3, in which we include both data collection practices in the app code/files and data leakage in the app traffic. The Table details the third parties by the name of the commercial entity, and by the company-owned Internet domain detected in the app analysis. Considering the practices in app code/files, a substantial fraction is triggered by Google services (e.g., Android operating system support), as demonstrated by the considerable presence of domains `google.com`, `support.android.com`. Besides Google-owned services, we note a significant presence of Facebook (14% of apps embed Facebook cookies), Flurry analytics (6.3% of apps) and PayPal payment service. Considering the app traffic leaks, we observe that Contact data is mainly transmitted to analytics services (e.g., Google's `crashalytics.com`), while the Location and DeviceID leaks are mainly towards ads (e.g., Liftoff app marketing) and smartphone notification services (e.g., Pushwoosh).

3.3 Privacy Conduct Issues

Privacy information disclosure: We check if mHealth app developers inform their users about the app privacy practices. To assess this, we check if mHealth apps provide a public privacy policy. Privacy policy is the main means to declare the collection and use of personal data and outline the app's privacy protection practices. Since 2018, Google requires app developers to disclose the collection and sharing of user data [3].

[15]Indicates the use of Google AdSense or Google Ad Manager for loading and managing ads.

Table 3. Main third-parties involved in user data collection practices

Collected data	Main third-party recipients of user data [% mHealth apps]			
Data collection practices in mHealth app code/files:				
Identifier Carrier	Google [34.0%] (google.com)	Facebook [10.1%] (facebook.com)	Verizon [6.26%] (flurry.com)	Amplitude [2.89%] (amplitude.com)
Identifier Cookie	Google [21.0%] (google.com)	Facebook [14.1%] (facebook.com)	Apache [10.8%] (apache.org)	PayPal [1.31%] (braintreepayments.com)
Contact Email	Google [22.5%] (google.com)	Google [1.06%] (androidquery.com)	Apache [1.06%] (apache.org)	Biznessapps [0.94%] (biznessapps.com)
Location Cell Tower	Google [11.8%] (support.android)	Google [4.43%] (appcompat.androidx)	Facebook [1.74%] (facebook.com)	PayPal [1.72%] (paypal.com)
Identifier IMEI	New Relic [4.37%] (newrelic.com)	acra.org [1.82%] (acra.org)	Verizon [1.68%] (flurry.com)	Google [1.08%] (fabric.io)
Identifier SSID BSSID	Facebook [7.41%] (facebook.com)	Google [1.82%] (google.com)	StartApp [1.31%] (startapp.com)	PayPal [1.31%] (paypal.com)
Identifier MAC	Leanium [1.62%] (learnium.com)	PayPal [1.52%] (paypal.com)	Google [1.17%] (fabric.io)	Pollfish [0.94%] (pollfish.com)
Contact Number	Learnium [1.58%] (learnium.com)	Digits Financial, Inc [0.32%] (digits.com)	mobimento.com [0.32%] (mobimento.com)	Paypal [0.26%] (paypal.com)
Identifier SIM Serial	Paypal [1.74%] (paypal.com)	Tencent [0.39%] (tencent.com)	Swelen [0.26%] (swelen.com)	Pushwoosh Inc. [0.24%] (pushwoosh.com)
Identifier IMSI	Paypal [1.72%] (paypal.com)	Ogury [0.24%] (presage.io)	Anywhere Software [0.18%] (b4a.anywheresoftware)	StartApp [0.18%] (startapp.com)
Data collection practices in Health app traffic:				
Contact	Google [1.81%] (crashlytics.com)	New Relic [0.05%] (newrelic.com)	AgileMD [0.04%] (agilemd.com)	Appioapp [0.04%] (appioapp.com)
Location zipcode	Stack [0.29%] (bidmachine.io)	Amazon [0.20%] (amazon-adsystem.com)	Tapatalk [0.08%] (tapatalk.com)	MobTech [0.07%] (mob.com) [0.07%]
Identifier Device ID	Pushwoosh [0.25%] (pushwoosh.com)	PushBots [0.02%] (pushbots.com)	InManage Ltd. [0.02%] (inmanage.com)	Insider [0.01%] (useinsider.com)
Identifier MAC	Google [0.14%] (crashlytics.com)	Axway [0.02%] (appcelerator.net)	Alibaba [0.01%] (umeng.com)	Jiguang-Aurora [0.01%] (jpush.cn)
Location GPS	Liftoff [0.04%] (liftoff.io)	Kiip [0.02%] (kiip.me) [0.02%]	Airnow Monetization Ltd [0.02%] (airpush.com) [0.02%]	Chukong Technologies [0.01%] (sdkbox.com) [0.01%]
Contact Password	Web Apps [0.04%] (fitnessitaly.com)	Artexe [0.01%] (zerocoda.it)	JVS Group [0.01%] (softcliniclive.com)	AlleDaags [0.01%] (samenvoeden.nl)

Table 4. mHealth apps with privacy policy on the Google Play store.

Apps	with Privacy Policy (%)	without Privacy Policy (%)
Medical:	**5,439 (67.36%)**	**2,635 (32.64%)**
Geoblocked	730 (13.42%)	208 (7.89%)
Paid	701 (12.89%)	887 (33.66%)
Free	4,008 (73.69%)	1,540 (58.44%)
Health & Fitness:	**9,648 (74.69%)**	**3,269 (25.31%)**
Geoblocked	745 (7.72%)	189 (5.78%)
Paid	910 (9.43%)	728 (22.27%)
Free	7,993 (82.85%)	2,352 (71.95%)
Number of installs:		
<100	1,713/2,929 (58.48%)	1,216/2,929 (41.52%)
100 – 1K	3,110/4,689 (66.32%)	1,579/4,689 (33.68%)
1K – 10K	4,066/5,692 (71.43%)	1,626/5,692 (17.63%)
10K – 100K	3,752/4,835 (77.60%)	1,083/4,835 (22.40%)
100K – 1M	1,891/2,257 (83.78%)	366/2,257 (16.22%)
>= 1M	556/589 (94.39%)	33/589 (5.61%)
Total (20,991)	**15,088 (71.87%)**	**5,903 (28.13%)**

Table 5. Consistency of data collection disclosure in the privacy policy with app traffic data leaks.

Category	Data Leaks	No Privacy Policy [%]	Complying [%]	Violating [%]
All	3148	28.7	47.2	24.1
Health & Fitness	2353	36.2	37.5	26.3
Medical	795	16.8	55.4	27.8

The top row of Table 4, reporting the fraction of apps with/without privacy policy, shows a worrisome picture. Out of the 20,991 mHealth apps, 5,903 (almost 28%) provide no valid privacy policy. Between the two mHealth categories, Medical apps are the ones that comply less with the privacy policy requirement, as only 67% of Medical apps provide privacy policy. Interestingly, we also find a positive correlation between the app popularity, i.e., number of installs, and the fraction of apps providing a privacy policy (see bottom row of Table 4). However, only for very popular mHealth apps, with one million downloads or more, the fraction of apps including a privacy policy on Google Play is dominant ($\approx 95\%$).

Non-compliance with privacy policies: Upon extracting the privacy practices disclosed in the app privacy policy, we reconsider the data leaking in the app traffic, and for each leak we check if the collection of this data is disclosed in the privacy policy. This way, we tag the leaks as either *complying* with or *violating* the privacy policy. We present the results in Table 5, where we also report the fraction of leaks occurring in apps with no privacy policy. We observe that 55.4% and 37.5% of leaks in Medical and Health & Fitness apps, respectively, comply with the privacy policies. The fraction of violations, above 26%, is consistent across the two app categories. However, for Health & Fitness apps, a larger portion of non-compliant leaks is associated with apps providing no privacy policy at all – 36.2% against 16.8% for Medical apps.

Measuring the fraction of compliant and non-compliant leaks for individual apps, we observe that the apps tend to either fully comply with the privacy policy or not to comply at all. While for 34% of apps we find full compliance, for 49% of apps we obtain no compliance due to either unavailable privacy policy (21.4%) or due to all the leaks violating the privacy policy (27.7%). To illustrate the compliance of individual apps, we manually inspect the compliance for 10 most popular mHealth apps with traffic leaks. As reported in Table 6, only four of these apps disclose and fully comply with their privacy policy. For the rest, no data leaks are declared in the privacy policy, if available at all.

Table 6. Consistency of privacy policy (PP) with user data leaks in popular mHealth apps.

App	Downloads	Traffic leaks	Has PP?	PP violation [%]	Leaking data
Period Tracker	100M+	3	Yes	66.67	Location zipcode (1stParty)
Calorie Counter	50M+	5	Yes	0.0	-
My SOS Family	10M+	4	No	-	-
Noom: Health & Weight	10M+	4	Yes	0.0	-
White Noise Lite	5M+	21	Yes	0.0	-
Push Ups Workout	5M+	7	Yes	100.0	Location zipcode (1stParty)
Smart Coach for Health	5M+	4	No	-	-
Linchpin Mobile	1M+	26	No	-	-
Baby Sleep	1M+	33	Yes	100.0	Contact (1stParty)
Health Mate	1M+	10	Yes	0.0	Location zipcode (1stParty) -

Table 7. Leaks of user data in HTTP and HTTPS traffic

Data leaks in app traffic	Leaks	Leaking apps	HTTP leaks[%]	HTTPS leaks[%]
Contact	1413	311	5.37	94.6
Location zipcode	1075	214	36.7	63.2
Identifier Device ID	248	68	8.87	91.1
Contact Password	116	59	75.8	24.1
Identifier MAC	57	42	26.3	73.6
Location GPS	149	20	42.2	57.7
Identifier Cookie	70	13	85.7	14.2
Demographics Gender	17	3	5.88	94.1
Contact Email Address	3	2	100	0.0

Insecure transmission of user data: We also assess whether traffic data leaks occur on secure communication. To this end, we calculate the portion of leaks on unencryped communications using the HTTP protocol and on secure communication using the HTTPS protocol. In the light of recent reports of widespread Internet surveillance [1] and legislation permitting internet service providers to sell user information extracted

from network traffic [2], HTTPS transmission of user data is essential for user privacy protection [18]. Analyzing the communication leaking personal data, we observe that as much as 23% of leaks are in unencrypted HTTP traffic. Table 7 reports the breakdown of leaks in HTTP and HTTPS traffic for the various types of leaking data. While for most data the leaks mainly occur in HTTPS, for some sensitive data, e.g, Contact Password, GPS location and Contact Email, significant leaks are observed in the un-encrypted HTTP.

3.4 User Perceptions of mHealth Apps

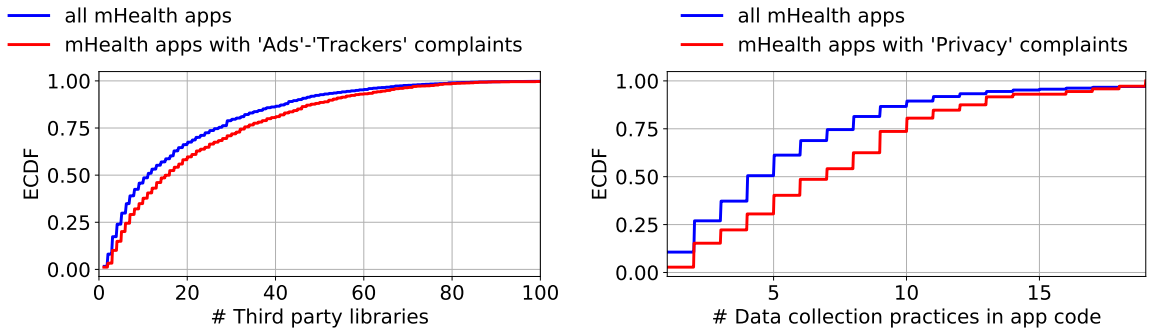
By analysing the mHealth app reviews, we quantify the users' perceptions of mHealth apps. In particular, we focus on the *negative* reviews (*i.e.,* ratings ≤ 2) to investigate the users' concerns around the app functionality and privacy conduct. To identify the complaints raised by users in the review text, we compiled a list of 59 keywords mapped to 12 complaint categories (full list in Appendix D). For example, the keyword `crash` is mapped to the category 'bugs', while the keyword `private` is mapped to 'privacy'. A scan of all the negative reviews (235,210) yielded a set of 391,642 user complaints, of which 67,057 referred to Medical apps and 324,585 – to Health & Fitness.

Overall user perception: In Table 8, we detail six user complaint categories, of which three refer to the application *usability*, and three refer to the app *privacy* conduct. As can be seen, most of the complaints (53%) point to app *usability* flaws: bug reports (e.g., unexpected crashes) or excessive data or battery consumption. In particular, close to 51% of the complaints are related to bugs, mentioned for 11% of mHealth apps. Compared to bugs, user complaints related to privacy are much less frequent. In particular, only 0.9% of negative reviews explicitly mention the privacy of personal data. This suggests that mHealth app users have a limited interest in (or awareness of) privacy issues.

Table 8. Breakdown of user complaints found in mHealth apps reviews.

Complaint Category	All mHealth (391,642 complaints)				Medical (67,057 complaints)				Health & Fit. (324,585 complaints)			
	#Compl.	%Compl.	#Apps	%Apps	#Compl.	%Compl.	#Apps	%Apps	#Compl.	%Compl.	#Apps	%Apps
Usability:												
Bugs	201,240	50.97	2,240	10.67	34,728	51.8	627	7.7	166,512	51.3	1,613	12.48
Battery	7,710	1.95	568	2.70	4,784	7.13	120	1.48	2,926	0.9	448	3.46
Mobile Data	2058	0.52	427	2.03	169	0.25	70	0.86	1787	0.55	305	2.36
Privacy:												
Privacy	3,609	0.9	351	1.67	990	1.48	80	0.99	2,619	0.81	271	2.09
Ads	43,794	11.09	1,128	5.37	10,702	15.96	262	3.2	33,092	10.2	866	6.70
Trackers	29,827	7.5	942	4.48	6,976	10.4	138	1.7	22,851	7.04	804	6.22

Privacy-related user complaints: Focusing on privacy-related complaints in Table 8, we find that the vast majority of these refer to ads and trackers, mentioned in 11.09% and 7.5% of negative reviews, respectively. Overall, complaints on intrusive ads and trackers are raised for 2,070 mHealth apps – almost 10% of the studied apps. Direct privacy complaints are much less frequent than bug- or ad-related complaints, as they only appear for 1.67% of mHealth apps.



(a) Number of 3rd-party libraries in apps code **(b)** Number of data collection practices in apps code

Figure 7. Relation between user complaints and the privacy conduct of mHealth apps, expressed in terms of third-party presence and data collection operations in each mHealth app.

We further investigate the apps targeted by privacy-related complaints, looking for correlation between the complaints and the actual conduct of the apps. For these apps, Figure 7 reports the number of third-party libraries (Figure 7a) found in the app code and the number of data collection practices (Figure 7b). For comparison, we also report the values obtained for the complete mHealth apps set.

We found that users' concerns around ads/trackers or privacy generally reflect a more pronounced ad/tracking penetration or a more intense data collection activity in the apps. As shown in Figure 7a, mHealth apps with complaints on ads or trackers embed more third-party libraries, likely associated with ads/trackers. In this case, the median number of embedded libraries is 16, 45% higher than the median for all mHealth apps. Similarly, considering the number of data collection practices in the apps, we observe that in cases where reviews include direct privacy complaints, the apps actually retrieve more personal information than expected. Based on Figure 7b, mHealth apps with privacy complaints include a median of 7 data collection practices, which is higher than the median of 4 for the complete mHealth app dataset.

Overall, the analysis of user reviews reveals that while mHealth app users have a limited interest in (or awareness of) the apps' privacy conduct, there exists a correlation between the users' concerns expressed in negative reviews and objectively measured aspects of privacy conduct, such as the presence of ads/trackers and the inclusion of user data collection operations.

4 Discussion

4.1 Overview of Findings

Our analysis of data collection practices, performed on a comprehensive set of 20,991 mHealth apps, revealed that the vast majority of the apps (88%) automatically retrieve and transmit personal user data. Despite being less prone to share user information than non-mHealth apps, each mHealth app conducted, on average, 4 user data collection practices. Amongst the data retrieved by these apps, we found a significant presence of persistent identifiers (e.g., IMEI, MAC address in that cannot be reset by users) and sensitive user information (e.g., contact name, email, phone number). While the former allows to track individuals over time and across different services, the latter refers directly to an individual's privacy. Moreover, by collecting and sharing cookies (64% apps), mHealth apps actively contribute to the creation of online user profiles for advertisers. Our analysis also showed that Health and Fitness apps were generally more prone to collecting and sharing user information than Medical apps. This is in accordance with the more pronounced integration of Health and Fitness apps with third-party ad and tracking services, shown in Figures 5 and 6.

The role of third-parties was predominant in the data collection practices of mHealth apps, as more than

80% of user information was retrieved on behalf of third-party services. In particular, our analysis found that a limited number of third-party services (fewer than 50) are recipients of close to 70% of user information. Among these, Google-owned services were the most recurrent in the analysed app set: GoogleAds software and DoubleClick advertisement service, GoogleAnalytics and Crashalytics analytics systems, and Android operating system support services. This indicates a concentration of data transmission towards a single entity, leveraging it for advertising and analytics services.

While the retrieval and sharing of user information by mHealth apps is routine, data collection practices are far from being transparent. Our comparative analysis of the app privacy policies and the actual user information leakage, depicted a worrisome scenario. Close to 30% of the apps do not offer any privacy policy text, and more than 25% of user data transmissions violate what is stated in the privacy policies. Another issue that raised particular concerns referred to the transmission of sensitive user information, such as GPS location (42%) or password (75%), using insecure communication channels. This is alarming, given the recent reports on Internet surveillance and unwanted commercialisation of user data [2,18]. Despite these issues being topical, our analysis of mHealth app reviews discovered that the app users have a limited awareness of (or interest in) the privacy conduct of the apps.

4.2 Strengths and Limitations of the Study

Our study is, to the best of our knowledge, the first attempt to profile the privacy conduct of mHealth apps *at scale* (more than 20,000 apps), *i.e.,* by considering most of the medical, health, and fitness-related apps available on the biggest existing mobile app store (Google Play). The study touches upon a multitude of viewpoints: the analysis of data collection practices and user data recipients, the dissection of apps privacy policies, and the analysis of publicly available app reviews. To identify as many data collection practices as possible, our study combined the analysis of static app resources (application code and files), with the inspection of the network traffic generated by the apps at run time. This allowed for a comprehensive view on the privacy handling of user information mHealth apps can retrieve and share.

To scale up the study and cope with a large number of mHealth apps, we leveraged automated analysis tools as well as state-of-the-art machine learning techniques. While these techniques provided high validation accuracy (above 96% for both the detection of privacy-leaks and disclosure of privacy practices), they might still generate limited false positives. To mitigate the scale of the app set, our live testing of mHealth apps heavily relied on extensive randomised interactions as opposed to hand-crafted app usage patterns/profiles, with the drawback that some parts of the applications (e.g., tabs, views, menus) might have not been triggered during testing. Lastly, it should be noted that we restricted the analyses to free apps only; however, we conjecture this does not strongly penalise the generality of our findings, since no more than 15% of mHealth apps on Google Play were paid.

4.3 Comparison with Prior Work

Mobile health applications and the associated privacy risks have received significant attention from the research community in the recent years. Huckvale et al. investigated the privacy of 79 health and wellness mobile apps accredited by UK National Health System [10]. They found that most of the apps transmitting user information (78%) did not describe their data collection practices in the privacy policies. Blenner et al. analysed 24 Android diabetes apps and discovered that 79% of these apps shared user information despite not providing any privacy policy [6]. Upon assessing the privacy practices of 36 top-ranked apps for smoking cessation and depression, Huckvale et al. revealed that only a small fraction of these apps (12 out of 29) disclosed the transmission of data to Facebook or Google in their privacy policies [11]. While these studies focused on consistency between the data collection practices and the privacy policies of mHealth apps, the work by Grundy et al. focused on the recipients of user information collected by 24 medical apps [9]. In line with our findings in Section 3.2, a prevalence of analytics and advertisement services among user data recipients was shown.

Compared to the aforementioned works, our study substantially differs in the scale of the analysis, which is orders of magnitude higher: more than 20,000 mHealth apps on Google Play, out of which 15,838 were analysed in-depth, as opposed to tens of apps assessed in previous studies [6,9–11]. To the best of our knowledge, the only study spanning a comparable range of mHealth apps was conducted in 2015 by Dehling et al. [7]. However, it is important to note that their analysis only categorised mHealth apps into classes of potential risk (e.g., low/medium/high risk of privacy leaks), while not providing any result on the (i) type of collected user information, (ii) recipients of this information, and (iii) consistency of the app practices with the disclosed privacy policies.

Our study also presents a broad assessment of mHealth apps that is missing in the existing analyses. In previous studies, the analysis was generally restricted to the data transmitted by mHealth apps [9] or to the consistency of the apps with their privacy policies [10] [6]). In contrast, our study offers a comprehensive view on the privacy risks associated with mHealth apps by collectively considering the information the apps transmit or can access through their code, the potential recipients of this information, the security of and transparency in user data transmission, and the perceptions of the app users around the performance of apps and their associated privacy conduct.

Our results showed that, compared to baseline non-mHealth apps, mHealth apps are generally less prone to access and transmit user data. However, considering the concentration of user data transmission towards the dominant third-party services, our findings are aligned with recent large-scale analyses of tracking and data sharing ecosystem in mobile apps [5,17,22]. The analysis of 959,426 apps found that most trackers embedded in the apps were linked to a small number of commercial entities, with Google being the most prominent [5] . Similarly, traffic analysis of 14,599 Android apps found that despite owning just 4% of all third-party tracking services, Google was present in 73% of the analysed apps [17].

5 Conclusions

This work investigated the privacy conduct of mHealth apps, belonging to the Medical and Health & Fitness categories on the Google Play store. To this end, we developed an infrastructure to analyze more than 20,000 apps. We found that the majority of apps collect and share data with third-parties, including advertising and tracking services. Interestingly, the apps collected user data on behalf of hundreds of third-parties, with a small number of service providers accounting for most of the collected data. Alarmingly, large portion of privacy leaks occurs on insecure communication protocols, putting user privacy at risk. The analysis also revealed that mHealth apps are far from transparent when dealing with user data, with only about half of the apps found to be compliant with their own privacy policies (if disclosed at all). Moreover, our review analysis suggests inadequate understanding of the apps' privacy practices by the end users.

Mobile apps are fast becoming sources of information and decision-support tools for clinicians and patients alike. Given that our analyses uncovered worrisome privacy issues and limited user awareness, we argue that it is important to surface our findings around potential privacy risks and bring them to the attention of clinicians. They should be cognisant of these risks and consider them carefully, to ascertain that the benefits of an app outweigh its risks. On top of this, it is important to articulate such privacy risks to patients and potentially make this an inherent part of the app usage consent. Moreover, it is critical to consider the trade-off between the benefits and risks of mHealth apps for any technical and policy discussion surrounding the services provided by such apps.

Addenda

Funding: This work was funded by Optus Macquarie University Cyber Security Hub; the research was also supported by the National Health and Medical Research Council (NHMRC) grant APP1134919 (Centre for Research Excellence in Digital Health) led by Prof Enrico Coiera. G.Tangari and K.Jiaz were supported by a postdoctoral fellowship from Macquarie University. Optus Macquarie University Cyber Security Hub and

the NHMRC Centre of Research Excellence (CRE) in Digital Health had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: this work was funded by the Optus Macquarie University Cyber Security Hub and the NHMRC Centre of Research Excellence (CRE) in Digital Health; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Data sharing: A sample of our dataset is available at https://mhealthapps2020.github.io/. Note that, upon publication, we will release all our dataset and analysis script for further research.

Transparency: The authors affirm that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned have been explained.

Ethical approval: Ethical approval was not sought as the study did not involve human subjects.

References

1. NSA and GCHQ target Tor network that protects anonymity of web users. http://www.theguardian.com/world/2013/oct/04/nsa-gchq-attack-tor-network-encryption, 2013.
2. Congress Overturns Internet Privacy Regulation. http://www.npr.org/2017/03/28/521831393/congress-overturns-internet-privacy-regulation, 2017.
3. User Data | Privacy, Security, and Deception - Developer Policy Center. https://play.google.com/intl/en-US/about/privacy-security-deception/user-data, 2020.
4. Saba Akbar, Enrico Coiera, and Farah Magrabi. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *Journal of the American Medical Informatics Association*, 27(2):330–340, 2020.
5. Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt. Third party tracking in the mobile ecosystem. In *WebSci*, 2018.
6. Sarah R Blenner, Melanie Köllmer, Adam J Rouse, Nadia Daneshvar, Curry Williams, and Lori B Andrews. Privacy policies of android diabetes apps and sharing of health information. *Jama*, 315(10):1051–1052, 2016.
7. Tobias Dehling, Fangjian Gao, Stephan Schneider, and Ali Sunyaev. 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):e8, 2015.
8. FDA. Digital health criteria. https://www.fda.gov/medical-devices/digital-health/digital-health-criteria, 23/03/2018.
9. Quinn Grundy, Kellia Chiu, Fabian Held, Andrea Continella, Lisa Bero, and Ralph Holz. Data sharing practices of medicines related apps and the mobile ecosystem: traffic, content, and network analysis. *bmj*, 364:l920, 2019.
10. Kit Huckvale, José Tomás Prieto, Myra Tilney, Pierre-Jean Benghozi, and Josip Car. Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment. *BMC medicine*, 13(1):214, 2015.

11. Kit Huckvale, John Torous, and Mark E Larsen. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. *JAMA network open*, 2(4):e192542–e192542, 2019.

12. Muhammad Ikram and Mohamed Ali Kâafar. A first look at mobile ad-blocking apps. In *16th IEEE International Symposium on Network Computing and Applications, NCA 2017, Cambridge, MA, USA, October 30 - November 1, 2017*, pages 343–350, 2017.

13. Muhammad Ikram, Narseo Vallina-Rodriguez, Suranga Seneviratne, Mohamed Ali Kaafar, and Vern Paxson. An analysis of the privacy and security risks of android vpn permission-enabled apps. In *Proceedings of the 2016 Internet Measurement Conference*, IMC '16, pages 349–364, New York, NY, USA, 2016. ACM.

14. Sarah J Iribarren, Kenrick Cato, Louise Falzon, and Patricia W Stone. What is the economic evidence for mhealth? a systematic review of economic evaluations of mhealth solutions. *PloS one*, 12(2):e0170581, 2017.

15. Misha Kay. mhealth: New horizons for health through mobile technologies. *World Health Organization 64.7*, 2011.

16. Mara Mercurio, Mark Larsen, Hannah Wisniewski, Philip Henson, Sarah Lagan, and John Torous. Longitudinal trends in the quality, effectiveness and attributes of highly rated smartphone health apps. *Evidence-Based Mental Health*, 2020.

17. Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *NDSS*, 2018.

18. Jingjing Ren, Martina Lindorfer, Daniel J Dubois, Ashwin Rao, David Choffnes, and Narseo Vallina-Rodriguez. Bug fixes, improvements, ... and privacy leaks: A longitudinal study of pii leaks across android app versions. In *NDSS*, 2018.

19. Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. ReCon: Revealing and controlling PII leaks in mobile network traffic. *MobiSys 2016 - Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 361–374, 2016.

20. Simon P. Rowland, J. Edward Fitzgerald, Thomas Holme, Powell John, and Alison MacGregor. What is the clinical value of mhealth for patients? *npj Digital Medicine*, 3(1):4, January 2020.

21. Joseph Tighe, Fiona Shand, Rebecca Ridani, Andrew Mackinnon, Nicole De La Mata, and Helen Christensen. Ibobbly mobile health intervention for suicide prevention in australian indigenous youth: a pilot randomised controlled trial. *BMJ open*, 7(1):e013518, 2017.

22. Narseo Vallina-Rodriguez, Srikanth Sundaresan, Abbas Razaghpanah, Rishab Nithyanand, Mark Allman, Christian Kreibich, and Phillipa Gill. Tracking the trackers: Towards understanding the mobile advertising and tracking ecosystem. *arXiv preprint arXiv:1609.07190*, 2016.

23. Sebastian Zimmeck, Peter Story, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. MAPS: Scaling privacy compliance analysis to a million apps. In *19th Privacy Enhancing Technologies Symposium (PETS 2019)*, volume 3, pages 66–86, Stockholm, Sweden, July 2019. Sciendo.

A User data collection practices

Table 9. Personal data types identified through the analysis of mHealth apps code/files and traffic.

Type of data	Description	#Apps (%)
Identifier Mobile Carrier	Identifier of the user's mobile network operator	3,266 (20.6)
Identifier Cookie	Cookies include a randomly generated user/client id, which identifies user's mobile app instance	3,108 (19.6)
Contact Email Address	User's email address registered with the app or with the operating system (Android)	1,636 (10.3)
Location Cell Tower	Location of the cellular base station to which the user's device is connected	1,224 (7.73)
Identifier IMEI	International Mobile Equipment Identity (IMEI) is a number, usually unique, that identifies a mobile device	1,143 (7.2)
Identifier MAC	Unique identifier of the network interface in the user's device	712 (4.49)
Identifier SSID BSSID	Name and MAC address of the network access point to which the user's device is connected	902 (5.69)
Contact Phone Number	User's device phone number	308 (1.94)
Identifier SIM Serial	The number of the physical SIM inserted in the user's device, used for international identification	263 (1.66)
Identifier IMSI	A number that uniquely identifies every user of a cellular network. To prevent eavesdroppers from identifying and tracking the subscriber on the radio interface, the IMSI is sent as rarely as possible.	221 (1.39)
Contact (name)	User's first or full name	311 (1.96)
Location zipcode (country)	Current user's zipcode registered by the user, including country code	214 (1.35)
Identifier Device ID	A unique identifier of the operating system (Android) instance on the user's device.	68 (0.42)
Contact Password	User's app login password	59 (0.37)
Location GPS	Exact GPS location of the user's device	
Demographics Gender	User's gender	3 (0.01)

B Accuracy of automated privacy policy analysis

Table 10. Prediction of data-collection disclosure in the app privacy policy text: validation accuracy on the APP-350 corpus[23].

Disclosed data-collection practice	Accuracy	AUC	Precision	Recall
Contact	0.98	0.97	0.75	0.66
Location zipcode	0.99	0.83	0.94	0.65
Identifier Device ID	0.99	0.90	0.84	0.81
Contact Password	0.99	0.80	0.93	0.62
Identifier MAC	0.99	0.93	0.96	0.87
Location GPS	0.99	0.93	0.89	0.87
Identifier Cookie	0.98	0.95	0.89	0.91
Demographics Gender	0.99	0.92	0.79	0.82
Contact Email Address	0.97	0.86	0.83	0.73

C mHealth apps with strong presence of ads

Table 11. Top-10 most popular mHealth apps (1M+ installs) with strong presence of advertisements.

Application	Category	Installs	Review Rating	Ads Requests (%)
Pull Ups Workout	Health and Fitness	1.000.000+	4.75	75.0
Squat Workout	Health and Fitness	1.000.000+	4.82	72.0
Abs Sit Ups Workout	Health and Fitness	1.000.000+	4.73	69.2
Androyal	Health and Fitness	1.000.000+	4.17	68.7
Ma grossesse by Doctissimo	Medical	1.000.000+	4.34	50.0
Lifesum Diet Plan	Health and Fitness	1.000.000+	4.47	9.09
Baby Sleep	Health and Fitness	1.000.000+	4.77	8.18
Boxing Interval Timer	Health and Fitness	1.000.000+	4.65	8.0
Linchpin Mobile	Health and Fitness	1.000.000+	4.29	7.52
PsyTests	Medical	1.000.000+	4.60	3.98

D User complaint categories

Table 12. Complaint categories defined for the app review analysis.

Complaint category	**Case-insensitive keywords**
Usability:	
Bugs	crash/bug/freez/glitch/froze/stuck/stick/error/disconnect/not work/not working
Battery	battery/cpu/processor/processing/ ram /memory
Mobile Data	mobile data/gb/mb/background data/
Mal-behaviour:	
Scam	scam/credit card/bad business/bad app
Adult	porn/adult/adult ad
Offensive/Hate	sexis/LGBT/trolling/racism/offensive/islamophobia/vile word/minorities/hate speech/shit storm
Privacy:	
Privacy	privacy/private/personal details/personal info/personal data
Ads	ads/ad/advertisement/advertising/intrusive/annoying ad/popup/inappropriate/video ads/in-app ads
Trackers	tracker/track/tracking
Security:	
Security	security/tls/certificate/attack
Malware	malware/trojan/adware/phishing/suspicious/malicious/spyware
Intrusive Permissions	permission

17-11-2020

Dear BMJ Editor in Chief,

We are submitting our paper titled "***Analysing Privacy Issues of Android Mobile Health and Medical Applications***".

In this study, we conducted the first, large-scale privacy analysis of mobile health (mHealth) apps. The privacy concerns in mHealth apps that led to this study are motivated by the sensitive types of information these apps can access, their business model centred on selling subscriptions or user data, and the recent enforcement of privacy standards around the world.

We deployed a suite of app analytics tools to perform a privacy audit of more than 20,000 mHealth apps available on the Google Play store. The scale of our analysis is orders of magnitude larger than previously reported analyses and virtually covers all mHealth apps on the Google Play store accessible from Australia. Our analysis provides a comprehensive view on mHealth privacy risks by studying the data collection practices performed by mHealth apps, the recipients of user information, the security and transparency in user data transmission, and the users' perceptions around the apps and the associated privacy conduct. *To the best of our knowledge, this is the first analysis of mobile health apps at scale.*

The results of the analyses show that the majority of mHealth apps retrieve and transmit personal user information, including sensitive data that allow tracking individuals over time and across services, or can directly be used to profile individuals and their preferences. Moreover, our results depict a concentration of user data transmission towards services owned by a (very) small number of commercial entities. While being routine, data collection practices of mHealth apps are far from being transparent and secure. Alarmingly, we found that less than half of the detected user information transmissions comply with the apps' privacy policies and that sensitive user data is often shared on insecure channels, directly exposing users to data interception and surveillance risks.

Collectively, our findings paint a worrisome picture, where numerous privacy and security breaches exist, whereas their awareness is minimal. Hence, it is important to bring our findings to the attention of clinicians, in order articulate the privacy risks to patients and be able to diligently weigh the benefits and risks of mHealth apps.

We hope you find our work publishable at BMJ.

Kind regards,

Gioacchino Tangari
Muhammad Ikram
Kiran Ijaz
Dali Kaafar
Shlomo Berkovsky

Macquarie University, Australia