# I'm 2.8% Neanderthal

## The beginning of genetic exhibitionism?

Lukasz Olejnik
INRIA, Rhone-Alpes
lukasz.olejnik@inria.fr

Agnieszka Kutrowska
Department of Biochemistry
Adam Mickiewicz University
Poznan, Poland
akutr@amu.edu.pl

Claude Castelluccia
INRIA, Rhone-Alpes
claude.castelluccia@inria.fr

## ABSTRACT

Direct-to-consumer genetic testing is gaining popularity. However, the sensitive nature of personal genomic sequencing results might not be fully understood by the general public. In this paper we study the examples of disclosure of this sensitive information on social networks. We found that Twitter users often post their results publicly. We observed that information on ethnic background is much more frequently released than other information, for example relating to disease risk. This data could be of potential value to entities such as insurance companies.

We found that about 24% of the analyzed tweets that mentioned ethnicity results also contained percentage data. In cases of users disclosing more details of their ethnic background, we found about 96% of these profiles also included identifying information and consequently can be attributed to individuals.

As a result, external entities such as insurance companies can gain an insight in the genetic test results and in the end the users could be subject to genetic discrimination.

## 1. INTRODUCTION

The dawn of publicly available commercial genetic testing is almost upon us. With the recent progress of genome sequencing and the achievement of the $1000 per genome milestone [36] by Illumina [24], the promise of fast extraction of data from whole genomes will ultimately be fulfilled. Genetic sequencing follows Moore's Law [7], which suggests that personal genome sequencing will soon be widely available and ubiquitous.

However, with the new possibilities new risks appear. People are often unaware of the consequences of private data disclosure and this is exemplified by their behavior while they use social networks. Users routinely post sensitive data without proper consideration; a good example is posting credit card photos to Twitter [20, 30], which are even disseminated by a dedicated Twitter feed [18]. People also frequently post pictures of their identity documents [26]. Unsurprisingly, social network users often regret the fact that they shared too much and/or inappropriate information [38].

In this paper, we study the potential problems arising from the disclosure of genetic test results on Twitter. The significance of our study is heightened by the fundamental lack of awareness of the millions of social media users in regards to guarding sensitive data. We stress that disseminating personal genetic test results can also have ramifications for the user's relatives [23]; it could also have consequences, known as *genetic discrimination*, from health insurers [28, 17]. Health records are also being used in ad targeting [12].

We study the Twitter users who disclosed their genome sequencing results obtained from 23andMe. Although according to a recent ethnographic study, people are typically concerned with ethical and privacy risks relating to the disclosure of genetic test results [14], during the course of our study, we found that many Twitter users had no qualms about publishing this information. Companies such as 23andMe should probably devote more effort to familiarizing their users with the actual risks of disclosure.

We acknowledge the practicality and ethical nature of disclosing the genetic risks to one's relatives [15], which is often encouraged by medical communities [16].

The paper's organization is as follows: in section 3, we discuss the background behind personal genomic sequencing and how companies such as 23andMe operate. In section 4, we highlight the hazards related to the disclosure of the genotyping results. Section 5 is devoted to the results and analysis of genotyping disclosures in Twitter network.

## 2. RELATED WORK

The phenomenon of oversharing private information has been observed since the beginning of the social network era. A lot of sensitive information about the users can be extracted from their public feeds. Acquisti et al. studied this problem using Facebook [2]. Mao et al. analyzed private data leaks over Twitter and among the studied information are examples related to health [31]. Cheng et al. showed that it is possible to infer the physical residence of the user with data obtained from

Twitter [8].

*Genetic discrimination* is a phenomenon of discriminating against patients or customers based on genetic data, including information on risks of developing certain conditions or one's ethnic background [28, 17].

# 3. BACKGROUND

Human genetics and genomics research generates vast amounts of data of possible clinical relevance. This data comes from the high-throughput association studies, from the analysis of allele frequencies and studies of natural selection, genetic variation and human migration, funded by public or private institutes [9]. On the basis of these results companies such as 23andMe create profiles for each customer interested in personal genetic testing.

Using genome shotgun sequencing methods [27], The 1000 Genomes Project [10, 11] sequenced 1092 genomes of individuals from 14 populations. The genomes were then analyzed with the aim of understanding the genetic contribution to various diseases. As a result, 38 million of SNPs (single-nucleotide polymorphisms; changes of a single DNA nucleotide occurring in the same species) [4] were discovered. SNPs collected in the databases come from multiple sources [9]. Geneticists use the frequencies of the identified SNPs to associate these single-base changes with susceptibility or tolerance to various conditions. In the era of high throughput sequencing, the approach shifted from the candidate gene studies (characterized by usually small sample sizes, population stratification issues, weak effects and low reproducibility rate) towards genome-wide association studies (GWAS). In this approach no prior hypotheses are needed and millions of common variants across the genome can be tested for association with the trait of interest. Rigorous criteria must be met before a SNP can be declared associated with a disease [33].

## 3.1 Genotyping companies

### *3.1.1 23andMe*

23andMe (`23andMe.com`) is a direct-to-consumer personal genotyping company. Customers receive a collection kit in the mail, provide a saliva sample, and send it back to the company where genomic DNA is extracted from the cheek cells by a CLIA[1]-certified laboratory. The sample is then sequenced and information on SNPs is extracted. In the end, the users can gain insight into their genetic and ethnic background, as well as see a breakdown of the risks of developing certain conditions.

After being tested, the customer receives: 1) *a raw list of his/hers SNPs*, 2) *a profile generated on the basis of SNP testing*, highlighting data that were previously associated with certain traits and 3) *links to information* about research studies and lifestyle changes.

About 650$k$ of customers used 23andMe's service as of March 2014 [1]. Since November 2013, 23andMe customers obtain only ethnicity data and raw results. Raw results can be analyzed using freely available tools on the Internet using information from databases such as openSNP.

### *3.1.2 Ancestry.com*

Ancestry.com (`ancestry.com`) enables users to discover their ethnic background using genotyping methods[2]. The employed techniques are similar to 23andMe's, i.e. the user sends a saliva sample and then can access his/her ancestry and ethnicity result on Ancestry.com's Web site. Ancestry.com does not provide health risk information.

## 3.2 SNP databases

The next step for a private recipient of the SNPs profile might be reaching out to others with whom he or she shares some of the genetic variants. This service is provided e.g. by the database openSNP (`opensnp.org`). OpenSNP enables customers to publish their test results, find others with similar SNPs and learn more about the genotyping results.

OpenSNP provides datasets of SNPs gathered from different sources, including SNPedia (`snpedia.com`). We analyzed the data in the repository in order to examine how much information on the risks connected with being a carrier of known SNPs are available to the lay public. Table 1 shows the results of our analysis. Out of the 9334 SNPs listed, 2635 (around 28%) are described in a manner enabling the assessment of the risks, after the exclusion of data labeled as *No summary provided*, *Average*, *Common*, *None*, etc. For example we found that 337 known SNPs (13%) are related to the increased risk of developing different types of cancer (as seen in Table 1). The openSNP and SNPedia are open source databases, updated regularly with the newly available studies. Every user with rudimentary knowledge of genetic testing can try to assess his/her risk of e.g. developing heart conditions. This is true also for third parties, which could accidentally, or due to lack of protection of this data, obtain raw profiles of individuals.

# 4. RISKS OF DISCLOSURE

Because of the potential for genetic discrimination, disclosure of the SNP profile might undermine one's ability to obtain insurance, but this danger is not limited to the individual in question [21, 25]. Risk alleles known to have familial inheritance might imply that some negative effects could be shared among one's family members and offspring [5]. Autosomal DNA is in-

[1] Clinical Laboratory Improvement Amendments

[2] `http://dna.ancestry.com`

| Informative SNPs from the SNPedia dataset: 2635 (100%) | | | |
|---|---|---|---|
| Cancer | *(different types)* | Increased risk | 337 (12.8%) |
| | | Decreased risk | 49 (1.9%) |
| | | Poorer survival | 9 (0.3% |
| Cardiovascular system conditions | Stroke | Increased risk | 16 (0.6%) |
| | | Decreased risk | 3 (1.1%) |
| | Cardiac event | | 16 (0.6%) |
| | Cariomyopathy | | 15 (0.6) |
| | Hypertension | Increased risk | 42 (1.6%) |
| | | Decreased risk | 8 (0.3%) |
| | Cholesterol | | 31 (1.2%) |
| | Heart attack | Increased risk | 16 (0.6%) |
| | | Decreased risk | 4 (0.2%) |
| | Heart disease | Increased risk | 36 (1.4%) |
| | Associated with heart attack/stroke | | 12 (0.5%) |
| Affecting life quality | Diabetes | Increased risk | 38 (1.4%) |
| | | Decreased risk | 3 (0.1%) |
| | Obesity | Increased risk | 28 (1.1%) |
| | | Decreased risk | 3 (0.1%) |
| | Lifespan | Longer | 5 (0.2%) |
| | | Shorter | 2 (0.1%) |
| | Crohn's disease | Increased risk | 34 (1.3%) |
| | | Decreased risk | 6 (0.2%) |
| | Parkinson's disease | | 16 (0.6%) |
| | Alzheimer's disease | | 60 (2.3%) |
| Other | | | 1846 (70%) |

Table 1: Analysis of the informative SNPs obtained from SNPedia dataset and available at `opensnp.org`

herited from both parents and every child gets 50% of DNA from the biological mother and 50% from the father. Thus, a parent and a child share 50% of DNA and siblings share around 50% of DNA. Half-siblings, grandparent and a grandchild, child and aunt/uncles share 25% of their DNA. Generally, it is assumed that with each increasingly distant branch in family tree the genetic likeness drops by *half.* Identical twins are exceptional and share 100% of DNA.

If genes are located on the same chromosome in short distance from one another, they are usually genetically linked, which means they tend to be inherited together. Mitochondrial DNA is inherited exclusively from the maternal line; any trait carried by the mother will be passed onto children. Some exceptions apply to the genes located on the X and Y chromosomes [34]. Therefore, if a person discloses that he/she has a SNP resulting in a 5 times higher risk of developing (e.g) arteriosclerosis, his/her offspring will have 50% chance of sharing this SNP, assuming SNP is on the autosomal chromosome and no other rules apply. Insurance companies could use this information to strategically prevent future losses due to costly treatments, although in certain countries such as the US, legislation like Genetic Information Nondisclosure Act (GINA) attempts to protect their citizens from these particular risks [37, 35].

However, if we consider other aspects, the problem becomes even more complex. Certain genes (or SNPs) are more prevalent in certain populations and diseases can have a different penetration rate (the chances of developing a phenotype for mutation carriers) [34, 6]. SNPs can mutate over time or appear spontaneously (*de novo*) in a child's genome. Genes located in chromosome regions with a higher possibility of crossover will be recombined more often. Dominant traits (like Huntington's disease) result in a phenotype, even if a patient is a heterozygote, while recessive traits can remain hidden [34]. Many phenotypes are a result of expression of multiple genes and many illnesses can be modulated by environmental conditions (diet, exposure, physical activity, etc.). Many traits are regulated on the level of epigenetics (histone acetylation, DNA methylation) [32].

## 4.1 Disclosure of ethnicity description

While it is clear that disease risks are sensitive information, it is also important to note that just disclosing the ethnicity of the user may likewise have serious consequences.

Many hereditary conditions are more prevalent in certain ethnicities. Individuals disclosing their ethnic background reveal, therefore, that they may be subject to this elevated risk. For example, Alzheimer's risk is higher for first-degree relatives of African Americans than Whites [19]. As a result of their European background, African Americans are also more likely to develop multiple sclerosis (MS), compared to Africans [13]. South Asians have more propensity for carotid atherosclerosis than Europeans or the Chinese [3]. Prostate cancer is more common in men of African ancestry; this phenomenon is attributed to a variation in intron of the *ZNF652* gene, which is more prevalent in this population than others [22].

While we acknowledge that the disclosure of a health risk is much more sensitive, and that ethnic background is a more general result, we want to emphasize that it consequently also reveals health information.

# 5. GENOTYPING RESULTS DISCLOSURES

Our main goal was to establish the extent of genetic test result dissemination by Twitter users. In order to gather this, we crawled Twitter and searched for specific keywords. Analyzed tweets span the period of February, 2008 and March, 2014. We present an analysis and examples of such disclosures.

## 5.1 Methodology

Twitter does not enable the search of their full repository of tweets using the standard developer API; therefore we created a custom crawler able to access the older tweets. We used a PhantomJS browser, which performed a Twitter search and the retrieval of the tweets containing specific keywords. We focused on the tweets describing the results obtained from 23andMe and Ancestry.com.

Thus, we generally searched for tweets with the keyword "*23andMe*", as well as other keywords (selectors), related to the interesting information. For example we searched for "*23andMe heart*" to target the disclosures of heart-related disease risks. We used this keyword-based approach with 36 keywords during the retrieval of all the twitter-related data, and retrieved 4,904 *candidate tweets*, potentially containing genotyping results. During manual analysis, we also observed that Twitter users often include information about the detected risks and other findings.

After performing a search query and retrieving all the *candidate tweets*, we semi-manually selected the ones disclosing genotyping results. In this process, we automatically selected tweets likely discussing the risks disclosures and ethnicity background in terms of percentage numbers, and then manually screened the candidate tweets for verification. In some cases, users even posted screen captures of information from the 23andMe's Web site, which contained their genotyping results.

Aside from reporting the numbers of the observed disclosures (tweets relating to the health data), we also discuss examples of those communications.

The tweets we analyzed come from unique users (we only considered a single tweet originating from a given user). If, for example, a user has repeated the message or included a similar one, we considered it as an only one tweet of this user.

In summary, we retrieved data from Twitter using specific keywords, we then searched for potential disclosures in terms of percentage numbers, and we selected the relevant data – a step often requiring manual work. A schematic diagram displaying the work flow is shown in Figure 1.

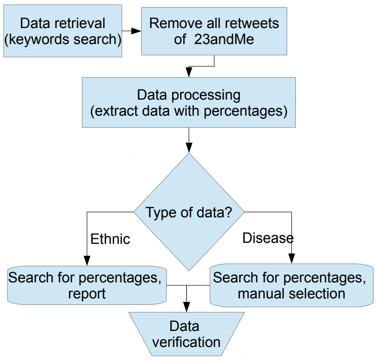


Figure 1: The methodology used throughout the analysis.

## 5.2 Ethnicity data disclosure

### 5.2.1 23andMe tweets

23andMe enables users to discover their detailed ethnic background. We used the keywords such as *Neanderthal, European, American, Scandinavian, African, Asian, unassigned, Oceanian, French, German, British* to retrieve tweets disclosing ethnicity. We retrieved 2004 of such tweets. We focused on the ones disclosing results in terms of percentage numbers; for example tweets such as "*i can confirm that i am a native american and from ibera/north africa. i'm also discovering about my genetic diseases*" were not considered (disclosure of ethnicity data, no percentage results). When a user just reports that he "*is American and sub-Saharan*" there is no way of ascertaining the actual extent of this ethnic background. We therefore assumed this was a disclosure of a minor significance. Instead, we focused on those users who actually disclosed detailed results *in terms of percentages*, for example, we considered informative tweets such as: "*I'm 83.0% African, 13.3% European, 0.6% East Asian and Native American, 0.4% South Asian, 2.6% unassigned*".

Out of the 2004 retrieved tweets, we found that 476 (23.8%) included data with one percentage value, and we manually verified that 98 (4.9%) included ethnicity information with two percentage values (e.g. "*i'm 60% spanish, 19.2% african*"). The remaining 1504 tweets were false positives such as "*I look forward to finding all my European, African & Asian relatives via @23andme and then visiting them*".

**Disclosing users are not anonymous.** We manually verified the profile pages of the 98 aforementioned Twitter users by visiting them. We established that in 95 (96.9%) of these profiles, identifying information was included either in the Twitter user names or Twitter descriptions. This suggests that users who are disclosing their genotyping results do not attempt to act anonymously, and in fact may be aware that they disclose the results of their tests publicly and openly, making it possible to connect them to their identities.

**Non-anonymous users are more popular.** We analyzed the profile pages of the 95 non-anonymous disclosing users. In case of each of these profiles, we retrieved the total numbers of tweets ($N_{tweets}$), followers ($N_{followers}$) and friends ($N_{friends}$), as well as the creation date of each profile. We then computed the difference in days between the profile creation date, and the date when we performed our test (5.06.14), i.e. how old were the profiles ($N_{days}$). For each profile, we computed the ratio of friends-to-followers ($R_{f2f} = \frac{N_{friends}}{N_{followers}}$). The median friends-to-followers ratio of our profiles was 0.44, which is about four times lower compared to a recent large scale study by Liu et al. [29], which reports a median ratio of 1.77 (for January 2012). This means that the non-anonymous users that disclosed their genetic ethnicity description are more pop-

| Ethnicity | 23andMe | Ancestry.com |
|---|---|---|
| Neanderthal | 62 | — |
| German | 5 | 2 |
| French | 5 | 2 |
| Scandinavian | 6 | 9 |
| American | 17 | 18 |
| British | 10 | 16 |
| African | 40 | 19 |
| Asian | 35 | 2 |
| European | 77 | 23 |

Table 2: Counts of tweets disclosing ethnicity results, for 23andMe and Ancestry.com.

ular than average Twitter users. Moreover, the average number of tweets per day ($R_{tpd} = \frac{N_{tweets}}{N_{days}}$) for these profiles was over 10. This means these users are relatively active, in addition to their apparent popularity.

### 5.2.2 *Ancestry.com tweets*

In the study of ethnicity result disclosures, we also considered tweets referring to the use of Ancestry.com's service, which provides data on ethnical background. We used similar methods as previously. For example, we searched for "*ancestry.com scandinavian*" in order to retrieve the potential tweets disclosing scandinavian ethnicities. We found 2,876 candidate tweets. In case of 110 of them, the disclosure mentioned just a one percentage value (e.g. "*i'm 9% (hey, you took 1 percent off!) native american.*"). We manually verified that 37 of them were disclosing genetic origin in terms of ethnicity description (two percentage values).

### 5.2.3 *23andMe and Ancestry.com: comparison*

We compared the number of times particular ethnicities were mentioned in the analyzed tweets. The results are in Table 2, which shows, for each company, the numbers of tweets mentioning an ethnicity for the tweets we verified manually (i.e. the ones containing at least two percentage values referring to a specific ethnicity). We counted the occurring keywords related to ethnicity in the candidate tweets (with percentages). For example a tweet "*@23andMe i am 2.7% neanderthal. i'm also 99.8% european*" contributed to an increase in the second column (rows for European and Neanderthal). In general, we found more tweets mentioning results obtained from 23andMe, than in the Ancestry.com's case, perhaps because 23andMe also offers their customers an insight into their genetic data and potential health risks.

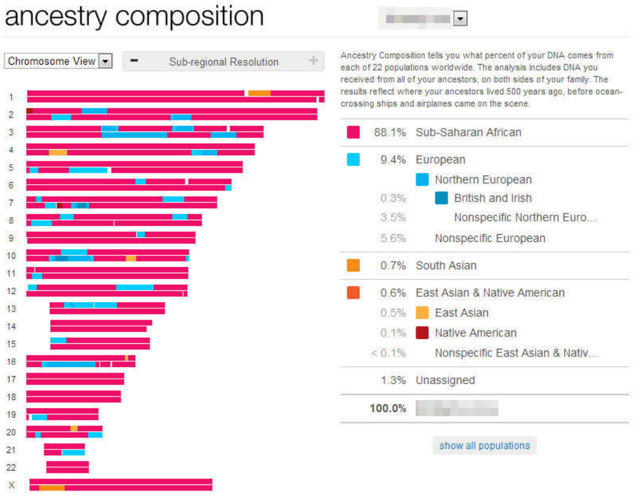
Generally, the most popular trait mentioned in all the tweets that contained at least one percentage number, was Neanderthal (324 tweets).

### 5.2.4 *Examples of ethnicity disclosure*

Below we show examples of analyzed messages.

- *23andme is now telling me i'm 'only' 99.7% European heritage. Surprise was i'm 0.1% Sub-Saharan African (on gene 7).#reallywhite*
- *i like 23andme's email updates. apparently i'm 99.8% European, 0.1% North African (neat!).*
- *finally got my @23andme results: 93.6% Ashkenazi Jewish (kind of duh), 6% nonspecific European, 0.2% Southern European, 0.2% unassigned*
- *results is as follows: my paternal line: haplogroup e1b1a my maternal line: haplogroup l1b1a7 1.4% neanderthal*
- *I'm less of a neanderthal (2.7%) than my parents (each 2.8%)!*
- *got my @ancestry dna results! 53% british isles, 34% scandinavian, 9% southern european, 4% uncertain.*
- *my @23andme: i'm 60% spanish, 19.2% african, 7.9% native american, 4.9% irish, 3.5% italian, 3% british, 1% jewish*

We also found some users that not only disclosed their hereditary details (e.g. *racial admixture on a chromosomal level*), but also posted full screen captures with detailed information like the user's full name, as seen in Figure 2.



ancestry composition

Ancestry Composition tells you what percent of your DNA comes from each of 22 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.

| | | |
|---|---|---|
| 🟥 | 88.1% | Sub-Saharan African |
| 🟦 | 9.4% | European |
| | | 🟦 Northern European |
| | 0.3% | 🟦 British and Irish |
| | 3.5% | Nonspecific Northern Euro... |
| | 5.6% | Nonspecific European |
| 🟧 | 0.7% | South Asian |
| 🟧 | 0.6% | East Asian & Native American |
| | 0.5% | 🟧 East Asian |
| | 0.1% | 🟥 Native American |
| | < 0.1% | Nonspecific East Asian & Nativ... |
| | 1.3% | Unassigned |
| | 100.0% | [redacted] |

Figure 2: Full screen capture of 23andMe ethnicity results posted by Twitter user. Note that both first and last names are included (blurred).

## 5.3 Disease risk disclosure

Using the same methodology of searching by keywords, we attempted to find if Twitter users disclose their diseases risks, obtained using 23andMe. The following keywords related to disease and internal organs were used: *prostate, obese, gout, glaucoma, stomach, schizophrenia, alcohol, coronary, cholesterol, attack, obesity, cancer, stroke, diabetes, heart, kidney, bowel, crohn, graves, parkinson, alzheimer, hirschsprung, meniere*. We tested explicitly for the keywords relating to disorders 23andMe is capable of reporting, choosing both prevalent (e.g. heart diseases) and rare conditions (e.g. Hirschsprung's disease). The motivation behind this keyword choice was an assumption that users might be more inclined to disclose "*popular*" conditions, as well as those "*rare*" they might have never heard before and think of them as being "*interesting*" and worth mentioning.

We first preprocessed the data as in the hereditary case (verification of disclosure in terms of percentage numbers) and obtained $2,900$ candidate tweets. We then manually verified all the candidate findings.

In this case, we did not encounter many disclosures related to the analyzed keywords. We saw numerous false positive tweets that we did not analyze (example for *crohn* keyword: "*Heritability of Crohn's disease is estimated to be 50-60%. This means that genetic factors contribute slightly more*"). Below we list some true positive examples of disclosed risks of developing a condition.

### 5.3.1 Diabetes

We want to emphasize that users are prone to disclose the actual risks related to diseases. For example some users disclosed that they have an increased risk of diabetes Type II (for 41%).

- *results back from #23andme and no great surprises - perhaps should lay off the buns though (chance of type 2 diabetes up 23%)*
- *23andme indicates that i have a 41% risk of developing type ii diabetes. feels like a much-needed kick up the arse.*
- *23andme.com says i have an elevated risk (36.1%) chance for type 2 diabetes. guess i should be better behaved with regard to sugar.*
- *increased risk of type 2 diabetes. 33.1% versus 25.7 average. makes sense.* ***my mother had it. other siblings too.***

We noticed that in the case of users who disclosed their health risks, some also noted that they might be forced to take precautions. In this sense, we can show that users actually understand and acknowledge the risks and suggest they will change their life habits.

### 5.3.2 Alzheimer Disease

- *got my gene results back from @23andme - 60% lower Alzheimer's disease risk than average, also hugely reduced chd risk. fascinating.*
- *got my @23andme results: 40% risk of Alzheimer's.*

In this case, users did not express having knowledge of lifestyle changes aiming to improve their situation, which contrasts with tweets mentioning other diseases (in particular, diabetes).

### 5.3.3 Cancer

- *discovering my genetic risk of prostate cancer (31.2%), and the hopes it's inaccurate (...)*
- *@23andme tells me i have a 33.9% risk of atrial fibrillation and 21.9% chance of prostate cancer - time for a check up.*

In these cases, we acknowledge the hopes for inaccurate result or the realization that testing one's health is important.

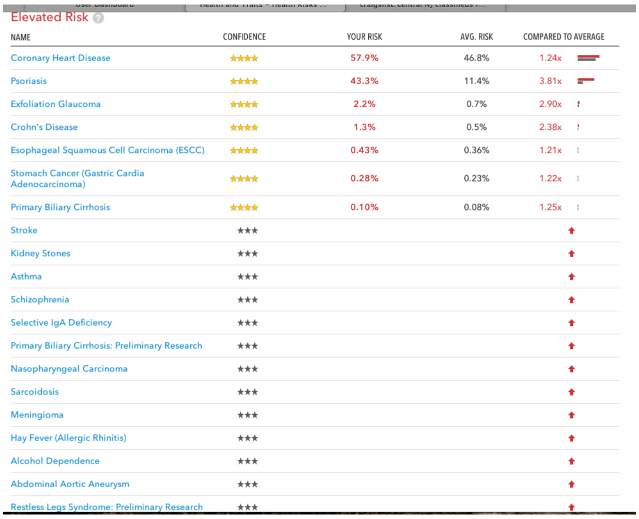
## 5.4 Other tweets with various content

In addition to previous records in this section, we encountered people posting different, but also sensitive, types of data.

- *wow, according to 23 and me john and my kids only have a 1% chance of brown eyes. #sowhite*
- *Coolest unexpected 23andMe result: I'm one of the 1% of Europeans resistant to the most common strain of HIV! Homozygous for Delta32 in CCR5*
- *23andme is freakily accurate. 4 star confidence that I have OCD + hypertension. Let's hope that 43% chance (usual 11%) Psoriasis is wrong.*
- *23andme says I'm prone to overeat and be lethargic*
- *23andme is telling me I have a 35.7% chance of getting Gout. So there's that to look forward to.*
- *Based on my genotype I have 59.6% chance of having coronary heart disease*

## 5.5 Discussion

We observed that Twitter users occasionally disclose their genetic ethnicity background (23.8% of analyzed tweets mentioning 23andMe and nationality such as "European" also include a percentage). Although according to a recent usability study, people are not expressing a specific inclination to the possibility of discovering their ethnic description [14], we detected more messages with this kind of information, compared to disclosing a risk of developing a condition or a disease. Perhaps when users receive two types of data (ethnicity, disease) and are faced with these seemingly different types of results, they conclude that ethnicity data is much less valuable or privacy-sensitive than information on disease risks. Therefore, they are more likely to disclose their ethnicity results, but keep the disease risks concealed. Similar reasoning can also apply to the disclosure of seemingly innocuous data like caffeine metabolism, to which the users might also assign lower significance (example: "*my mom got her 23andme dna results back! 11% french ancestry with genes for high caffeine metabolism.*"). Moreover, perhaps the users are more likely to understand the risks related with the disclosure of this data. However, we still found specific examples referring to a number of diseases. Some users even posted full screen shots of ethnicity data (Figure 2), as well as disease risks, seen in Figure 3.



Elevated Risk

| NAME | CONFIDENCE | YOUR RISK | AVG. RISK | COMPARED TO AVERAGE |
|---|---|---|---|---|
| Coronary Heart Disease | ★★★★ | 57.9% | 46.8% | 1.24x |
| Psoriasis | ★★★★ | 43.3% | 11.4% | 3.81x |
| Exfoliation Glaucoma | ★★★★ | 2.2% | 0.7% | 2.90x |
| Crohn's Disease | ★★★★ | 1.3% | 0.5% | 2.38x |
| Esophageal Squamous Cell Carcinoma (ESCC) | ★★★★ | 0.43% | 0.36% | 1.21x |
| Stomach Cancer (Gastric Cardia Adenocarcinoma) | ★★★★ | 0.28% | 0.23% | 1.22x |
| Primary Biliary Cirrhosis | ★★★★ | 0.10% | 0.08% | 1.25x |
| Stroke | ★★★ | | | ↑ |
| Kidney Stones | ★★★ | | | ↑ |
| Asthma | ★★★ | | | ↑ |
| Schizophrenia | ★★★ | | | ↑ |
| Selective IgA Deficiency | ★★★ | | | ↑ |
| Primary Biliary Cirrhosis: Preliminary Research | ★★★ | | | ↑ |
| Nasopharyngeal Carcinoma | ★★★ | | | ↑ |
| Sarcoidosis | ★★★ | | | ↑ |
| Meningioma | ★★★ | | | ↑ |
| Hay Fever (Allergic Rhinitis) | ★★★ | | | ↑ |
| Alcohol Dependence | ★★★ | | | ↑ |
| Abdominal Aortic Aneurysm | ★★★ | | | ↑ |
| Restless Legs Syndrome: Preliminary Research | ★★★ | | | ↑ |

Figure 3: Full screen capture of 23andMe disease risk results posted by a Twitter user.

### 5.6 Ethical considerations

In this study we analyzed the phenomenon of user-sharing of sensitive genotyping data. For the purpose of this analysis we crawled Twitter in search for relevant tweets. We did not use the standard Twitter API because it is limited in terms of past history: only relatively recent tweets can be searched. Therefore, using a custom solution was the only way to perform this study. We then retrieved the contents of the tweets for analysis.

We understand the sensitive nature of the topic. In order to ensure that user-identifying data is not abused, we anonymized all the occurrences of identities in the images and tweets that we included in this paper.

Furthermore, all tweets were securely stored on a computer which never front-faced the Internet, and they were removed immediately after the study (in order to enforce the right to modification/deletion).

## 6. CONCLUSION

Direct-to-consumer genetic testing has arrived to the market. Performing a test may be as easy as sending a saliva sample in the mail and then reading the results on a Web site. It is important to note that unlike other personal data, genetic information is hereditary and disclosing this information have implications for the user's relatives. Social networks enable unconstrained communication between people. However, sharing personal data carries certain specific risks. Although most of the social networks introduced detailed authorization models which can limit the disclosure of personal data to only specific sets of relatives, it is may not be enough.

We found a considerable number of Twitter users who decided to share their genetic information results, especially their ethnicity. Moreover, users are also sharing disease risks. This in turn may have consequences for them and their relatives; for example, insurance companies can take advantage of this data, raising insurance premiums or refusing coverage of certain customers.

This kind of genetic discrimination could potentially be difficult to fight as insurers could easily obtain this information on social networks.

**Acknowledgements.** We thank Dr. Katarzyna Koscielska for critically reading the manuscript and helpful comments.

## 7. REFERENCES

[1] 23andMe. 23ANDME, INC. PROVIDES UPDATE ON FDA REGULATORY REVIEW. `http://mediacenter.23andme.com/press-releases/23andme-inc-provides-update-on-fda-regulatory-review/`.

[2] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.

[3] S. S. Anand, S. Yusuf, V. Vuksan, S. Devanesen, K. K. Teo, P. A. Montague, L. Kelemen, C. Yi, E. Lonn, H. Gerstein, et al. Differences in risk factors, atherosclerosis, and cardiovascular disease between ethnic groups in canada: the study of health assessment and risk in ethnic groups (share). *The lancet*, 356(9226):279–284, 2000.

[4] L. B. Barreiro, G. Laval, H. Quach, E. Patin, and L. Quintana-Murci. Natural selection has driven population differentiation in modern humans. *Nature genetics*, 40(3):340–345, 2008.

[5] B. A. Bernhardt, C. Zayac, and R. E. Pyeritz. Why is genetic screening for autosomal dominant disorders underused in families? the case of

hereditary hemorrhagic telangiectasia. *Genetics in Medicine*, 13(9):812–820, 2011.

[6] S. R. Browning and B. L. Browning. Population structure can inflate snp-based heritability estimates. *American journal of human genetics*, 89(1):191, 2011.

[7] R. H. Carlson. *Biology is technology: the promise, peril, and new business of engineering life*, volume 2010. Harvard University Press Cambridge, 2010.

[8] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[9] M. K. Cho. Understanding incidental findings in the context of genetics and genomics. *The Journal of Law, Medicine & Ethics*, 36(2):280–285, 2008.

[10] . G. P. Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[11] . G. P. Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[12] C. Cooper. Hospital records used to 'target ads on twitter and facebook' say privacy campaigners, in latest nhs data concerns. `http://www.independent.co.uk/life-style/health-and-families/health-news/hospital-records-used-to-target-ads-on-twitter-and-facebook-say-privacy-campaigners-in-latest-nhs-data-concerns-9166633.html`.

[13] B. Cree, O. Khan, D. Bourdette, D. Goodin, J. Cohen, R. Marrie, D. Glidden, B. Weinstock-Guttman, D. Reich, N. Patterson, et al. Clinical characteristics of african americans vs caucasian americans with multiple sclerosis. *Neurology*, 63(11):2039–2045, 2004.

[14] E. De Cristofaro. An exploratory ethnographic study of issues and concerns with whole genome sequencing. *arXiv preprint arXiv:1306.4962*, 2013.

[15] K. Forrest, S. Simpson, B. Wilson, E. Van Teijlingen, L. McKee, N. Haites, and E. Matthews. To tell or not to tell: barriers and facilitators in family communication about genetic risk. *Clinical genetics*, 64(4):317–326, 2003.

[16] L. E. Forrest, M. B. Delatycki, L. Skene, and M. Aitken. Communicating genetic information in families–a review of guidelines and position papers. *European Journal of Human Genetics*, 15(6):612–618, 2007.

[17] L. N. Geller, J. S. Alper, P. R. Billings, C. I. Barash, J. Beckwith, and M. R. Natowicz. Individual, family, and societal dimensions of genetic discrimination: a case study analysis. *Science and Engineering Ethics*, 2(1):71–88, 1996.

[18] J. Gilbert. 'need a debit card' twitter account proves infinite stupidity of humans. `http://www.huffingtonpost.com/2012/07/03/need-a-debit-card-twitter_n_1645892.html`.

[19] R. C. Green, L. A. Cupples, R. Go, K. S. Benke, T. Edeki, P. A. Griffith, M. Williams, Y. Hipps, N. Graff-Radford, D. Bachman, et al. Risk of dementia among white and african american relatives of patients with alzheimer disease. *Jama*, 287(3):329–336, 2002.

[20] A. Greenberg. Yes, people actually post pictures of their credit cards online. this twitter account was created to shame them. `http://www.forbes.com/sites/andygreenberg/2012/07/03/yes-people-actually-post-pictures-of-their-credit-cards-online-this-twitter-account-was-created-to-shame-them/`.

[21] W. W. Grody, B. H. Thompson, and L. Hudgins. Whole-exome/genome sequencing and genomics. *Pediatrics*, 132(Supplement 3):S211–S215, 2013.

[22] C. A. Haiman, G. K. Chen, W. J. Blot, S. S. Strom, S. I. Berndt, R. A. Kittles, B. A. Rybicki, W. B. Isaacs, S. A. Ingles, J. L. Stanford, et al. Genome-wide association study of prostate cancer in men of african ancestry identifies a susceptibility locus at 17q21. *Nature genetics*, 43(6):570–573, 2011.

[23] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1141–1152. ACM, 2013.

[24] Illumina. Population power. extreme throughput. $1,000 human genome. `http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn`.

[25] Y. Joly, H. Burton, B. M. Knoppers, I. N. Feze, T. Dent, N. Pashayan, S. Chowdhury, W. Foulkes, A. Hall, P. Hamet, et al. Life insurance: genomic stratification and risk classification. *European Journal of Human Genetics*, 2013.

[26] L. Kharouni. The dangers of posting credit cards, ids on instagram and twitter. `http://blog.trendmicro.com/trendlabs-security-intelligence/the-dangers-of-posting-credit-cards-ids-on-instagram-and-twitter/`.

[27] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[28] E. V. Lapham, C. Kozma, and J. O. Weiss.

Genetic discrimination: perspectives of consumers. *Science*, 274(5287):621–624, 1996.

[29] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. `http://www.ccs.neu.edu/home/amislove/publications/Profiles-ICWSM.pdf`, 2014.

[30] S. Malenkovich. Posting photos of your debit card... is a terrible idea. `https://blog.kaspersky.com/the-next-time-you-feel-like-posting-a-picture-of-your-debit-or-credit-card-dont/`.

[31] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: An analysis of privacy leaks on twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, pages 1–12, New York, NY, USA, 2011. ACM.

[32] C. Ober and D. Vercelli. Gene–environment interactions in human disease: nuisance or opportunity? *Trends in genetics*, 27(3):107–115, 2011.

[33] S. L. Pulit, M. Leusink, A. Menelaou, and P. I. de Bakker. Association claims in the sequencing era. *Genes*, 5(1):196–213, 2014.

[34] T. Strachan and A. Read. *Human molecular genetics*. Number Ed. 2. Wiley-Liss; New York, 1996.

[35] J. H. Tanne. Us senate outlaws genetic discrimination. *BMJ: British Medical Journal*, 336(7652):1038, 2008.

[36] A. Vance. Illumina's dna supercomputer ushers in the $1,000 human genome. `http://www.businessweek.com/articles/2014-01-14/illuminas-dna-supercomputer-ushers-in-the-1-000-human-genome`.

[37] M. Wadman. Genetics bill cruises through senate. 2008.

[38] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "i regretted the minute i pressed share": A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.