A Rapid Test for Accuracy and Bias of Face Recognition Technology Supplementary Materials

A Image sourcing

Focus on recent images. Our method tests cloud services on recent photos – older photos are more likely to have been used for training by cloud providers, which would bias the results. Our method uses images that had been published on the web within 12 months before the test is run.

Our method does not filter photos using EXIF data, which is typically present in the photograph's file and often contains the date on which the photo was taken. That is because we found EXIF data to be sometimes misleading and reports the date a photo was uploaded/hosted rather than the date it was taken. Additionally, some images do not carry EXIF data.

Google News. The process starts by inputting each name into the Google News Search API, which, in turn, yields up to 100 recent news articles related to the specified query. For each news article, the "newspaper" Python package outputs a link for "the best image to represent this article (the first image in the HTML markdown where the main article lies)." This process yields an average dataset of between 30 and 80 images per input name, depending on the popularity of the name. The variability in the dataset size is contingent upon the popularity of the individual's name and the corresponding availability of relevant images in the news articles. Overall, we found that few articles are available for individuals of Asian origin, and thus, this method for sourcing images may not work well in general.

Google Images. The image retrieval process involves issuing requests to the Google Custom SearchAPI for each input name. Each query is designed to return a maximum of 10 items. The parameters of the request offer flexibility in specifying the number of results, their position, the date of the results, the result type (in this case, images), and more. To accumulate a total of 100 image links for a single name, a series of 10 requests is made, systematically varying the position of the results in each subsequent query. The variability of the number of images is contingent upon the popularity of the individual's name and the corresponding availability of relevant images.

Lists of names. The full list of names and meta information for the two datasets sourced in this study can be found here. We experimented with two methods of generating the lists. One of the authors manually generated the list of celebrities (mostly singers and actors), balancing different demographics. The list of athletes was automatically generated by sampling from the 2020 Summer Olympics Wikipedia page. The *Celebrities* dataset was constructed using Google News with the individual's name as a search query. The *celebrities* dataset includes 80 names and is divided into eight demographic groups. We compiled the list by selecting 10 names for each group, determined by gender (male/female), racial background (Asian/Black/White), and age (junior/senior) for Whites only. Demographic information on gender, age and race (for Celebrities) was obtained from Wikipedia and matched other public information. The number of images obtained for each identity is histogrammed in Fig. S.1 (left). For the *Athletes* dataset, we used Google Images for the above-mentioned reasons. The query was constructed using the athlete's name and country (i.e., "<FirstName> <LastName> <Country>"). We did not use race information but rather the athletes' country's continent. The dataset contains 2755 names originating from 74 countries, strategically selected to achieve gender balance within three distinct ethnic origins: Africa, East Asia, and Europe. The criteria for country selection included a deliberate effort to maintain an approximate equilibrium between males and females within three distinct ethnic origins: Africa, East Asia, and Europe. Notably, the chosen countries were characterized by historical homogeneity, ensuring a focus on regions where demographic mixing has been limited. The countries within each are:

- **Africa**: Angola, Bahamas, Bahrain, Barbados, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cayman Islands, Central African Republic, Chad, Democratic Republic of the Congo, Eritrea, Eswatini, Ethiopia, Gabon, Ghana, Grenada, Guinea, Guinea-Bissau, Guyana, Haiti, Ivory Coast, Jamaica, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Republic of the Congo, Rwanda, Senegal, Sierra Leone, Solomon Islands, Somalia, South Sudan, Sudan, Tanzania, The Gambia, Togo, Trinidad and Tobago, Uganda, Zambia, Zimbabwe.

- **East Asia**: China, Chinese Taipei, Hong Kong, Japan, Mongolia, South Korea.

- **Europe**: Austria, Belarus, Belgium, Czech Republic, Denmark, Estonia, Finland, Germany, Iceland, Latvia, Liechtenstein, Lithuania, Norway, Slovakia, Slovenia, Sweden, Ukraine.

The number of images obtained for each identity is histogrammed in Fig. S.1 (right).

Duplicate removal. Duplicate and quasi-duplicate photos are unnecessary and artificially distort accuracy estimates. We identify and remove duplicate images by computing their cosine similarity in MobileNet [22] embeddings, pre-trained on ImageNet. Pictures were organized into similarity groups if the cosine similarity was greater than 0.9, and the medoid image of each group was retained. We eliminated duplicates twice: first, on sourced images and then again later for the cropped faces (Sec. 4).

Challenges in identity consistency. A potential issue occurs when the same individual is inadvertently listed multiple times with variations in their name (e.g., "Barack Obama" and "Barack Hussein Obama"). This situation will result in overlapping sets of images treated as different identities for what is technically the same identity. Such overlaps will cause the evaluation process to underestimate model accuracy. It is the responsibility of users to avoid such duplicates.

Manual label annotation. Manual annotation was done using a self-developed browser-based interface. All faces from each query were presented together one query at a time on the browser, with options to display either the full or cropped image (refer to Sec. 4). A first pass was made based on facial appearance. Ambiguous faces, where it was otherwise impossible to identify the person solely based on the image, were reviewed, and a determination was made using meta information (from the website where the image was published or from image captions). There are rare cases of non-photorealistic images (e.g., signal processing filter, paintings, caricatures, pixel art) being assigned a positive label as well, as we have observed that FRT services usually manage to identify those correctly. The manual annotation process took a total of 200 hours, about 12 seconds per image on average. The annotator responsible for labeling all the images is one of the co-authors and resides in Europe, potentially increasing the likelihood of annotation errors for Asian faces. To assess the accuracy of our labels, all disagreements between the manual annotation and our method's automatic ID assignment (Sec. 6) were reviewed by two people, reducing the likelihood of errors to a very low level. We specifically reviewed about 600 disagreements, revising approximately 500 for Athletes and about 10 for Celebrities (out of about 25,000 total labeled by our method). Thus, we estimate that our labels are $> 98\%$ correct for Athletes and $> 99.5\%$ correct for Celebrities.

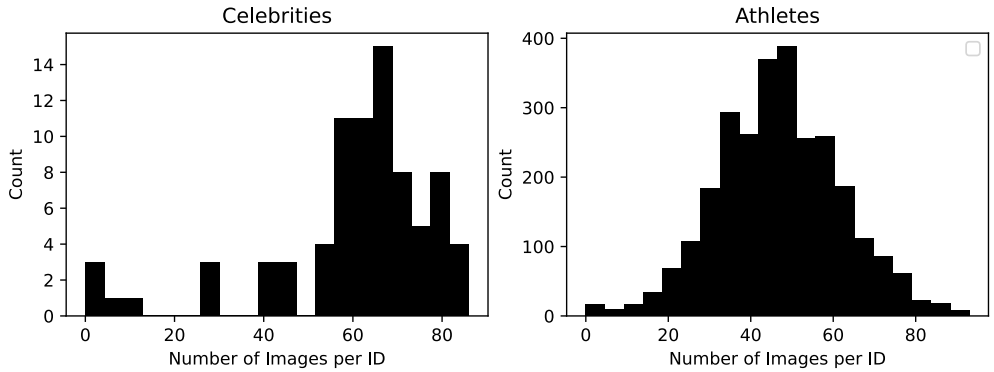
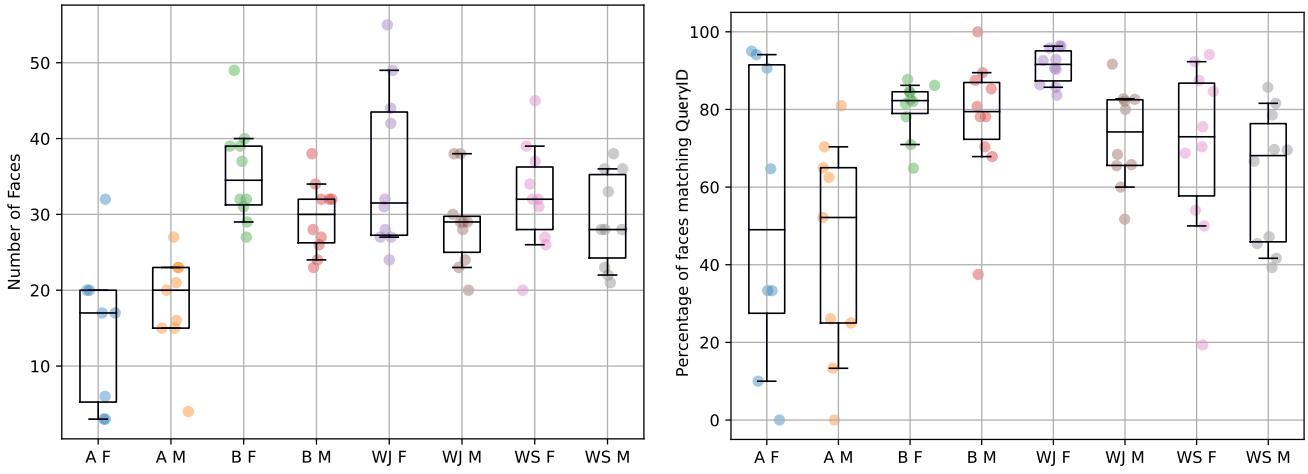
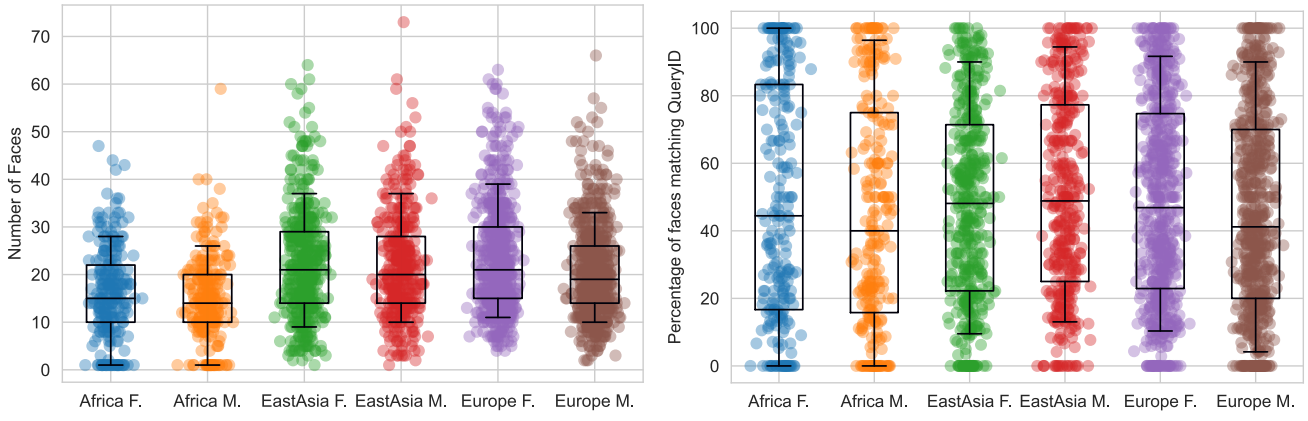


Figure S.1. **Number of images per identity.** The histograms show that, on average, more images were obtained for the Celebrities than for the Athlete datasets.



(a) **Celebrities** N. of images (b) **Celebrities** % of correct images

Figure S.2. **Statistics for the Celebrities dataset.** (a) Number of face images per demographic group (nomenclature below). (b) Percentage of correct face images (the identity of the person in the image matches the name queried) per name by group, as established by hand labeling. Each marker represents an individual. The box plot spans the interquartile range (75th and 25th percentiles of the data), and the whiskers extend to the 90th and 10th percentiles. Nomenclature: F: Female, M: Male, A: Asian, B: Black, WJ: White Junior, WS: White Senior.



(a) **Olympics** N. of images (b) **Olympics** % of correct images

Figure S.3. **Statistics for the Athletes dataset.** (a) Number of downloaded face images per name by group. (b) Percentage of correct face images (the identity of the person in the image matches the name queried) per name by group. Each marker represents one individual. The box plot spans the interquartile range (75th and 25th percentiles of the data), and the lines extend to the 90th and 10th percentiles. F. stands for Female, and M. for Male.

B Face detection

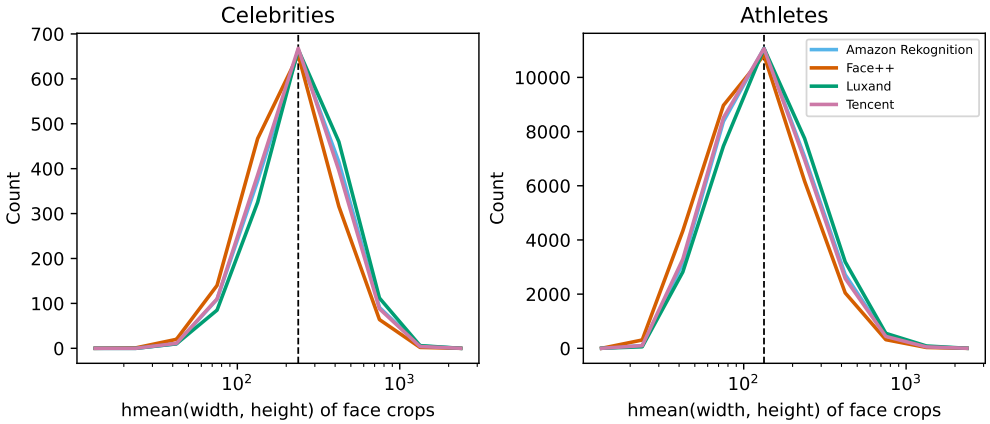


Figure S.4. **Histogram of face sizes.** The harmonic mean of width and height (in pixels) for each face image is used as a proxy of image size. Only images that are included in the validation experiments (Sec. 8) are considered here. Notice that the Athletes dataset face images have slightly lower resolution than the Celebrities dataset's. Face++ crops faces more tightly and Luxand more loosely than the other services.

Fig. S.4 shows a distribution of the sizes of the face-bounding boxes for each provider. Not all services detect every face in an image. We only keep those face images that are detected by all services. Our method maps one bounding box per provider to a common FaceID as described in Sec. 4. Across all retrieved images, our method is able to assign 78.3% (Celebrities) and 39.4% (Athletes) of the detected bounding boxes faces to a common FaceID. If we additionally employ the restriction to only keep images that show exactly one face (see Sec. 4) we keep 19.4% (Celebrities) and 39.0% (Athletes) of the detected bounding boxes, respectively.

Some 1:1 face matching services might not offer a face detection service (e.g., Verigram). In this case, one can use the "Multiple faces detected" error response of the matching service to efficiently determine the subset of single-face images as follows:

Algorithm 1 Procedure to find images containing exactly one face for services without a face detection API.

```
let valid_imgs be a set
let invalid_imgs be a set
for each (i1, i2) in image_pairs do
    if i1 in invalid_imgs or i2 in invalid_imgs then
        continue
    end if
    result = compare_imgs(i1, i2)
    if result == "invalid" then
        if i1 in valid_imgs then
            invalid_imgs.add(i2)
        end if
        if i2 in valid_imgs then
            invalid_imgs.add(i1)
        end if
    else
        store_confidence_value(i1, i2, result)
        valid_imgs.add(i1)
        valid_imgs.add(i2)
    end if
end for
```

C Identity label estimation

Table S.1. **Confusion matrices for annotations vs estimations.** See Sec. 6 for details on how these labels are assigned. y denotes the label that was assigned by hand, and $\hat{y}$ is the label that was assigned by our algorithm. $y = -1$ was assigned when faces were not unambiguously identifiable by the human annotator (e.g., occluded faces). In addition, we report the number of faces that were crawled but excluded from the analysis ("n excluded") as they did not meet the minimum requirements: at least 8 crawled faces must be present per query, and all services must have been able to make an estimate based on the image.

(a) Celebrities

	$\hat{y}=1$	$\hat{y}=0$	$\hat{y}=-1$
$y=1$	1213	2	422
$y=0$	3	338	218
$y=-1$	0	0	0

n excluded: 13

(b) Athletes

	$\hat{y}=1$	$\hat{y}=0$	$\hat{y}=-1$
$y=1$	12311	117	16576
$y=0$	254	4351	21076
$y=-1$	1	0	26

n excluded: 3913

(c) Celebrities 2024 (Appx. F)

	$\hat{y}=1$	$\hat{y}=0$	$\hat{y}=-1$
$y=1$	874	3	65
$y=0$	9	145	81
$y=-1$	0	0	0

n excluded: 50

Split-IDs. A practical challenge when using clustering-based methods for pseudo-annotation are clusters of images belonging to the same identity but marked as separate IDs. In practice, this can occur for individuals who undergo significant physical changes (e.g., due to facial surgery) or actors whose images are strongly associated with particular roles (e.g., an actor widely known as "Batman"). Our methodology addresses such cases by discarding identities that fail to form a single coherent cluster during the grouping process (see Sec. 6 and Fig. S.5). By discarding these challenging cases our algorithm slightly overestimates the accuracy of the models.

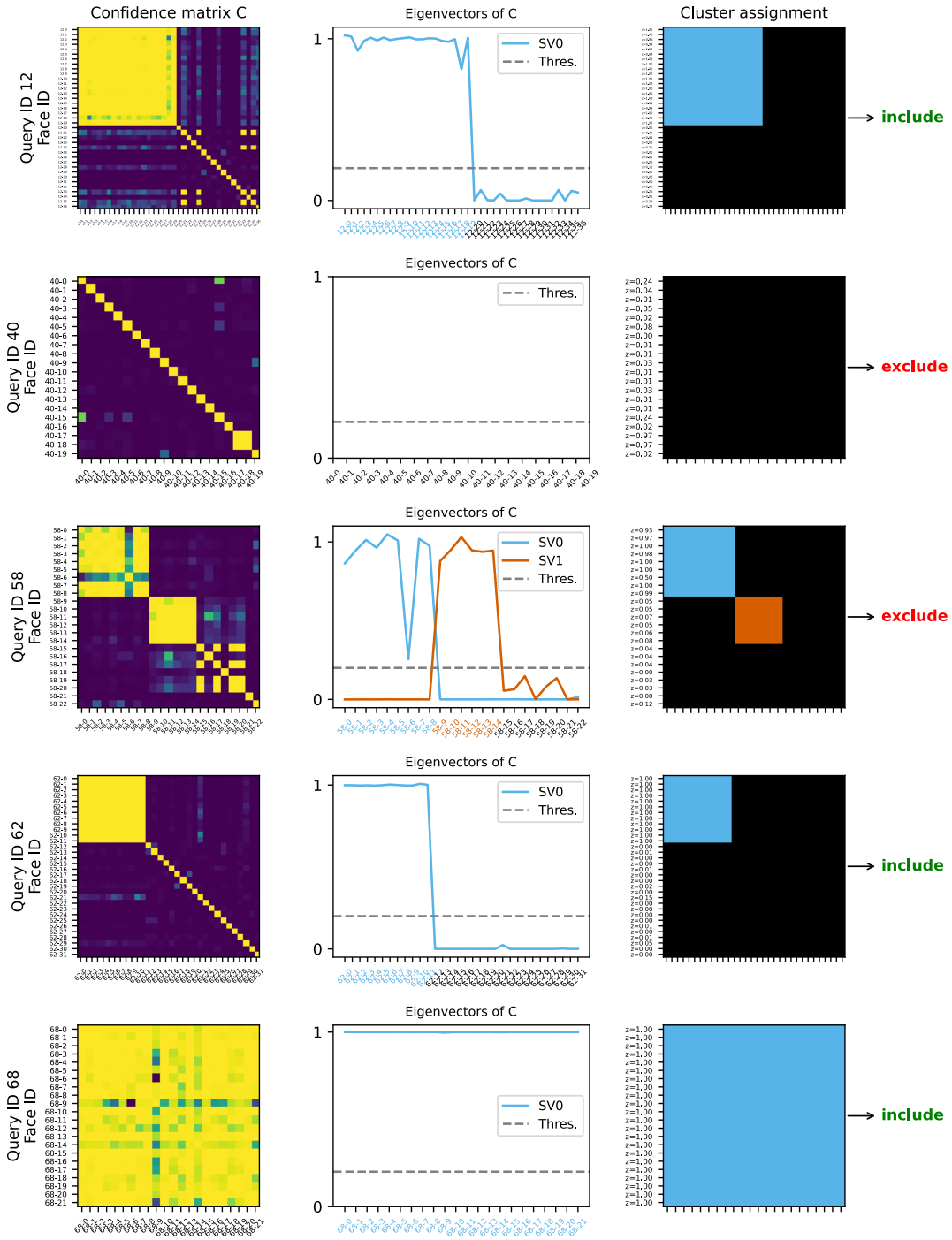


Figure S.5. **ID label estimation examples.** The procedure is described in Sec. 6 and this figure provides additional examples to supplement Fig. 3. The first column shows the pairwise confidence matrix C that was obtained from one service by comparing all pairs of faces that were associated with a given name query (reported on the left). The second column shows the eigenvector(s) of C that meet the criteria described in Sec. 6. The third column shows the groups, as well as the final decision, that are computed by our algorithm. The first row shows an easy case with about half the faces belonging to a dominant identity and the rest belonging to unrelated identities. The second row shows a case where all the identities are unrelated. The third row shows two, perhaps three, dominant identities (our algorithm recovers two). The fourth row is similar to the first row, with fewer images belonging to the dominant identity. The last row shows an easy case, where all the images are associated with the same identity.

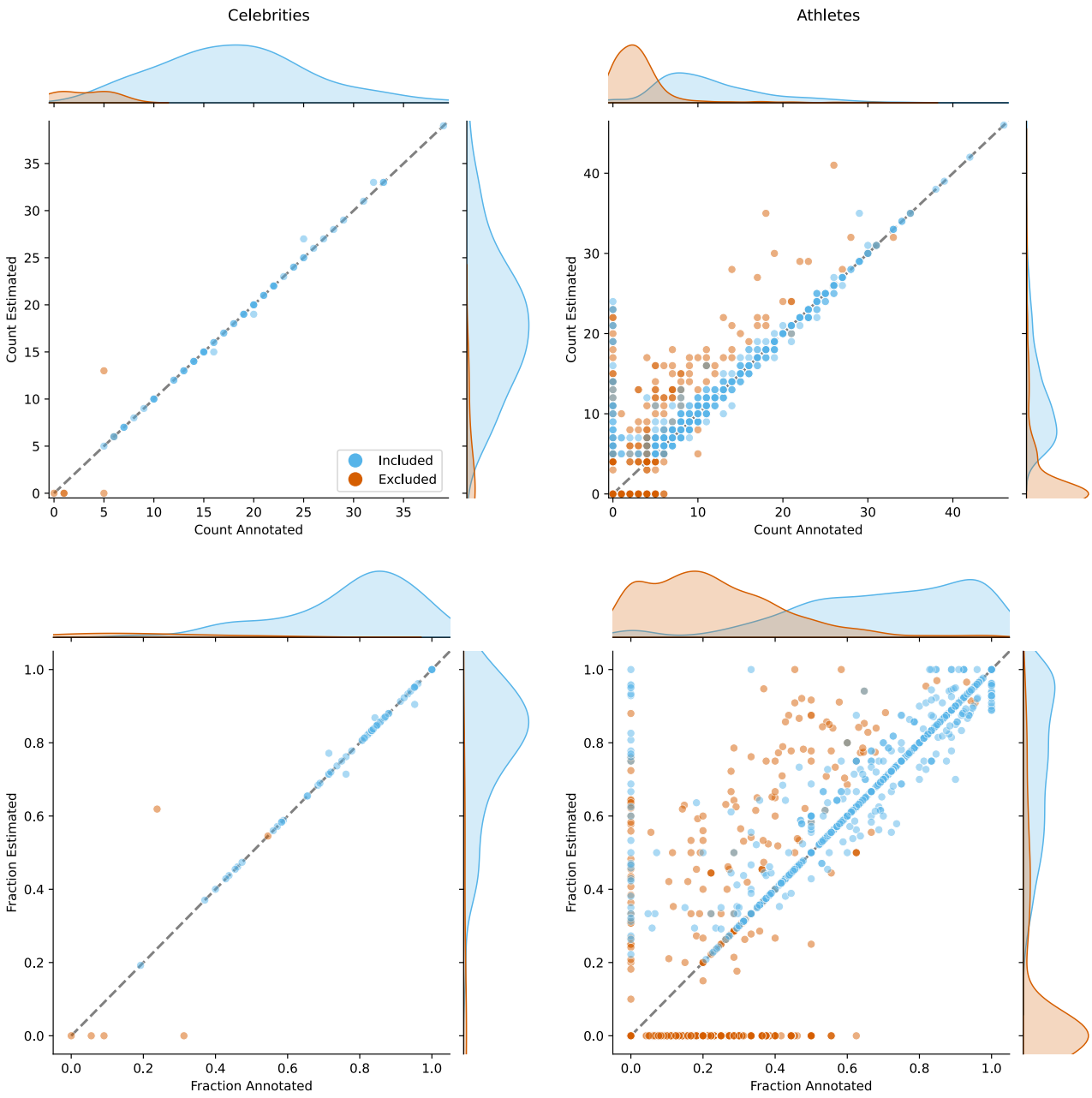


Figure S.6. **Number (count) and percentage (fraction) of *correct identity* faces per query.** The plots show the absolute numbers (top row) and fraction (bottom row) of correct identity faces for each query q – one dot per query. The color of each dot shows which queries were excluded from further consideration by our algorithm as described in Sec. 6. This plot does not include queries that do not fulfill minimum requirements, which means that all queries contain at least 8 images.

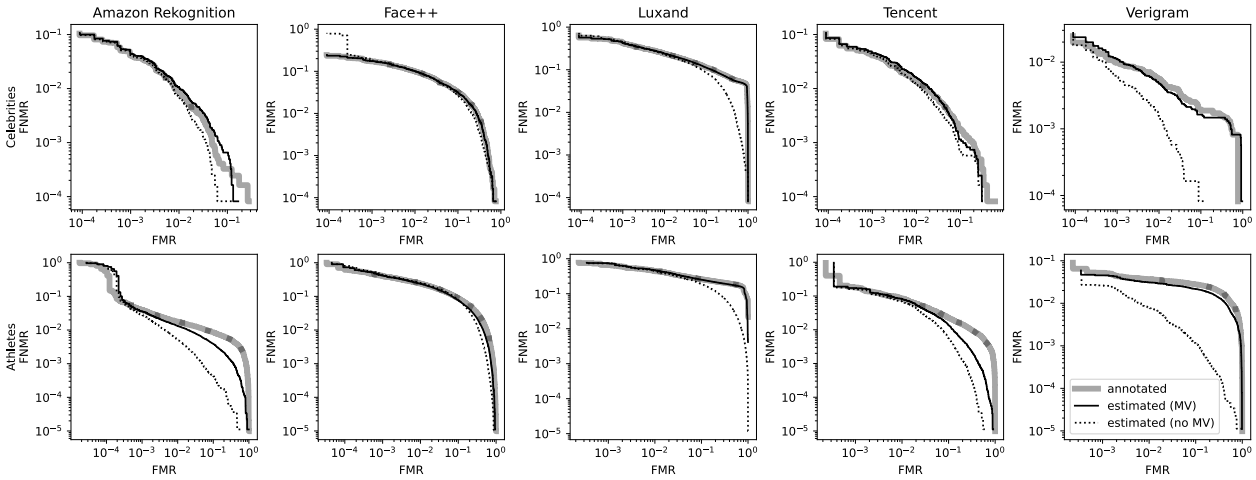


Figure S.7. **Effect of majority voting on identity label estimation.** Estimated FMR-vs-FNMR curves are shown before (no MV) and after (MV) consolidation between services. Majority voting yields estimates that are closer to those obtained through hand-annotation. We use majority voting in our method.

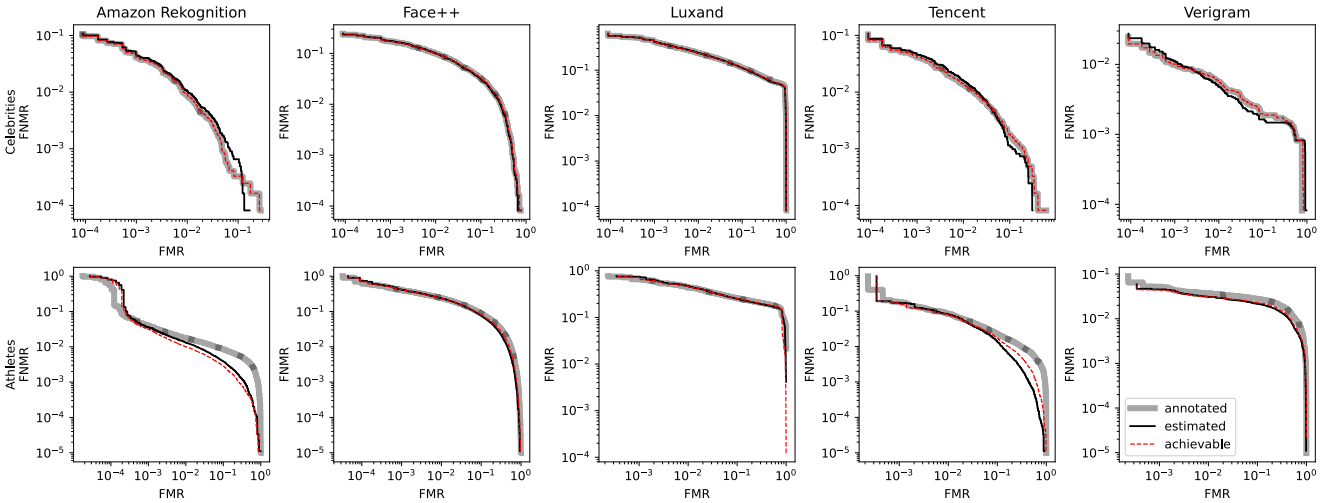


Figure S.8. **Achievable accuracy for the estimated test set.** The red dashed line indicates the maximum achievable accuracy when $\forall(\hat{y}_i \mid \hat{y}_i \neq -1) : \hat{y}_i = y_i$. Compared to the annotated curve, the difference is due to the fact that our method leaves out certain faces where preconditions for correct ID assignment are not given. We can conclude that the error between the annotated and estimated curve for Celebrities mainly stems from wrong assignments ($y_i \neq \hat{y}_i \neq -1$, *Type B errors*, see Sec. 6 for explanation). In contrast, for the Athletes dataset, the visible error is caused by the smaller intersection of annotated and estimated face image sets as we drop significantly more faces in this dataset (*Type A error*, see Tab. S.1).

D Additional service evaluation results

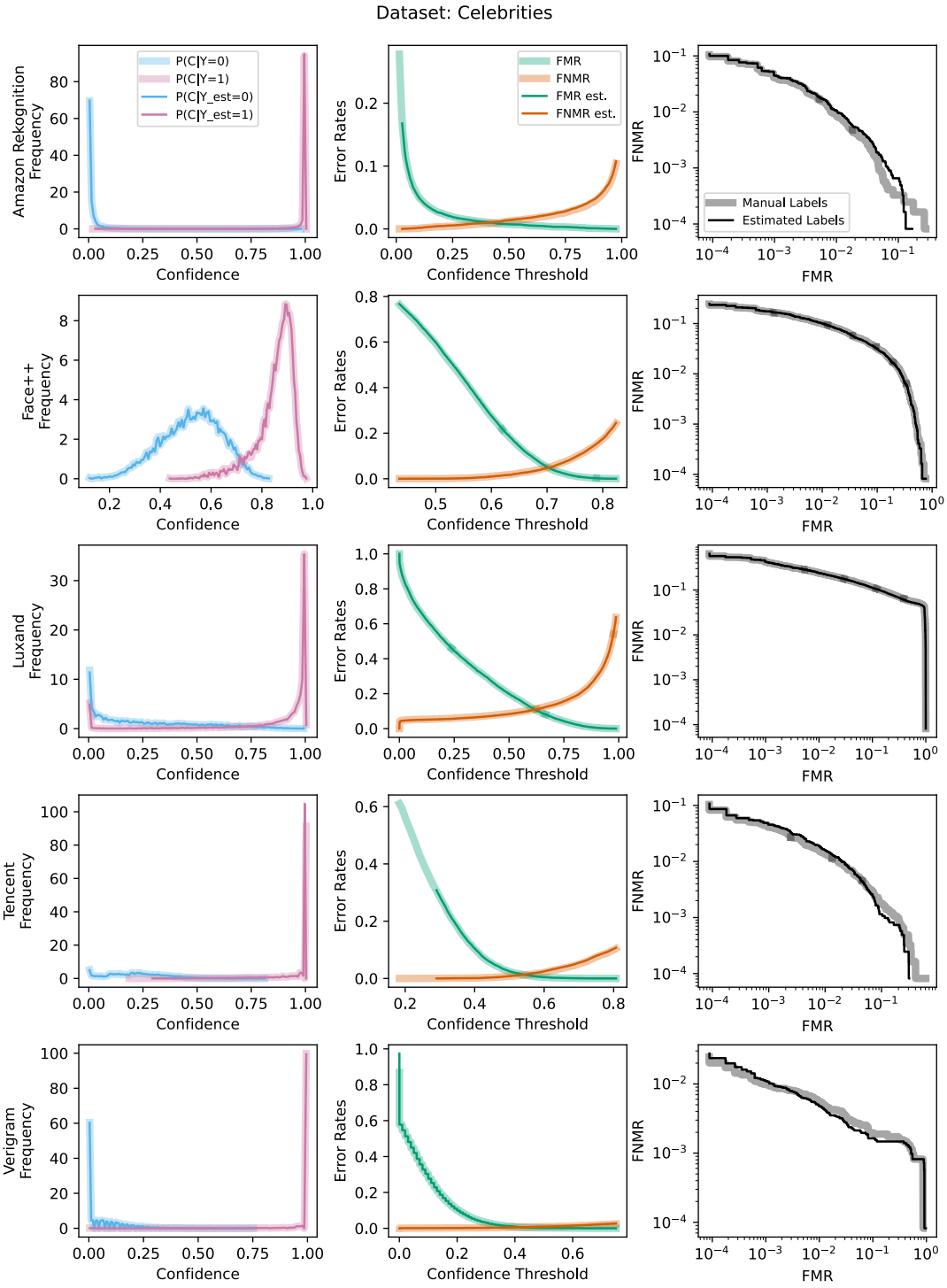


Figure S.9. **Verbose service evaluation results for the Celebrities dataset.**

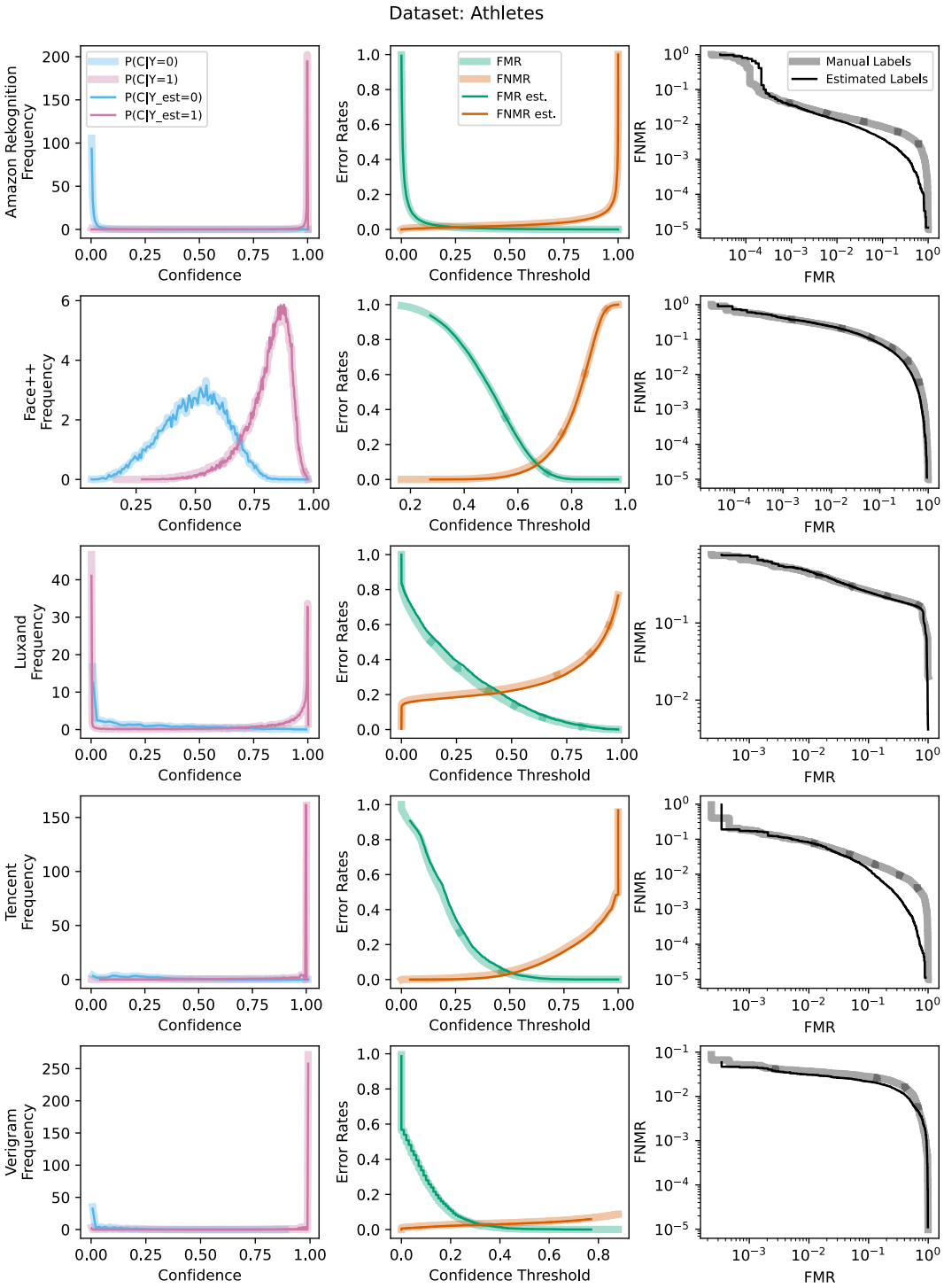


Figure S.10. **Verbose service evaluation results for the Athletes dataset.**

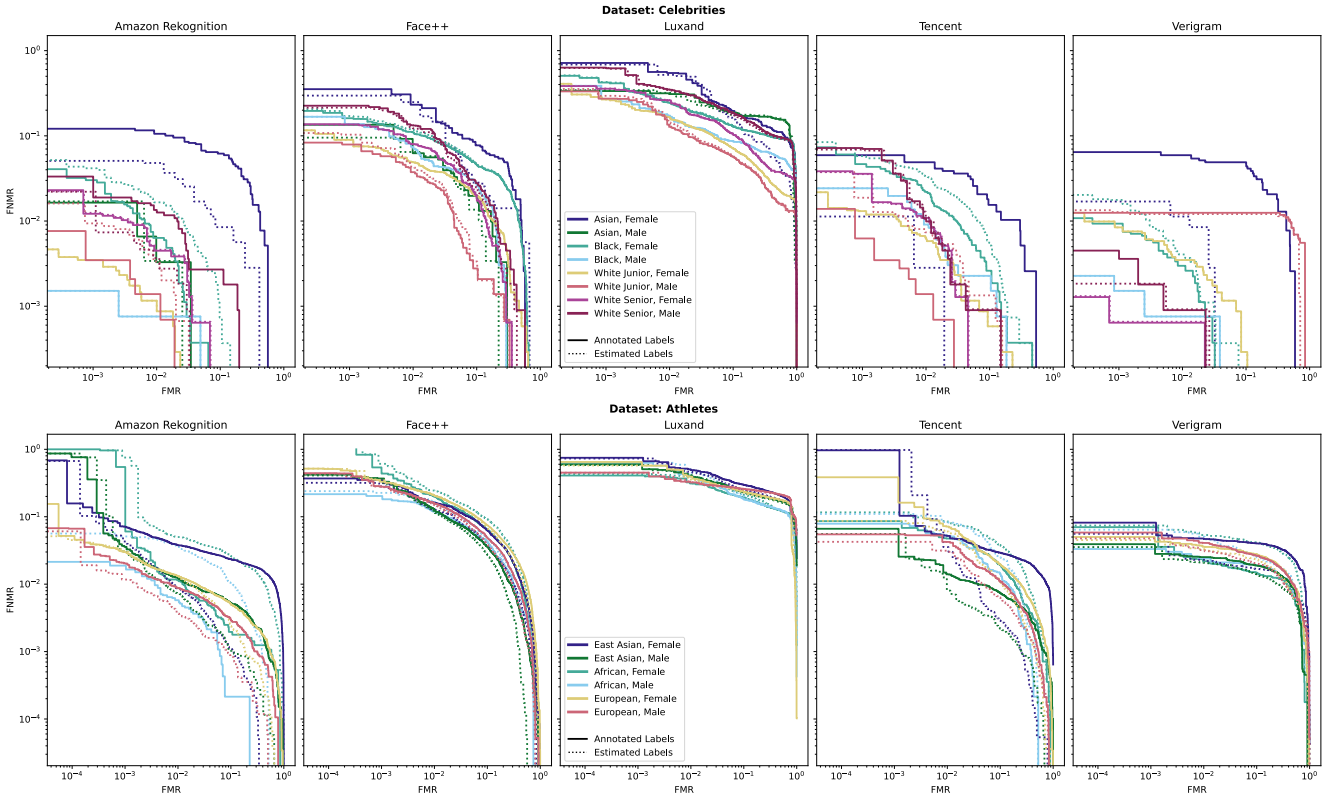


Figure S.11. **FMR-FNMR curves by demographic groups.** Each panel shows results for a single service. The top row is based on the celebrities dataset, and the bottom row on the athletes dataset. See also Fig. 5.

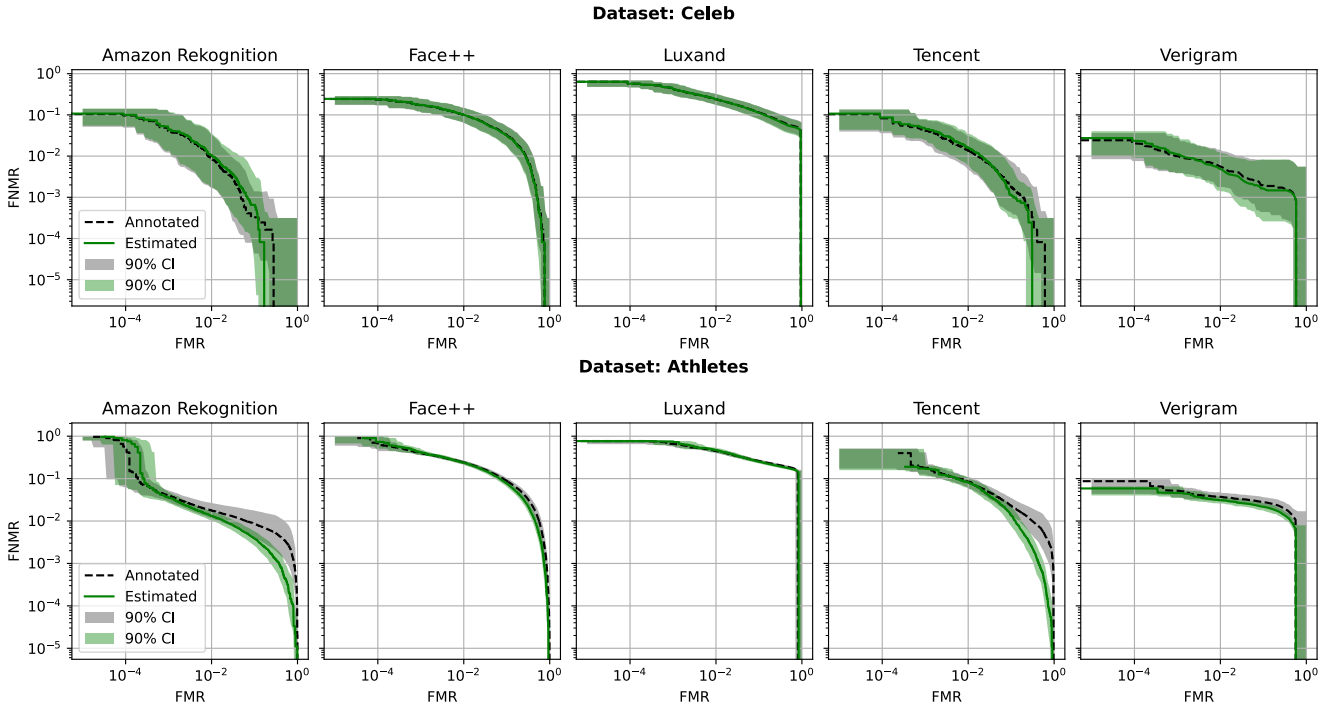


Figure S.12. **FMR-FNMR curves with Wilson confidence intervals [17].** The solid line is our method's estimate, and the dashed line is the hand-labeled annotation. Each panel shows results for a single service for the Celebrities (top row) and Athletes dataset (bottom row).

E Robustness regarding service composition

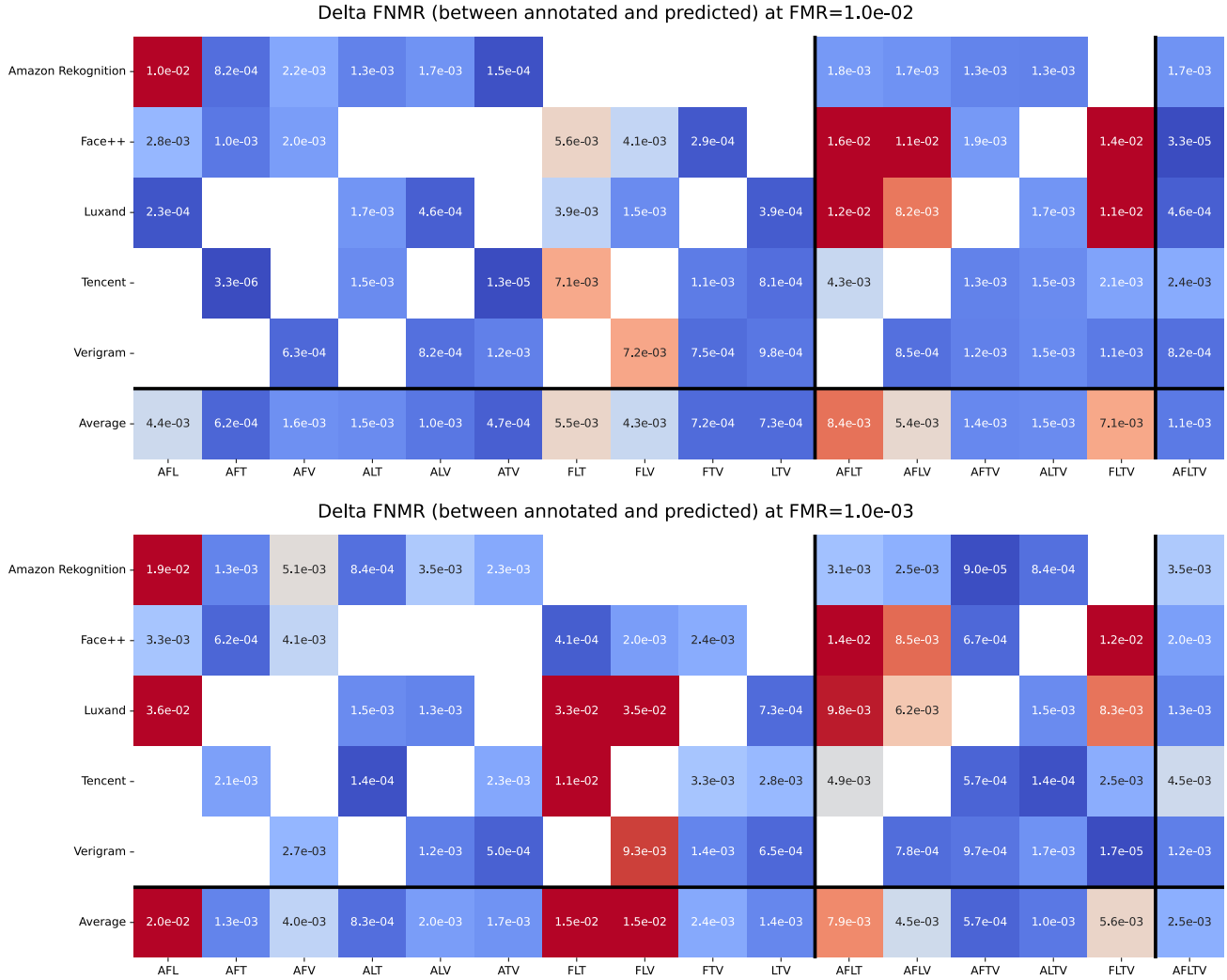


Figure S.13. **Effect of service composition on label estimation accuracy using the Celebrities dataset.** We want to test how sensitive our method is w.r.t. the set of included services. To measure the accuracy of our predictions, we calculate $\Delta\text{FNMR} = |\text{FNMR}_{estimated} - \text{FNMR}_{annotated}|$ at fixed FMRs of 0.01 (top) or 0.001 (bottom). Columns indicate different sets of included services abbreviated with their first letter. White squares indicate that the particular service (row) was not included in this subset (column). We test configurations of 3, 4, and 5 included services and find that the inclusion of services that have the lowest accuracy (Luxand) leads to a larger error in many 3-service and 4-service settings, while it can be compensated in the 5-service setting.

F Robustness over time

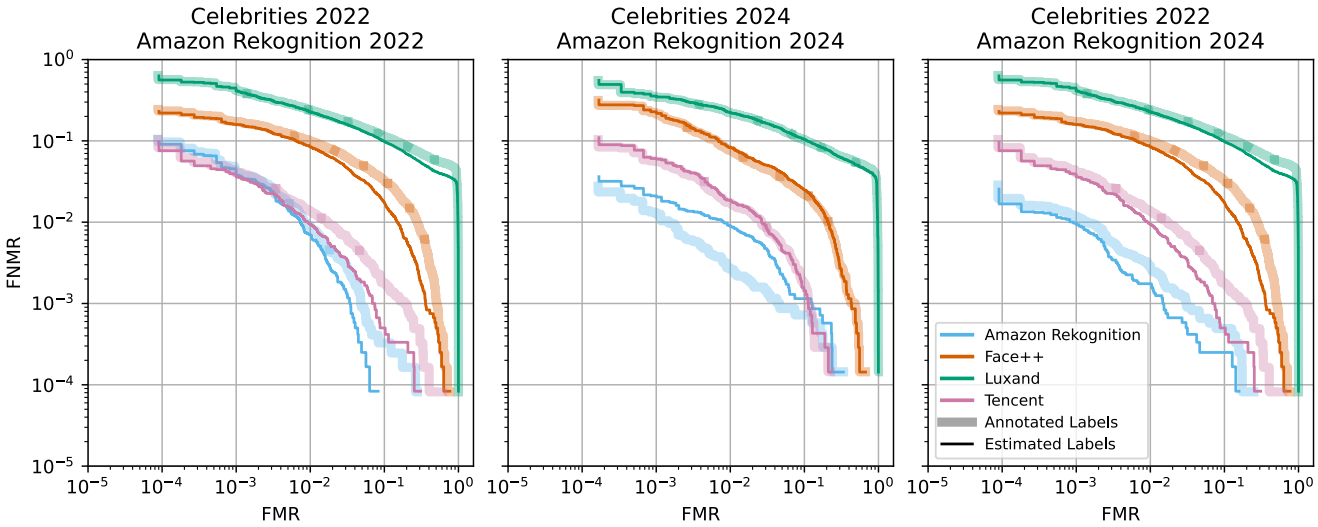


Figure S.14. **Probable model update in the Amazon Rekognition service between 2022 and 2024.** In this analysis, we focus on the robustness of our method over time with changing datasets and/or models. The left panel shows estimations based on the Celebrities dataset used in the main paper originally collected in 2022. We did a rerun of our method using the same services and the same list of names in 2024 (mid-panel). As described earlier, this results in a different set of images and possibly a change in the service's underlying model. The 2024 rerun shows similar results for three out of four services (Face++, Luxand, Tencent) and improved accuracy for Amazon Rekognition. To determine if this improvement is a result of a possible model change, we ran the 2024 version of Amazon Rekognition on the 2022 dataset combined with the other three 2022 services (right panel) and found that the improved service accuracy persists. Therefore, we conclude that a model change has likely happened for the Amazon Rekognition service between 2022 and 2024. We show that our method is generally robust over time, even if the underlying evaluation dataset is dynamic by design. Note: Verigram results are omitted as we did not have API access anymore when the 2024 experiments were conducted.

G Semi-supervised results

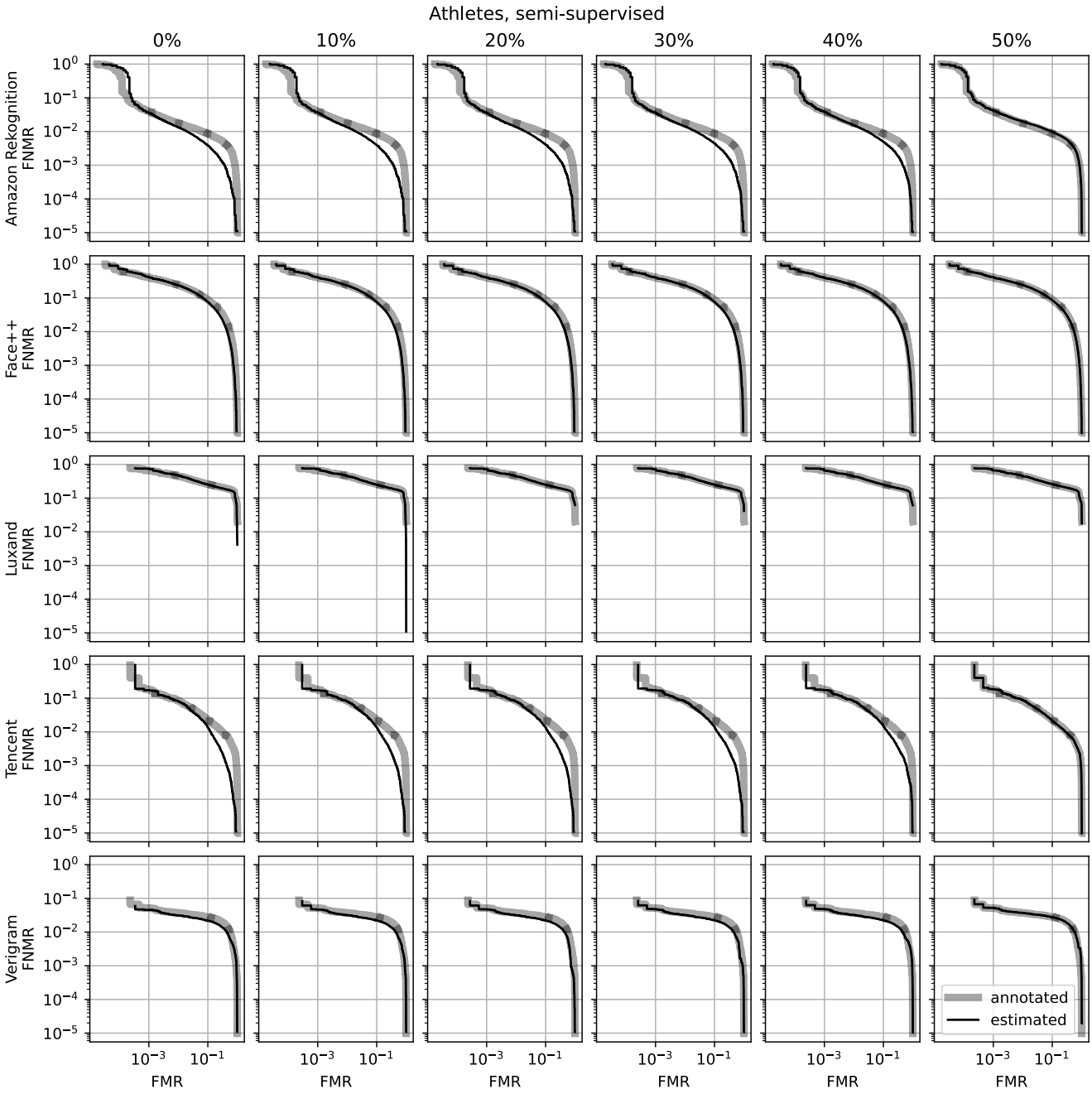


Figure S.15. **Semi-supervised FMR-FNMR curves using the Athletes dataset.** Our method allows the combination of estimated and annotated labels for semi-supervised accuracy estimation. Columns indicate the fraction of faces where annotated labels are used. As explained in Sec. 6 and shown in Fig. S.8, the disagreements between annotated and estimated curves stem from two types of errors. We prioritize to correct *Type A* errors by including faces to the estimation dataset that were initially excluded by our method ($\hat{y} = -1$). Once there are no more excluded faces to add, we correct *Type B* errors by replacing the estimated labels of those faces that have the highest degree of ambiguity according to their z-values. One can see that the curves gradually align with the 100%-annotated ones and reach near-perfect alignment with 50%-annotated labels.