ANALYSIS &
INFERENCE
215 Haverford Avenue
Swarthmore, PA 19081
610-543-0159

Report Concerning Data in Two Leaked Copies of Federal Government Watch Lists

MUTASEM JARDANEH, et al., Plaintiffs,
v. MERRICK B. GARLAND, et al. - Case 8:18-cv-02415

Prepared for Counsel for CAIR

by

William A. Huber, Ph.D., PSTAT®

March 22, 2023

Contents

Figures

Tables

1 Introduction

¶1 In January 2023, files named "nofly.csv" and "selectee.csv" were discovered on a server maintained by the US airline CommuteAir, Thalen and Covucci (2023). These appear to be versions of the Transportation Security Administration's (TSA) "watch lists" from 2019:

> *The server contained data from a 2019 version of the federal no-fly list that included first and last names and dates of birth," CommuteAir Corporate Communications Manager Erik Kane said.*

Cushing (2023).

¶2 Counsel for the Council on American-Islamic Relations (CAIR) requested my help in

a. summarizing the entries in these two files,

b. designing a rigorous program to sample entries and classify them according to their likely religious affiliation, and

c. analyzing the classifications to assess the proportions of entries in each file (and overall) associated with Muslims.

2 Muslim representation on the watch lists

¶3 By "watch lists" I mean the combined entries in the "nofly.csv" and "selectee.csv" files, called the "No-Fly" and "Selectee" lists, respectively. Appendix A to this declaration describes these files and analyzes their contents.

¶4 The records in these watch lists contain (apart from a small number of exceptions) only names and dates of birth. However, Appendix A presents strong evidence that many individuals are referenced more than once in the watch lists. These lists should therefore be considered collections of *aliases*. This perspective is consistent with their principal use in determining from an individual's identification papers whether there is any corresponding watch list record for them.

¶5 The watch lists do not explicitly indicate nationality or religious affiliation. To estimate the proportions of Muslims on the watch lists, I conducted a statistical study in which the watch lists were randomly sampled and a panel of CAIR evaluators classified the names in the records according to whether they thought they corresponded to a Muslim. The panel was also aware that, to verify the accuracy of their classifications, some additional records of names with known religious affiliations had been randomly spread through the list of records they evaluated, but they did not know how many such records there were.

¶6 This study provides statistical estimates of the proportions of *records* in the watch lists that correspond to Muslims. Table 4 in Appendix B gives the results. It reports the best statistical estimate for the proportion of such records in the No-Fly list, the Selectee list, and the combined list. It also gives 90% credible intervals for those estimates. Credible intervals indicate how each estimate is affected by variation due to random sampling. In each case, there is only a 5%

chance that the true proportion of Muslim records in the corresponding watch list is less than the lower limit of the credible interval and only a 5% chance that the true proportion is greater than the upper limit.

¶7 **The main result is there is a 95% chance that at least 98.3% of the records in the combined watch lists correspond to Muslims.**

¶8 The information in these files does not indicate how many *individuals* are named in the watch lists, because many individuals appear more than once. The estimates of the proportion of Muslim *records* do not necessarily translate to comparable estimates of the proportion of Muslim *people* on the watch lists. However, the extremely high estimated proportion of Muslim records suggests a high proportion of people on the watch lists are Muslims. (The exploratory statistical study described in Appendix B at ¶35 bolsters this conclusion.)

¶9 Names on the watch lists are sequences of *tokens* (sequences of letters) representing given and family names and, perhaps, sometimes with titles or honorifics. Nearly seven million tokens appear in the combined watch lists. To illustrate the typical names in the watch lists, Table 2 in Appendix A lists the 50 most frequent tokens and the number of times they appear.

3 Materials Used

Nofly.csv (a 77,029 KB text file of 1,566,062 lines). Obtained from CAIR attorneys.

Selectee.csv (a 11.788 KB text file of 251,169 lines). Obtained from CAIR attorneys.

Carleton College Office of the Chaplain. “Student Religious Groups and Contact Information.” October 21, 2022. https://www.carleton.edu/chaplain/students/groups/.

Cushing, Tim (2023). “US Airline Leaves No Fly List Details Accessible On The Open Web.” January 23, 2023 at https://www.techdirt.com/2023/01/23/us-airline-leaves-no-fly-list-details-accessible-on-the-open-web/.

Hashmi, Heraa (2017). “Worldwide Muslims Condemn List.” July 14, 2017. https://data.world/sya/worldwide-muslims-condemn-list.

Mahdawi, Arwa (2017). “The 712-page Google doc that proves Muslims do condemn terrorism.” The Guardian, March 26, 2017. https://www.theguardian.com/world/shortcuts/2017/mar/26/muslims-condemn-terrorism-stats.

National Archives (2007). “The Soundex Indexing System.” May 30, 2007. https://www.archives.gov/research/census/soundex.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

The Stan Development Team Stan Modeling Language User's Guide and Reference Manual. https://mc-stan.org/.

Thalen, Mikael and David Covucci (2023). "U.S. airline accidentally exposes 'No Fly List' on unsecured server." Posted Jan. 19, 2023 at https://www.dailydot.com/debug/no-fly-list-us-tsa-unprotected-server-commuteair/.

United States Census Bureau (n.d.). "Soundex". https://www.census.gov/history/www/genealogy/decennial_census_records/soundex_1.html

Wikipedia (n.d.) "List of Catholic bishops in the United States." Accessed March 1, 2023. https://www.wikiwand.com/en/List_of_Catholic_bishops_in_the_United_States.

4 Appendix A: What is in the watch list files

¶10 The contents of NoFly.csv are the “No-Fly list” and the contents of Selectee.csv are the “Selectee list”.

¶11 Both files begin with header lines naming the data fields: "SID", "CLEARED", "LASTNAME", "FIRSTNAME", "MIDDLENAME", "TYPE", "DOB", "POB", "CITIZENSHIP", "PASSPORT.IDNUMBER", "MISC".

¶12 In NoFly.csv,

a. “CLEARED”, “MIDDLENAME”, “TYPE”, “POB”, and “MISC” are entirely empty.

b. “CITIZENSHIP” is non-blank on only 1791 records. Its non-blank values are “AF” (81 records), “PK” (1637 records), and “TH” (73 records).

c. “PASSPORT.IDNUMBER” is non-blank on only 75 records. Its values are “1996201”, “A624665”, “PP3391495” (one record each) and “Z025502” (72 records).

¶13 In Selectee.csv, “CLEARED”, “MIDDLENAME”, “TYPE”, “POB”, “CITIZENSHIP”, “PASSPORT.IDNUMBER”, and “MISC” are entirely empty.

¶14 Consequently, the vast majority of the records in both lists have non-empty values only for “SID”, “LASTNAME”, “FIRSTNAME”, and “DOB”.

¶15 There is little overlap in records between the lists: in the combined list, only 508 records (less than 0.03%) have the same values for DOB, FIRSTNAME, and LASTNAME.

¶16 None of these fields explicitly indicates religious or national affiliation.

4.1 Contents of the SID field

¶17 This field contains whole numbers ranging from 1 through 99900029. Only 72 of them exceed 11191355 and these range from 99700012 to 99900029.

¶18 In the combined NoFly.csv and Selectee.csv data, there are no duplicated values of SID. **These SID numbers thereby serve as unique identifiers of the records.**

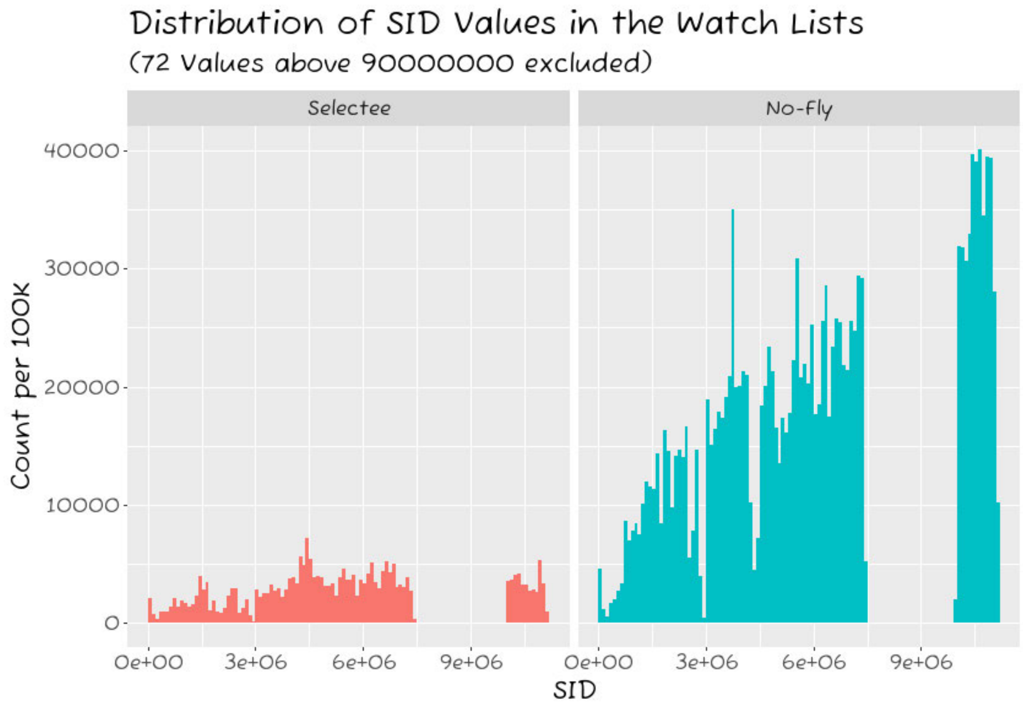
¶19 **When the combined data are sorted by SID, many sequences of contiguous records appear to refer to single individuals, varying in the construction and spelling of their names and their date of birth.** Table 1 displays the first such example, showing 11 records with similar names and dates of birth.

Table 1 Example of consecutive records (by SID) in the combined lists

SID	LASTNAME	FIRSTNAME	DOB
113	[redacted]	[redacted]	[redacted]
115	[redacted]	[redacted]	[redacted]
117	[redacted]	[redacted]	[redacted]
119	[redacted]	[redacted]	[redacted]
121	[redacted]	[redacted]	[redacted]
124	[redacted]	[redacted]	[redacted]
126	[redacted]	[redacted]	[redacted]
128	[redacted]	[redacted]	[redacted]
130	[redacted]	[redacted]	[redacted]
132	[redacted]	[redacted]	[redacted]
134	[redacted]	[redacted]	[redacted]

¶20 Table 1 also reveals there are gaps in the sequence of SIDs. Figure 1, a histogram of the SID values, shows that the frequency of such gaps diminishes with increasing SID. Initially there are only a few thousand SIDs out of the first hundred thousand possible numbers. For instance, among the possible values from 1 through 100,000, there are only 2095 SIDs in the Selectee list and only 4582 SIDs in the No-Fly list, for a total proportion of only 6.677% of the first 100,000 numbers. At the other extreme, among the possible numbers from 10,900,001 through 11,000,000, the proportion of SIDs is 44.682%.

Figure 1 Distribution of SID values in the watch lists

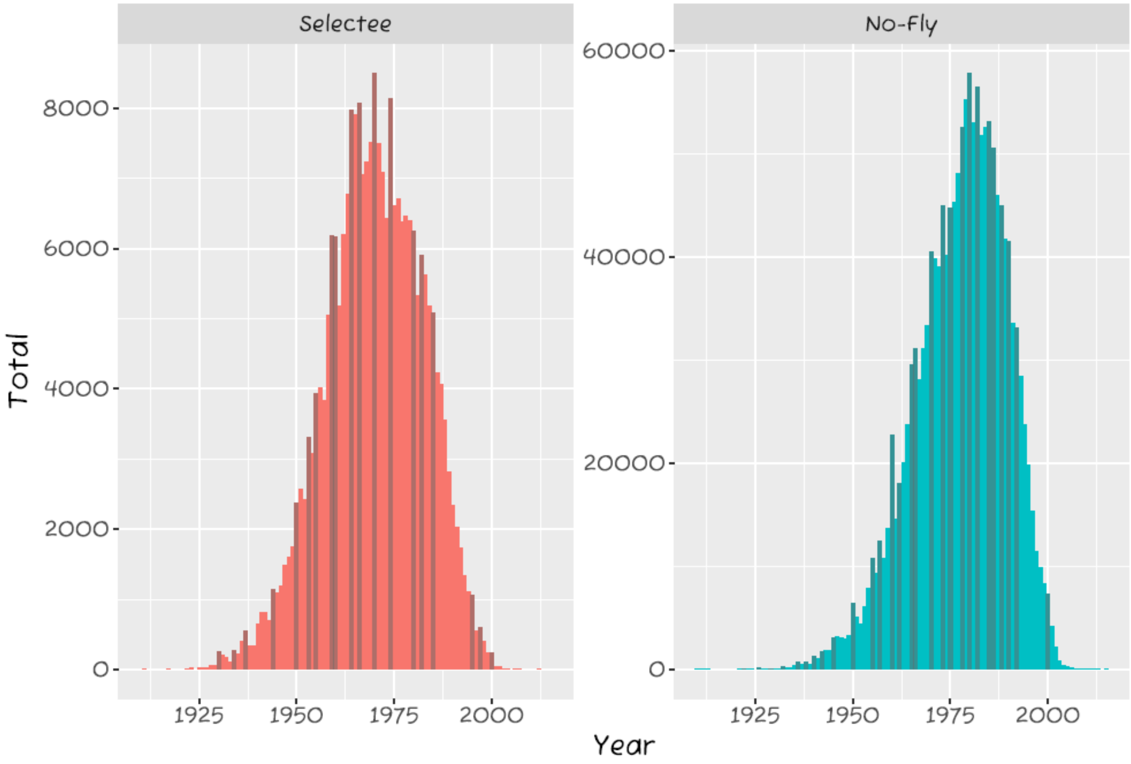


¶21 Figure 1 also reveals two large gaps in the SID sequence. There are no SIDs from 7,413,688 through 9,976,829, a span of over 2.5 million numbers. There are no SIDs from 11,191,356 through 99,700,011, a span of over 88 million numbers.

4.2 Contents of the DOB field

¶22 The values of this field are all valid dates ranging from 1910 through 2015. Figure 2 shows their distributions by list. The name of this field and these distributions indicate this is a Date Of Birth value.

Figure 2 Histograms of DOB values in the watch lists



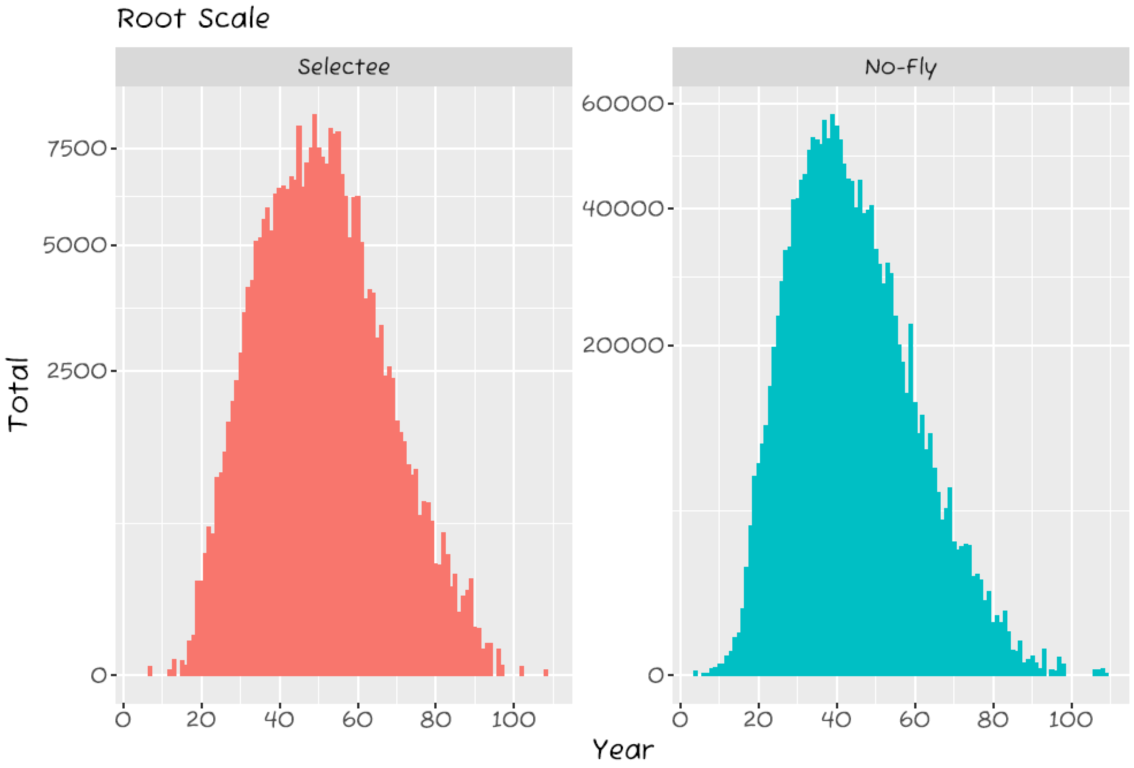
¶23 Some birth years are unusually frequent. These are darkened in Figure 2. Although many of these years are multiples of five (*e.g.*, 2000, 1995, 1990, *etc.*), not all of them are. For example, the No-Fly list contains no entries giving 1925 as the birth year, 128 entries for 1926, and only nine entries for 1927.

¶24 The ages of the watch list entries at the end of 2019 can be inferred by subtracting the birth years from 2020. Figure 3 shows their distributions (using one bar for each year of age). To reveal details of the very young and very old, it uses a square root scale for the counts by year of age. It shows that **some minors and centenarians are on the watch lists:**

a. 1,679 entries (0.09%) have DOB values for people who were under the age of 18 at the end of 2019.

b. 28 entries (0.0015%) have DOB values for approximately six individuals 101 years or older. The youngest was under five years old (three records with distinct names indicate a birth year of 2015) and the oldest was over 109 years old (born in 1910).

Figure 3 Inferred ages in the watch lists in 2019



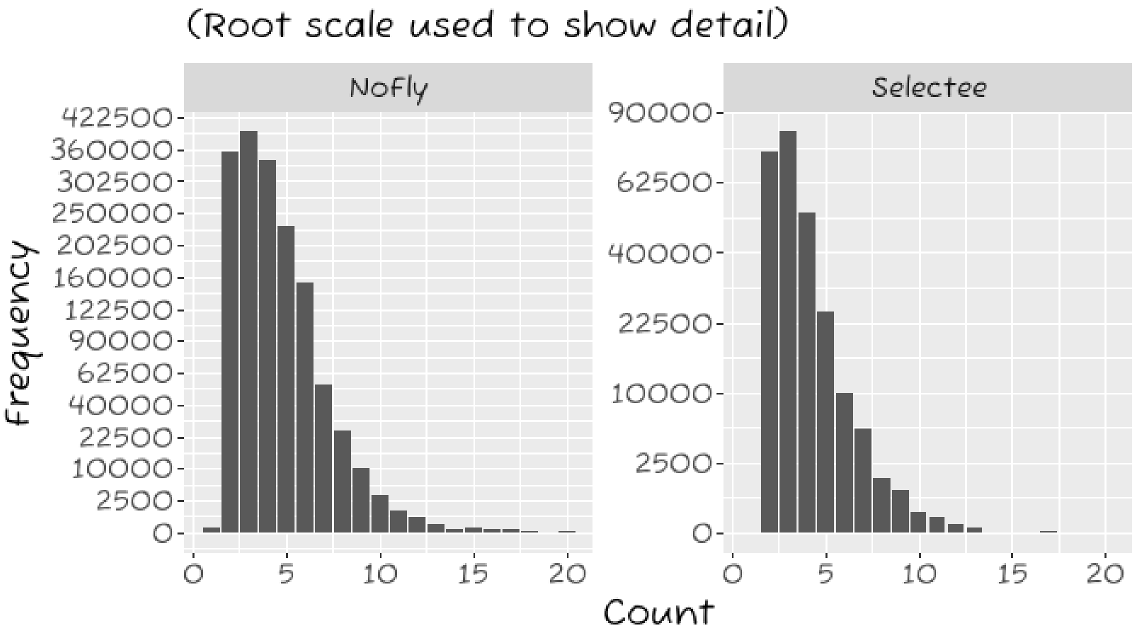
¶25 The birthdays are unevenly distributed: 23% of them are in January, of which more than half are January 1. 21% of the birthdays occur on the first of a month. Early days (2, 3, 4, 5, 6, 7) and some other days (11, 12, 20) are unusually frequent, too.

¶26 **The uneven distribution of birthdays indicates many of the birth dates are guesses.** It is evident that even the year of birth is a guess. The guesses can range over many years or decades. For instance, [redacted] (the value of FIRSTNAME; LASTNAME is blank) appears on three records with DOB given variously as [redacted] and [redacted] Many instances of birthdays ranging over multiple years for a given name or set of similar names can be found. See Table 3 below for another example of varying birthdates for what appears to be the same individual.

4.3 Contents of the FIRSTNAME and LASTNAME fields

¶27 These fields are sequences of letters – “tokens” – separated by blanks, as shown in Table 1 above[1]. The total number of tokens in both fields typically is three, but ranges from one through twenty as summarized by the frequency chart in Figure 4 below. For example, four No-Fly records have [redacted] for the value of FIRSTNAME and a blank for the value of LASTNAME, a total of one token. At the other extreme, the No-Fly list has four entries with [redacted] for the value of FIRSTNAME and [redacted] for the value of LASTNAME: 20 tokens altogether.

Figure 4 Token frequencies in the watch lists (FIRSTNAME and LASTNAME combined)



¶28 To illustrate typical values of tokens, **Table 2 lists the 50 most frequently appearing tokens** (excluding one-letter tokens, which are likely initials, and two-letter tokens, which tend to be small particles like “EL” and “AL”). The frequencies are the total number of occurrences: some tokens may appear more than once in a watch list record.

[1] Most tokens are names, but some, notably “SHAYKH”, may be honorifics.

Table 2 The 50 most frequently appearing tokens of three or more letters in the watch lists

Token	Frequency	Token	Frequency	Token	Frequency
MUHAMMAD	**200656**	RAHMAN	34037	ABDEL	16002
ALI	144044	SALIH	32914	KHAN	15785
ABD	126177	**MAHMUD**	**31901**	ISMAIL	15640
ABU	120005	HASSAN	29010	KARIM	14601
AHMAD	109844	HAMID	27882	HAMAD	14497
BIN	67270	KHALID	26373	KHALIL	14264
MOHAMMED	**59639**	SAID	25535	**MAHMOUD**	**13399**
ABDALLAH	58109	UMAR	21957	SULAYMAN	13248
IBRAHIM	57481	AZIZ	21297	SALAH	13012
AHMED	47673	SALIM	20723	DIN	12760
ABDUL	46729	MUSTAFA	19928	SHAYKH	12527
MOHAMMAD	**44465**	JASIM	18825	SALEH	12492
MOHAMED	**42050**	ABBAS	17824	SAAD	11987
BEN	36403	OMAR	17783	HADI	11925
HASAN	36290	JUBURI	17654	NASIR	11743
ABDULLAH	35978	KHALAF	16984	SALMAN	11505
HUSAYN	35898	HUSSEIN	16808		

¶29 **Names do not identify records.** There are only 913,656 unique combinations of FIRSTNAME and LASTNAME among the 1,817,231 total records, for an average of two records per full name.

4.4 How the records might be related to individual people

¶30 **There is no discernible way to identify, with any reliability, when two records refer to the same person.** Although many sets of records *appear* to correspond to a common individual, they vary in many ways, including:

a. How the name is split between the FIRSTNAME and LASTNAME field, such as (FIRSTNAME, LASTNAME) values of [redacted] or [redacted]).

b. The order of the tokens in the names, such as ("[redacted]").

c. The spellings of the tokens. For instance, there are 1,416 versions of "MUHAMMAD" as identified by the Census Bureau's "Soundex" representation of names, US Census Bureau (n.d.) and National Archives (2007). (Soundex is a brief code that captures the initial letter of a name and up to three of the following consonants classified within six phonetically similar groups.) The most frequent spellings are "MUHAMMAD", "MOHAMMED", "MOHAMMAD", "MOHAMED", and so on. (Table 2 above highlights the six most frequent versions of this name.) Some of these names are variants, such as "MEHMET". Some of the variant spellings appear to be truly different names, such as "MONTE", but these are infrequent.

d. The date of birth, as described in section 4.2 above.

e. The SID. Although the SID is a unique record identifier, there are numerous instances where sequences of closely related names, with the same or similar birthdate, are associated with a contiguous sequence of SIDs, as mentioned in ¶19 above. However, this is not invariably the case. Table 3 below provides an extreme example showing 110 watch list entries (all from the No-Fly list) that appear to refer to one individual, as determined from similar birthdates in 1926 and similar pronunciations of the tokens. In that table, “Record” is the rank of each entry when the combined watch lists are sorted by increasing SID. Thus, for instance, the nine entries with SID ranging from 775978 through 775996 were evidently made consecutively because they are records 31760 through 31768, with no gaps. Note that they do not all have the same DOB value, but that all values of DOB are either “9/9/1926” or “1/1/1926”. Indeed, many of these entries duplicate both LASTNAME and FIRSTNAME and differ only by DOB. *It is possible to identify these entries with the same individual only because of the rare birthdate and the relatively consistent use of just two values of DOB.*

¶31 At the other extreme, many individuals appear to be represented by unique records in the watch lists. Some evidence of this is supplied by grouping all watch list records by codes for the decade of DOB and the Soundex values of their FIRSTNAME and LASTNAME (regardless of order). Out of the 420,795 unique combinations of these decades and Soundex versions of the full name, 172,882 (41%) appear only once. An example is [redacted] for which no close variant can be found.

¶32 Even this analysis is approximate, because (a) many clearly related names can resolve to different Soundex codes (*e.g.*, “[redacted]” has a Soundex of “a426 y210” while [redacted]” has a different Soundex of “k631 y210”) and (b) many clearly related names have slightly different birthdates that fall into different decades.

¶33 Consequently, **it is not possible to determine how many individuals are represented in these watch lists**, or even to provide a reliable range for that number.

Table 3 Example of multiple watch list records for (apparently) one individual

Record	SID	DOB	FIRSTNAME	LASTNAME
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Record	SID	DOB	FIRSTNAME	LASTNAME
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Record	SID	DOB	FIRSTNAME	LASTNAME
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

5 Appendix B: A statistical sampling study of religious affiliation in the watch lists

5.1 Design of the study

¶34 To assess religious affiliation in the watch lists, I carried out a two-phase statistical study of the combined watch list entries. In this study, full names obtained from the watch lists (constructed by concatenating the FIRSTNAME and LASTNAME fields) along with names of people with known religious affiliations were randomly presented to a panel of three evaluators who were blinded concerning the affiliation and asked to determine it from the names alone.

¶35 Phase I was exploratory. I attempted to identify groups of entries likely to correspond to an individual and drew a random sample of 30 such groups, comprising 71 watch list records. Informal evaluation of those records by CAIR suggested 29 of those individuals likely were Muslims and one was not. Provisionally assuming those evaluations were accurate and that I correctly grouped the watch list records by individuals, **I concluded with 95% confidence** (using standard statistical procedures for Binomial random samples) **that between 82% and 99.9% of all individuals on the watch lists are Muslims.**

¶36 The objectives of Phase II were to refine this result and to measure the classification error, because there is no certain connection between a name and religious affiliation. The Phase I result and further study of the contents of the watch lists, as recounted in Appendix A, indicated it is not possible to group watch list records by individuals with any reliability. (Such grouping could be achieved if individual identifiers, such as nationalities and passport numbers, were available, but they appear to have been stripped out of these files: see ¶13.) I therefore changed the objective of Phase II to *estimating the proportions of watch list entries likely to refer to Muslims*.

¶37 Because CAIR wished to obtain a precise result with a high degree of confidence (at least 95%), I designed Phase II to use FIRSTNAME and LASTNAME fields from a sample of watch list records, of which approximately 10% would be dedicated to assessing the accuracy with which CAIR is able to determine whether a watch list entry references a Muslim.

¶38 Assessing accuracy is a matter of estimating two kinds of errors: a "false positive" in which a non-Muslim is identified as Muslim and a "false negative" in which a Muslim is not identified as Muslim. To this end I located lists of names of people with known religious affiliations to create "reference lists:"

a. Names of known Muslims were obtained from the "Worldwide Muslims Condemn List," Hashmi (2017) and Mahdawi (2017). It is a list of Muslims and Muslim organizations who have condemned terrorist acts. This list contains duplicates, names of organizations, and some revealing titles such as "Professor," "Foreign Minister", "Prince", and so on. I systematically removed all duplicates, all names of organizations, and erased all titles and honorifics, yielding a list of 3116 names.

b. Names of Catholic bishops emeriti and eparchs[2] were similarly formatted, yielding a list of 188 names; and names of leaders of student religious groups at a US college yielded a list of 26 names, of whom 4 are Muslim leaders and the other 22 are leaders of non-Muslim groups.

c. The combination of these lists provided 210 names of non-Muslims and 3120 Muslims.

¶39 I randomly sampled 146 of the Muslim names and 152 of the non-Muslim names (without replacement) from the reference lists. These quantities, 146 and 152, were randomly generated from a distribution averaging a count of 150, thereby totaling around 300, approximately 10% of the planned total of 3000 names.

¶40 I randomly sampled 3000 – (146 + 152) = 2702 entries from the combined watch lists.

¶41 The 3000 sampled names were cast into the same full name format (all capitalized and stripped of most punctuation), randomly ordered, and saved as a spreadsheet.

¶42 I provided the spreadsheet to the panel along with instructions for evaluating the names in it, reproduced *verbatim* and in full in Appendix B1 below. The instructions revealed that some non-watch list names would be included as a check of evaluation accuracy, but they did not indicate how many such names there would be, nor how they were obtained.

5.2 Results

¶43 Table 4 (below) presents the raw counts and proportions, uncorrected for the classification error rates. The "raw proportion" column is the fraction of records in each sample *classified as* Muslim. The "estimate" column estimates the fraction of Muslims in each list. To not over-estimate the Muslim representation, the few records not explicitly classified as Muslim by the panel were counted as non-Muslim. (See Appendix B1 below.)

Table 4 Raw and Corrected Estimates

List	Sample size	Classified Muslim	Raw proportion	Estimate (posterior mode)	90% Credible interval
Both	2702	2530	93.6%	98.9%	98.3 – 99.3%
No-Fly	2330	2211	94.9%	99.9%	99.6 – 99.99%
Selectee	372	319	85.8%	92.4%	88.6 – 95.4%

[2] An eparch rules an eparchy, which is an ecclesiastical unit of Eastern Christianity. Although the eparchs on this list rule US eparchies, many have names characteristic of Eastern Europe and the Middle East, like Yousif Habash, Ibrahim Ibrahim, and Abdallah Elias Zaidan.

List	Sample size	Classified correctly	Raw proportion	Estimate (posterior mode)	90% Credible interval
Muslim[3]	146	135	92.5%	92.5%	87.8 – 95.2%
Non-Muslim	152	145	95.4%	95.4%	91.6 – 97.3%

¶44 **The raw results require correction.** To appreciate the need for a correction, notice that the classification error rates are low, because the rates of correct classification in the known lists are both above 90%. The high raw proportions of records *classified* as Muslim on the watch lists indicate there is likely a high proportion of *true* Muslims in the sample (and, accordingly, in the watch lists themselves). That large proportion of true Muslims will result in a relatively large number of records mis-classified as non-Muslim, whereas the relatively small proportion of true non-Muslims will result in relatively few records mis-classified as Muslim. Consequently, *a great majority of the records classified as non-Muslim are likely errors.*

¶45 A Bayesian analysis[4], implemented on the **Stan** statistical computing platform, was used to correct the results. This analysis estimates four numbers, or "parameters:"

a. p_1 is the proportion of Muslim records in the No-Fly list.

b. p_2 is the proportion of Muslim records in the Selectee list.

c. p_+ (referred to as "pp" by the **Stan** program, Appendix C) models the rate at which Muslims are correctly classified. The *observed* rate of 92.5% is subject to random sampling variation and so estimates this parameter with a little uncertainty.

d. p_- (referred to as "pm" by the software) models the rate at which non-Muslims are correctly classified. The *observed* rate of 95.4% is subject to random sampling variation and so estimates this parameter with a little uncertainty.

¶46 The proportion of Muslim records in the combined watch list can be estimated from the weighted average of the proportions in each list, $q = (1566062\, p_1 + 251169\, p_2)/(1566062 + 251169)$. The coefficients are the numbers of records in each list.

¶47 The result of a Bayesian analysis is a (posterior) *probability distribution* for each estimated parameter. The mode of a distribution is the most likely value of the parameter, given the data. The spread of the distribution provides a *credible interval* within which the parameter value is likely to lie, with any specified degree of credibility. Figure 5 displays histograms of the

[3] The Muslim and Non-Muslim estimates are straightforward and need no correction. These estimates and their credible intervals can be computed with simple formulas not requiring the Stan software. The credible intervals given here are those computed with Stan in the course of obtaining the corrected estimates of *p* for the watch lists. They agree with the formulas to an accuracy of ±0.1%, serving as a verification of the accuracy of the Stan calculations overall.

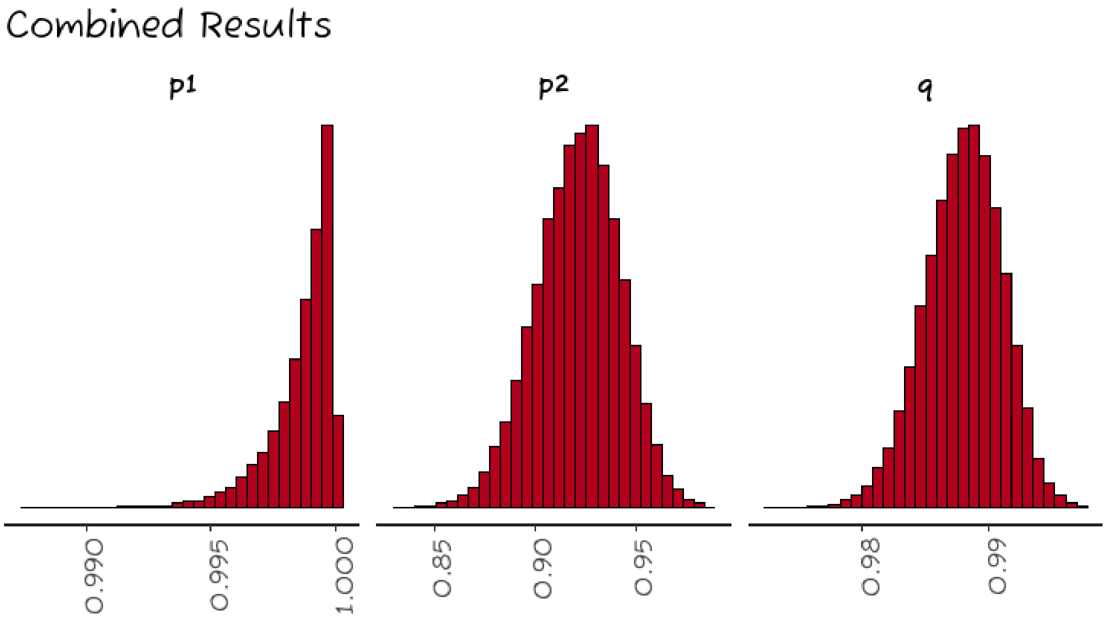
[4] Such analyses require assumed "prior distributions" of the parameters. With the large sample sizes employed here (from each watch list and the reference lists), the results are insensitive to these assumed prior distributions.

distributions of p_1, p_2, and q along with schematics of the symmetric 90% credible intervals for all three parameters in the No-Fly and Selectee lists. The value of 90% was chosen because there is only a (100 – 90)/2 = 5% chance that the true parameter value is *smaller* than the left endpoint of the interval.

¶48 Consequently, **this study estimates there is a 95% chance that at least 99.6% of the records in the No-Fly list identify Muslims and there is a 95% chance that at least 88.6% of the records in the Selectee list identify Muslims[5]. Combined, these results give a 95% chance that at least 98.3% of all records in the combined watch lists identify Muslims.**

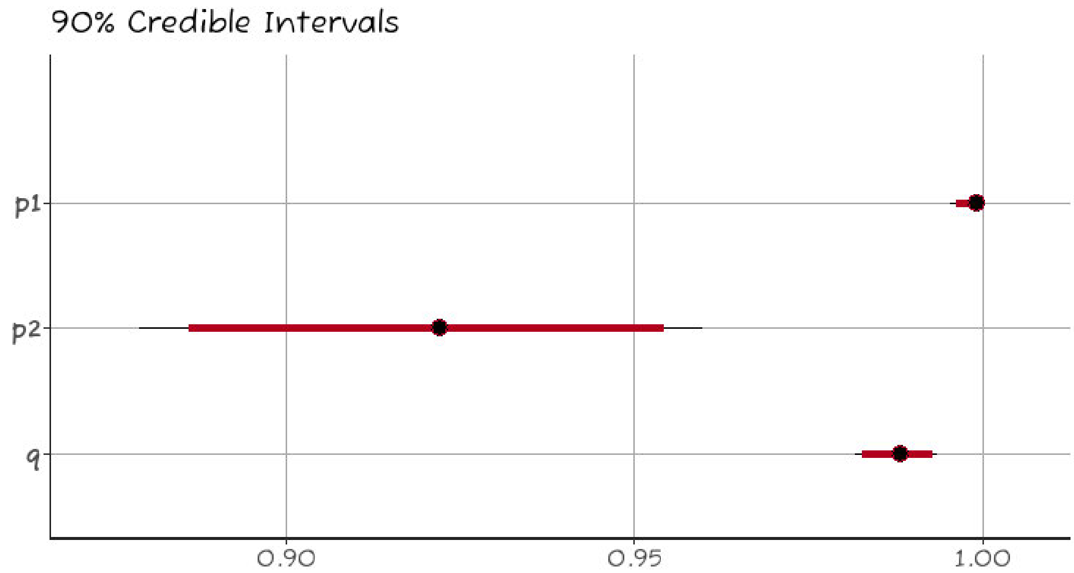
Figure 5 Posterior distributions of the fraction of Muslims on the No-Fly and Selectee lists

"p1" is the fraction of Muslim records in the No-Fly list, "p2" is the fraction of Muslim records in the Selectee list, and "q" is the fraction of Muslim records in the combined lists. The bar heights indicate relative probabilities for the true values in the watch lists. Thus, the location of the highest bar in each histogram is the most likely value (see Table 4 *above).*



[5] These chances are not statistically independent, because they both rely on the same classification error rate estimates. This correlation causes the combined chance and its credible intervals to be similar to that of the No-Fly list. The correlations between q and each of the p_i are both close to 80%. They in turn induce a correlation of 33% between p_1 and p_2.

"p1" is the fraction of Muslim records in the No-Fly list, "p2" is the fraction of Muslim records in the Selectee list, and "q" is the fraction of Muslim records in the combined lists. There is a 90% chance that the correct value lies within the thick red portion of its interval.



5.3 Limitations of the study

¶49 The number of records created for any individual in the watch lists could depend on how reliable the information is about their name and date of birth and how comfortable the agent entering this information was with the pronunciation and spelling of the name. To the extent Muslims are represented (on average) with more records than non-Muslims, the proportion of Muslims in the watch lists will be *less* than the proportion of records classified as Muslim; and conversely, if more records are used per person on average for non-Muslims, the proportion of Muslims on the watch lists will be *greater* than the proportion of records classified as Muslim.

¶50 The accuracy of these results also depends on the accuracy with which the reference lists of known Muslims and known non-Muslims are correct and on the implicit assumption that the classification error rates for names on these known lists will be the same as the classification error rates for names on the watch lists.

5.4 Appendix B1: Instructions to the panel

I have examined the lists in greater detail and taken a large random sample for you to classify according to religious affiliation.

This is a perfectly random sample from the rows of the combined No-Fly and Selectee tables. The "Name" column in the attached spreadsheet file concatenates "FIRSTNAME" and "LASTNAME" in that order. I have cleaned up the names (which originally are in all caps) to remove most punctuation and accented characters.

To this sample I have added a few more names randomly sampled from lists of people with known religious affiliations, both Muslim and non-Muslim, taken both from US lists and worldwide lists, and processed them in the same manner. This yields 3,000 names altogether.

Your results in classifying these names, assuming they are generally correctly classified, will help justify claiming that you can identify religious affiliation (Muslim or not) accurately. To maintain rigor, I will not disclose how many such additional names there are. (There aren't so many as to greatly increase your efforts, but there are enough to keep you alert!)

I have put the entire collection of names in random order. This randomization gives you flexibility, because it means you don't have to work through all three thousand entries. If you go through them in the order given and have to stop short, you will be okay. If you complete just a few hundred you should achieve your minimal requirements. That would still be a legitimate random sample.

The worksheet has a second column. The "Recno" column is a numeric identifier I can use later to match your results to my tables that indicate what the religious affiliation is and what the source of each name is. (This identifier is redundant, because I can use the name alone for the matching, but the identifier is backup in case any of the names happen to get changed during your processing.) You can also sort the data on the identifier any time to restore the original random order, should the lines ever get out of order.

The worksheet has a second column. The "Recno" column is a numeric identifier I can use later to match your results to my tables that indicate what the religious affiliation is and what the source of each name is. (This identifier is redundant, because I can use the name alone for the matching, but the identifier is backup in case any of the names happen to get changed during your processing.) You can also sort the data on the identifier any time to restore the original random order, should the lines ever get out of order.

The additional names come from several sources and some of them include organizations and honorifics. I tried to remove all organizations and eliminated the honorifics, but it's likely I missed a few. It is possible, then, that some of these names will obviously not be people or will have other problems. (Note, though, that all the sources, including the watchlists, have misspellings.) **My suggestion is to classify each name into three categories: Muslim, non-Muslim, and not a person ("other").**

For the classification process, it can be useful for each reviewer (or small group of reviewers) to work independently, contributing one new column to this worksheet for each independent assessment. Then you can get together, paste your columns into this sheet along one another, and compare your results to produce a consensus classification, which you will record in a new column. This will give us more detailed insight into your classification by revealing names that might have been difficult for you to classify.

So, to summarize, the result of your classification work will be to augment the attached file with additional columns: one for each independent reviewer and a final column for the consensus. The columns will all be filled out down to the same row, which need not be the last row. The values in them will be whatever codes you find convenient to indicate whether a name is Muslim, non-Muslim, or does not seem to be a valid name. Down to that final completed row, the consensus column should never be blank: force yourselves to make a determination for every name you examine. Otherwise, to be conservative, we would have to analyze blank entries in whatever way is least favorable for you.

6 Appendix C: The **STAN** program

¶51 This program was executed in the **R** programming environment, **R** Core Team (2022). It generated a Monte-Carlo sample of 120,000 random values of the parameters. Table 4 and Figure 5 summarize its results.

```
stanmodelcode <- "
data {
  int<lower=1> N1;  // population sample size
  int<lower=1> N2;  // population sample size
  int<lower=1> Np;  // Muslim sample size
  int<lower=1> Nm;  // Non-Muslim sample size
  int<lower=0> X1;  // Numbers classified Muslim
  int<lower=0> X2;  // Numbers classified Muslim
  int<lower=0> Yp;  // Number classified Muslim among Muslim sample
  int<lower=0> Ym;  // Number classified non-Muslim among non-Muslim sample
}
parameters {
  real<lower=0, upper=1> p1;  // True proportion of Muslims
  real<lower=0, upper=1> p2;  // True proportion of Muslims
  real<lower=0, upper=1> pp;  // True success classification rate
  real<lower=0, upper=1> pm;  // True failure classification rate
}
model {
  Yp ~ binomial(Np, pp);
  Ym ~ binomial(Nm, pm);
  X1 ~ binomial(N1, p1 * Yp * 1.0 / Np + (1-p1) * (1 - Ym * 1.0 / Nm));
  X2 ~ binomial(N2, p2 * Yp * 1.0 / Np + (1-p2) * (1 - Ym * 1.0 / Nm));
}
generated quantities {
  real q;
  q = (p1 * 1566062 + p2 * 251169) / (1566062 + 251169); // Overall
}"
fit <- stan(model_code = stanmodelcode,
        model_name = "Watchlists",
        data = list(N1 = 2330, N2 = 372, X1 = 2211, X2 = 319,
                Np = 146, Nm = 152, Yp = 135, Ym = 145),
        iter = 1e4, warmup = 2e3, chains = 12,
        sample_file = "STAN/Watchlists.csv",
        verbose = TRUE)
```