Filling in data gaps on refugee children: Statistical imputation for missing count data

Jan Beise, Yukun Pei (UNICEF); Sebastian Steinmueller (UNHCR)

UNHCR's annually published population dataset with the number of refugees and their age/sex composition by country of origin and country of asylum contained missing age data for 23 per cent of the global refugee populations at end-2021. Figures of the global proportion of refugees who are children disseminated to date used the available data only. We show that this approach likely leads to an upward biased estimate of the global proportion of children due to the omission of refugee groups in the Americas and Europe with lower proportions of children. We propose a multilevel Bayesian model to create a dataset with multiple imputations of age/sex counts for country pairs with missing data, resulting in standardisation of global and regional estimates by countries of origin and regions of asylum. The estimate of the proportion of children in the global refugee population derived from the imputed dataset is 42 per cent compared to 47 per cent in the available-data figure.

Table of contents

List of Figures

List of Tables

Demographic data on refugees

Reporting process and data structure

UNHCR compiles official statistics on stocks and flows of forcibly displaced and stateless persons twice a year, once for mid-year figures (Mid-Year Statistical Reporting, MYSR) and once for end-year figures (Annual Statistical Reporting, ASR). For these reporting exercises, UNHCR's country offices compile aggregate population figures for refugees and asylum-seekers hosted in their country from a range of sources and data producers such as governments, UNHCR's own refugee registration database proGres and sometimes non-governmental actors. The geographic level on which theses statistics are being produced is therefore always the country country of asylum. The figures undergo a statistical quality control process at the country, regional and global level of the organisation and are disseminated on the publicly available refugee data finder (UNHCR 2022) after applying statistical disclosure control to suppress very small counts of persons that could identify individuals. The models and results presented in this paper use data only on the end-year stock of refugees under UNHCR's mandate (including others in need of international protection and refugee-like populations[1], excluding Palestine refugees under UNRWA's mandate), that is, not on asylum-seekers, internally displaced persons and stateless persons.

The end-year figures compiled with reporting date 31 December contain sex- and age breakdowns of the stocks refugees under UNHCR's mandate. The data is available by country of origin and country of asylum. The variables [sex]_[agebracket] contain the counts of refugees as of 31 December 2021 in the individual sex and age brackets in the respective country pair. For example, *female_12_17* contains the number of female refugees aged 12 to 17. Variable *total* is the total number of refugees over all sex/age categories.

[1] For better readability, we will refer to refugees, refugee-like populations and others in need of international protection as refugees throughout this text.

Data gaps

Pre-defined sex-specific age brackets are 0-4, 5-11, 12-17, 18-59 and 60 years and older. For some origin/asylum combinations however, only the total end-year count without any demographic information is available. Table 1 displays six rows of this dataset for illustration (age brackets other than 0-4 and 5-11 not shown for better readability).

Table 1: UNHCR end-2021 data

origin	asylum	missingAge	female_0_4	female_5_11	male_0_4	male_5_11	total
LBY	GBR	Age missing	NA	NA	NA	NA	4, 468
VEN	CUW	Age missing	NA	NA	NA	NA	17, 000
ECU	USA	Age missing	NA	NA	NA	NA	1, 525
ETH	SSD	Age not missing	419	422	442	466	4, 532
NGA	CMR	Age not missing	9, 386	17, 550	9, 465	17, 504	120, 928
COL	CAN	Age not missing	103	175	94	190	2, 651

Since reporting of official figures is undertaken at the level of the country of asylum, the reasons for missingness of age- and sex-disaggregated data in UNHCR's official statistics are a function of statistical capacity and coordination in the country in which refugees are hosted. This might in many countries be a lack of priority or technical capacity to produce such disaggregated figures, or in some instances unwillingness to share detailed data. Table 2 provides an aggregate overview of missing age data in the dataset.

Table 2: Missingness of age data

Data availability	Number of refugees	% of refugees	Number of country pairs
Age missing	5, 984, 840	23	2, 639
Age not missing	19, 748, 877	77	2, 023

Figure 1 shows for what proportion of the refugee population living in each region age-disaggregated data was available at the end of 2021. While demographic coverage is close to universal for refugees hosted in Africa, it is available for 41 per cent of refugees in Europe, 93 per cent in Asia and only for 40 per cent in the Americas. This is to a large extent a result of the differing population data sources in these regions: While the individual demographic details of refugees in many countries in Africa and MENA are recorded in UNHCR's own case registration system proGres, population data in other regions often comes from government offices with varying degrees of availability of demographic data.

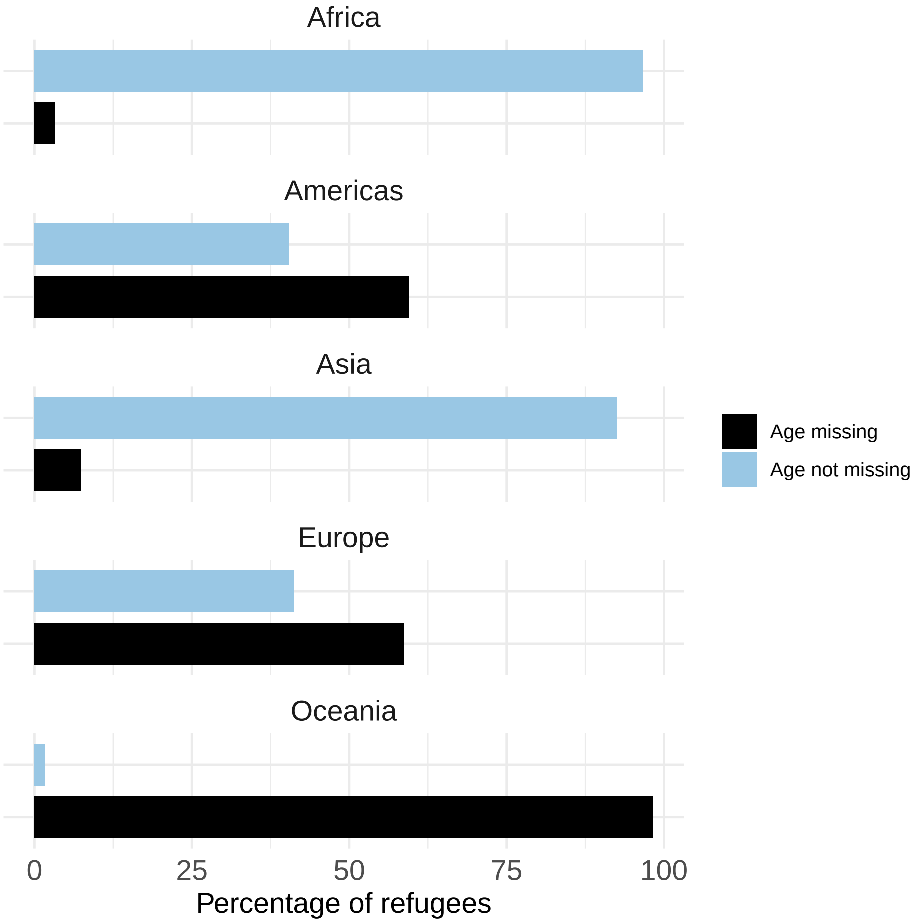

Figure 1: Missingness by region of asylum

UNHCR has in the past typically reported the sex/age breakdown in the available data as global and regional aggregates of the demographic distribution of all refugees. By reporting the observed demographic distribution as the sex/age structure of the entire refugee population including the part without available data, we are assuming that the 23 per cent for whom no age information was available at the end of 2021 have the same age distribution as the ones with available data. It is difficult to check this very strong assumption of ignorability of the missing data without further information on the sex/age distribution in the missing part of the data. We can however compare the distribution of other, fully available variables between refugees with and without demographic information. If such variables can be assumed to be correlated with the sex/age distribution at least to some extent, this can give us an indication whether the ignorability assumption is likely to be justified or not.

In particular, we can look at the distribution of data availability by country and region of asylum, and we can furthermore compare the distribution of origins of refugees in the observed and the unobserved part of the population. If missingness of demographic data was entirely random and thus ignorable, we would expect the geographic origins of refugees to be similar in the observed and the unobserved part of the demographic data, that is, we would see a similar distribution of origin countries.

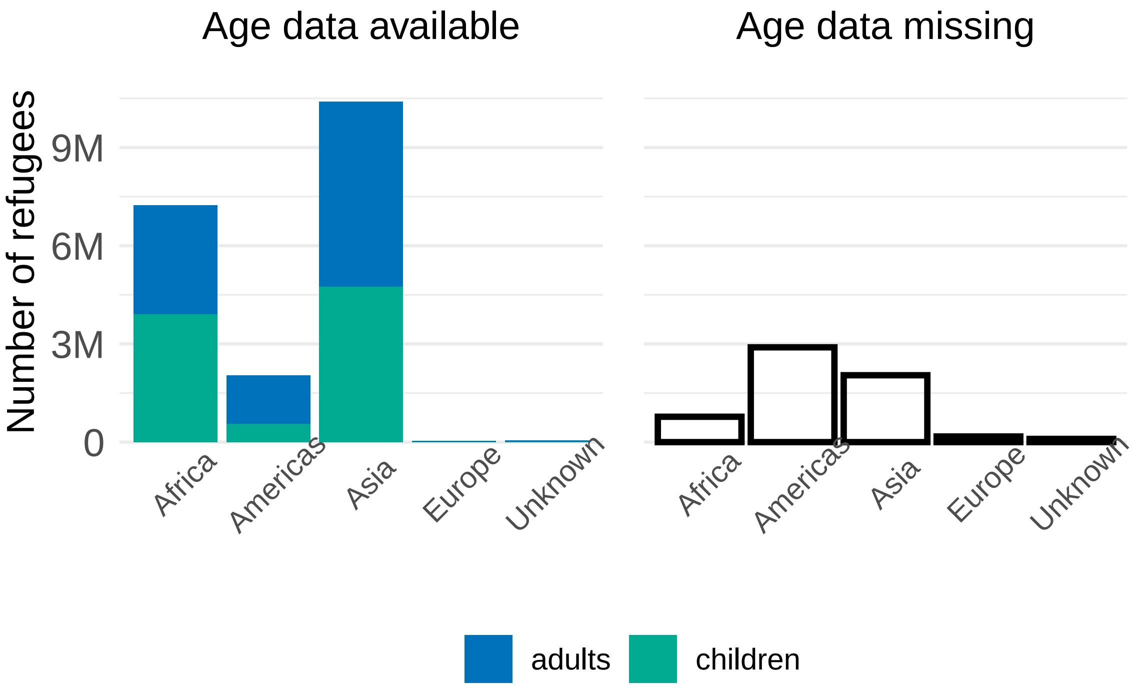

Figure 2: Missingness by origin region

Figure 2 shows the distribution of refugees by origin regions separately for the two subsets of the global refugee population with and without age-disaggregated data. The most common origin regions are Sub-Africa and Asia for refugees with available demographic information.Those without demographic data availability have most commonly fled from countries in the Americas and Asia. The left part of the graph shows that in the available data, the proportion of children is lower in these two regions. This provides a first indication that refugees with available sex/age-disaggregated data have different demographics from those without such data, and that we cannot simply assume the same demographic distribution between these two groups. The varying missingness in the data across regions of origin is indirectly related to the way demographic data is more available for countries of asylum in Africa and Asia: Since 72 per cent of refugees are hosted in neighbouring countries of asylum (Global Trends 2021), it is not surprising to see better data availability not only for refugees in, but also for those from these regions.

The fact that the distribution of origins of refugees for whom demographic data is missing differs from those with available data would not necessarily be an issue for global and regional estimates of the poportion of children among refugees if that proportion did not differ much between countries and regions of origin. Figure 3 displays the proportion of children in each country of asylum (each dot is one origin/asylum population, only shown for populations of at least 500 refugees with available age-disaggregated data) for the four largest origin countries at the end of 2021. Each of the four countries of origin has its own distinct distribution of the proportion of children among refugees in different countries of asylum: Refugees from South Sudan have very high proportions of children across countries of asylum in Africa, while the percentage of children among Venezuelans residing primarily in South American countries is mostly below 30 per cent. This is not very surprising: even without seeing the data, we would expect populations from the same origin country to have at least in many cases similar demographic distributions due to the same or comparable reasons and periods of displacement and fertility patterns. There is more variability across countries of asylum for refugee populations displaced from Afghanistan. Additionally, we can observe a clustering of the proportions by region of asylum for each country of origin. For example, there is a relatively clear pattern of Syrian refugees in Asian countries having higher proportions of children than those living in African and European countries of asylum.

Figure 4 shows the proportion of children as a function of the distance between country of origin and country of asylum for the nine largest refugee producing countries. Dots for countries of asylum that neighbour the respective origin country are blue. The proportion of children by and large decreases for countries of asylum that are further away, and it is higher for neighbouring countries.

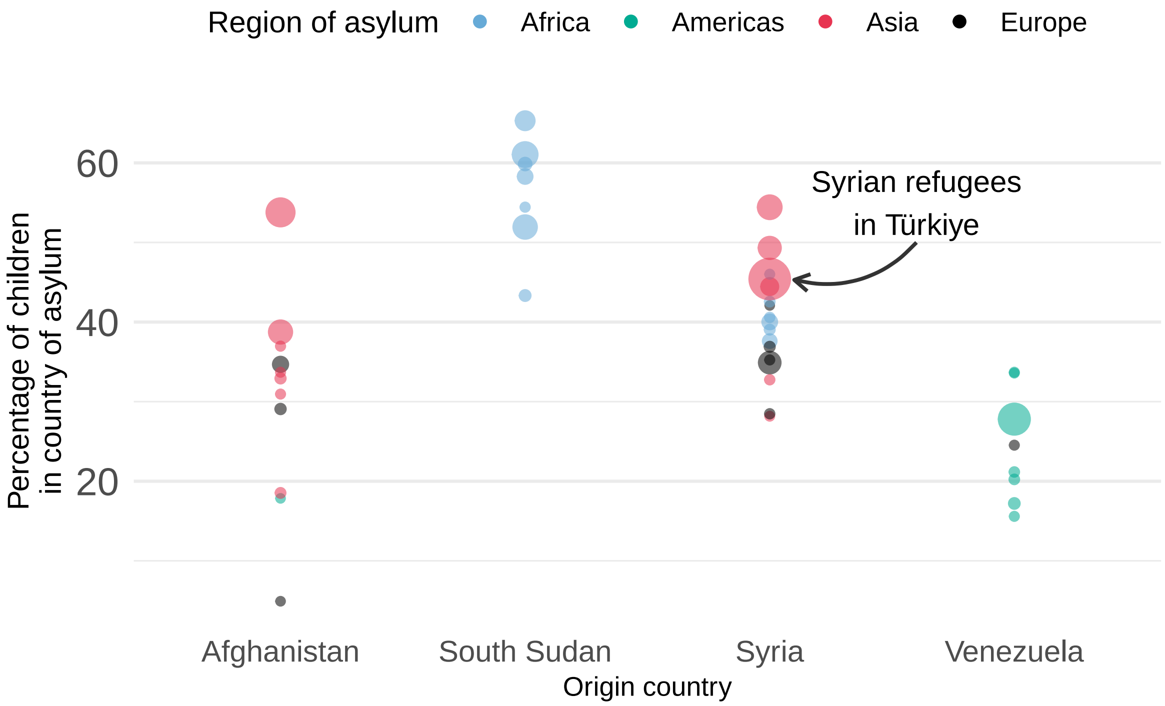

Figure 3: Proportion of children among refugees in each country of asylum by country of origin (dot size is proportional to number of refugees in the country of origin/asylum pair)

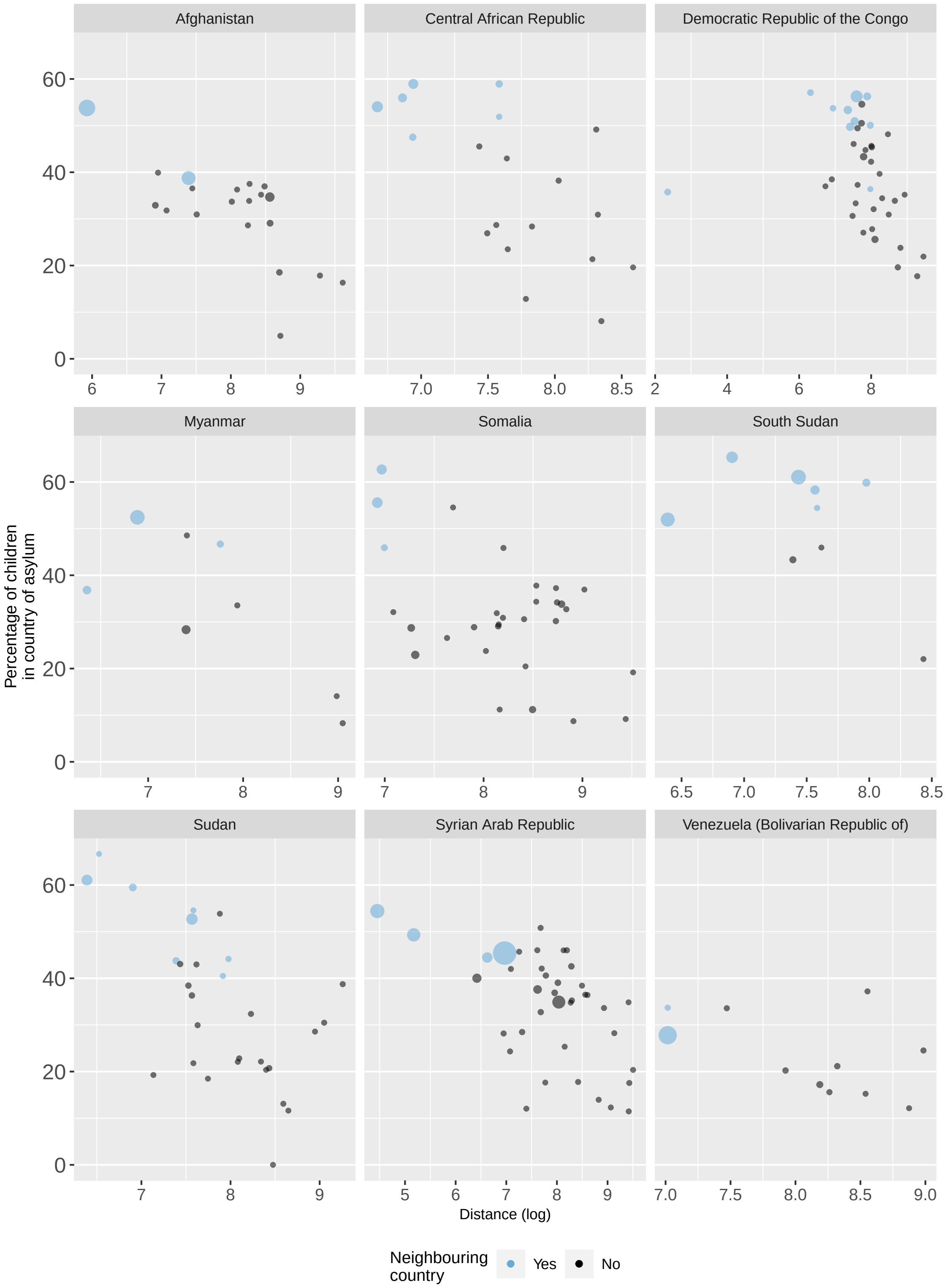

Figure 4: Proportion of children among refugees for nine large refugee origins by distance to country of asylum (dot size is proportional to number of refugees in the country of origin/asylum pair)

Model

Considering the findings of the descriptive analysis with considerable differences in the distribution of countries of origin and asylum between refugees with and without available data, and the explanatory power of origins and countries of asylum for the demographic characteristics of refugee populations, ignoring the missing data (i.e. treating it as if it was the same as the available data) risks introducing bias into regional and global estimates of the proportion of refugee children. In the following, we propose a model to predict counts for the age/sex brackets of those country pairs with missing demographic data (the *NAs* in Table 1), which in turn will allow us to estimate regional and global proportions of children. Summarising the descriptive analysis of missingness and available age data, there are three main conclusions to be considered in predicting the proportion of children for refugee populations without available age data:

1. We cannot assume the global proportion of children among refugees for whom demographic data is available is the same as the proportion we would see if data was available for populations with currently missing data.

2. Knowing which country of origin refugees are from gives us some information regarding the proportion of children we would expect. For example, looking at Figure 3 we would generally assume that there is a higher proportion of children in a refugee population from South Sudan than in refugees from Venezuela.

3. For each origin population, knowing which region of asylum a population is from can help us refine the estimate for the proportion of children based on the country of origin. If we know a population is from Syria and they reside in a country of asylum in Europe, we would assume the proportion of children to be lower than if they were living in a country in Asia.

When presented with the task of predicting the percentage of children among refugees in a country of asylum A from country of origin C, in the absence of any other information it would intuitively make sense to use the global proportion of children among refugees from that country for whom we have demographic data, i.e. to use data on populations from C from other countries of asylum. We can go further and only use data on refugees from C for countries of asylum in the same geographic region as A, making use of the clustering of proportions seen in Figure 3. While this procedure works for origins with a sufficient number of countries with available data in all world regions, in some cases we will encounter country pairs with missing data for which there is little or no data available for the same origin in the same region. In these cases, the next best option would be to use the global mean of the proportion of children for that origin country, or, if there is little data even across all countries of asylum, to fall back to the global mean from all origin countries. There are also going to be mix cases, where only data from a low number of countries of asylum in the same region with low population sizes is available. While we do not want to discard such data for the same origin entirely, it would make sense to additionally use data from countries of asylum in other regions for the same origin to avoid relying entirely on little and potentially unreliable data.

Multilevel models (also called hierarchical, or sometimes random effect, models) are well suited to pool information across geographic levels as described above. We fit a multilevel model with varying nested intercepts at the levels of country of origin, region of asylum and country of asylum. Since the data is available as aggregated counts in ten age/sex brackets that sum up to the population total, we choose a multinomial distribution to model the underlying probabilities of a refugee falling in each of the age/sex brackets depending on their country of origin and country of asylum. When estimating these probabilities, the model treats populations from the same origin country as being more similar to each other than those from different origins, and for each origin considers those residing in countries in the same region of asylum as more similar than those in different regions of asylum. We fit the model in a Bayesian framework that allows us to incorporate uncertainty in parameters and in predictions from those parameters in the final predictions.

For each refugee population i with a total of n_i people from country of origin c in country of asylum a in region r, we want to predict the vector of counts in the ten age/sex brackets $y_i = (y_{1,i}, ..., y_{k,i}, ...y_{10,i})$ ($y_{1,i}$ is the count of girls in the 0-4 age bracket, $y_{2,i}$ are girls in the 5-11 bracket, $y_{5,i}$ are women aged 60 and older, $y_{6,i}$ are boys aged 0-4 and so on). y_i follows a multinomial distribution:

$$y_i \sim M(n_i, \pi_i) \tag{1}$$

We fit a multilevel multinomial logit model to estimate from observed age/sex data the vector of probabilities $\pi_i = (\pi_{1,i}, ..., \pi_{k,i}, ...\pi_{10,i})$ expressing the probability of a refugee falling in age/sex bracket k given their country of origin, country and region of asylum and whether they live in a country neighbouring their origin country:

$$log(\frac{\pi_{k,i}}{\pi_{4,i}}) = \eta_{k,i} = \beta_{k,0} + \beta_{k,neighbor} + b_{k,c} + b_{k,c:r} + b_{k,c:r:a} \tag{2}$$

where the counts of the 4th age/sex bracket, women aged 18-59, are the reference category (therefore $\pi_{4,i}$ in the logit denominator). The subscripts for the varying intercept of region and country of asylum, c:r and c:r:a, indicate nesting within country of origin c (and country of asylum within region of asylum), meaning there are no crossed intercepts of region and country of asylum over countries of origin. The π_k sum up to one:

$$\sum_{k=1}^{10} \pi_{k,i} = 1 \tag{3}$$

We use the following priors for the model parameters, where Π_k is the proportion of the global population in that age/sex group in 2021 (UN DESA 2022):

$$\beta_{k,0} \sim N(log(\frac{\Pi_k}{\Pi_4}), 1) \tag{4}$$

$$\beta_{k,neighbor} \sim t(7, 0, 2.5) \tag{5}$$

$$sd(b_{k,c}) \sim t(5, 0, 2.5) \tag{6}$$

$$sd(b_{k,c:r}) \sim t(7, 0, 2.5) \tag{7}$$

$$sd(b_{k,c:r:a}) \sim N(0, 0.01) \tag{8}$$

The prior on the global intercepts with the $log(\frac{\Pi_k}{\Pi_4})$ mean aims to provide sensible levels for the age/sex bracket proportions. It incorporates our prior belief that, for example, it is unlikely that all refugees in a country are girls aged 0 to 4, but otherwise is not overly narrow. The student-t priors on the group-level standard deviations of the intercepts of country of origin and region of asylum are not informative and only serve to regularise the parameter space. The prior on country of origin is slightly wider to allow less pooling (i.e. closer to a separate model for each country of origin) and thereby take into consideration the general idea of the model that age/sex distributions are idiosyncratic to populations of an origin country. The very narrow $N(0, 0.01)$ prior on standard deviation of the country of asylum term $b_{k,a}$ merits further discussion: Including the term might seem unusual to start with since in most cases there is only one observation in the form of age/sex counts and total (i.e. one row in the dataset) per country of origin/asylum to start with (exception are countries in Latin America and the Caribbean with two populations displaced from Venezuela, refugees and others in need of international protection). On one hand, leaving out this parameter leads to overconfidence in predicting counts in country pairs with missing data since in that case the counts are drawn from the varying intercepts of the regions (which are observed in the data in most country of origin/region of asylum pairs), ignoring the uncertainty that comes with predicting a whole new country pair not used in fitting the model. On the other hand, allowing a wider standard deviation leads to unreasonably wide prediction intervals (sometimes including counts of zero children in populations of several thousand refugees). Restricting the prior by way of a very small standard deviation helps us constrain the predictive intervals to sensible ranges. We include a fixed effect for neighbouring country pairs since Figure 4 shows that we can on average expect higher proportions of children among refugee populations in neighbouring countries of asylum. Additional covariates such as the distance and GNI difference between country of asylum and origin would be interesting additions to the model, but were left out for this version since they led to too high a number of parameters and difficulties in convergence of the Markov chains used to sample from the posterior distribution.

To predict the count vector of age/sex brackets $\widetilde{y_{i^*}}$ for a country pair i^* without reported data and account for uncertainty of parameters θ and prediction, we draw for each country pair with missing data m values from the posterior predictive distributions:

$$p(\widetilde{y_{i^*}} \mid y) = \int p(\widetilde{y_{i^*}} \mid \theta) \cdot p(\theta \mid y) \, \mathrm{d}\theta \tag{9}$$

We can then analyse the m posterior draws y_{j,i^*} $(j = 1, ..., m)$ descriptively and calculate for example uncertainty intervals and location parameters to estimate regional and global proportions of age groups such as children. We will call each of the m draws of values for the missing part of the dataset an imputation.

For the group-level varying intercept of region of asylum, we choose a modified version of the SDG regional groupings in which some of the sub-groups are aggregated. The choice of region groups is somewhat arbitrary but also important because, for regions with little available reported data for a country of origin, the exact regional definition becomes influential for the distribution of the draws from the posterior predictions of country-level parameters. With the modified SDG groups, we try to create groups that are neither too small nor too big and, crucially, contain comparable countries of asylum. Table 3 summarises these regions including refugee population sizes with missing and available age data.

Table 3: Regional groupings and number of refugees with missing and available age data

Region	Age available	Age missing

Australia and New Zealand	0	57,400
Central and Southern Asia	3,187,108	345,496
Eastern and Southern Europe	9,166	475,087
Eastern Asia	600	308,174
Latin America and the Caribbean	2,032,849	2,642,796
Northern Africa and Western Asia	7,025,896	231,395
Northern America	47,456	421,848
Northern and Western Europe	1,306,267	1,398,875
Oceania	1,209	11,601
South-eastern Asia	238,718	4,685
Sub-Saharan Africa	5,899,608	87,483

The model was fit with R-package brms (Bürkner 2017). Reproducible code is available here. Annex I contains the model summaries.

Results

Global and regional estimates

Table 4: Regional and global model estimates for the proportion of children

Region	% available data	% estimated	Lower 95% CI	Upper 95% CI
Africa	54.4	53.8	53.4	54.4
Americas	27.7	25.3	24.3	26.3
Asia	47.2	46.2	45.2	47.8
Europe	32.3	30.4	29.5	31.5
Oceania	19.7	27.6	18.4	38.2
WORLD	46.8	42.2	41.7	42.9

Table 4 shows the percentage of children among refugees living in the five major world regions estimated with imputations from the model and compares them to the percentage we obtain using the available data only. The global estimate for the proportion of children drops by almost 5 per cent, from 47 per cent to 42 per cent, when using the imputed dataset. The

95% uncertainty interval for the global point estimate ranges from 41.7 to 42.9 per cent and does not include the available-data figure of 47 per cent. This indicates the standardisation by origins and regions of asylum through the model is relevant, and previously disseminated figures for the proportion of refugees who are children were most likely overestimates. Most of the decrease from the available-data to the modelled estimates is due to the fact that the latter uses data on all refugees residing in Europe and the Americas with lower proportions of children. Regional estimates from the imputed data are below the available-data figures with narrow uncertainty intervals for all regions apart from Oceania. The available-data figures in Oceania are based on only 1,209 refugees with reported data. The model pools the low estimate from this small dataset towards the global means for each origin, resulting in the comparatively large increase compared toe the available-data estimate. When interpreting the width of the uncertainty intervals, it is important to keep in mind that only predicted (formerly missing) counts vary over the imputations, but age/sex counts for country pairs with reported data are fixed over the imputations and do therefore not add to uncertainty. This means that regions like Africa with a low proportion of missing data have very few country pairs with imputed data and therefore have narrow uncertainty intervals.

Uncertainty checks for country level predictions

To assess whether the model estimates uncertainty sensibly, we can look at the range of predictions at country-pair level. Figure 5 shows the distribution of the proportion of children over the 200 imputed datasets among Syrian refugees in four countries of asylum without reported data. The posterior predictive distributions are relatively wide, with the 95% posterior predictive interval for the proportion of children among Syrians in the USA ranging from 15 per cent to 43 per cent. This is a desirable property of the models since, in the absence of reported data on an origin population in a country of asylum, we would not be confident in overly narrow prediction intervals. This is also the reason why it is not recommended to disseminate point estimates of demographic proportions in country pairs with imputed values. Showing uncertainty intervals, or even better the full posterior predictive distirbution, is necessary to provide a sense of the range of plausible values. The same recommendation holds for analysis on the level of country of origin (aggregating over countries of asylum for an origin), country of asylum. Generally, the country-pair level imputations with high uncertainty are best understood as building blocks to be used to aggregate up to and estimate demographic proportions and their uncertainty in regions of asylum and origin. The imputation dataset comes with iso3 codes for countries of asylum and origin that allow to easily create any regional grouping.

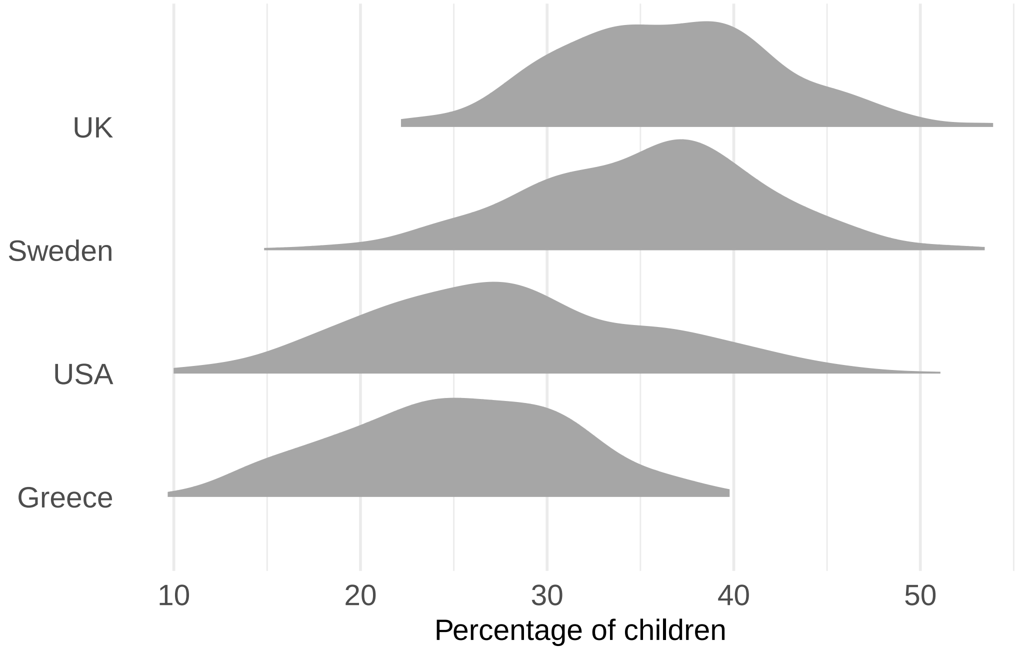

Figure 5: Uncertainty in posterior predictive distribution of the proportion of children among Syrian refugees in four countries of asylum

Figure 6 shows the same plot for Venezuelans in four South American countries. The distributions for the proportion of children among Venezuelans in Ecuador and Argentina are much narrower than for those in Brazil and Chile, or for Syrians in any of the countries shown above. In contrast to many other countries, demographic data is available in Argentina and Ecuador for some Venezuelan populations, leading to narrower prediction range in these countries.

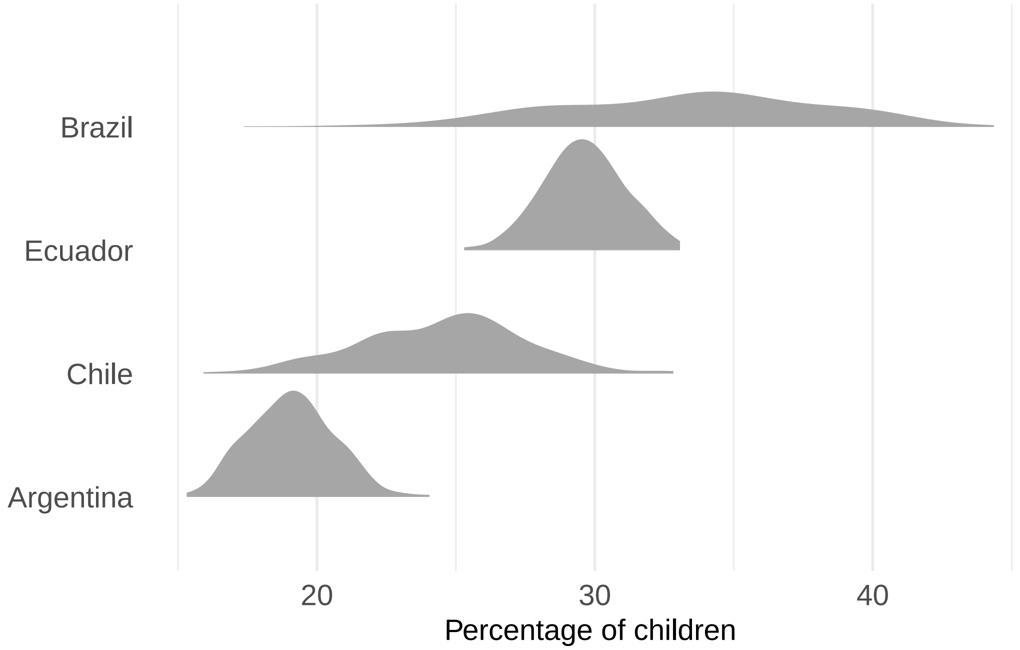

Figure 6: Uncertainty in posterior predictive distribution of the proportion of children among Venezuelan refugees in four countries of asylum

Figure 7 shows the distribution of reported (point only) and predicted (intervals) proportions of children among Syrian refugees in all countries by world region. Demographic data is reported from all countries in Africa and there is consequently no uncertainty in data from these countries. In Oceania in contrast, no age data on Syrian refugees was reported at all, leading to very large uncertainty intervals for Australia and New Zealand. The sub-regional groupings used in the model (see Table 3) are clearly visible in the plot panel for Europe, where estimates for countries with missing data are split into the two subgroups within the larger Europe region. Conditional on country of origin and the neighbour indicator, the main determinant of the central location of the posterior predictive distribution of the proportion of children in any one country of asylum is the region that country is in. In other words, Figure 7 is a visualisation of the two main functions of the model: To standardise global and regional counts by origin countries and regions of asylum, and to provide a measure of uncertainty coming from both model parameters and predictions in the imputed dataset.

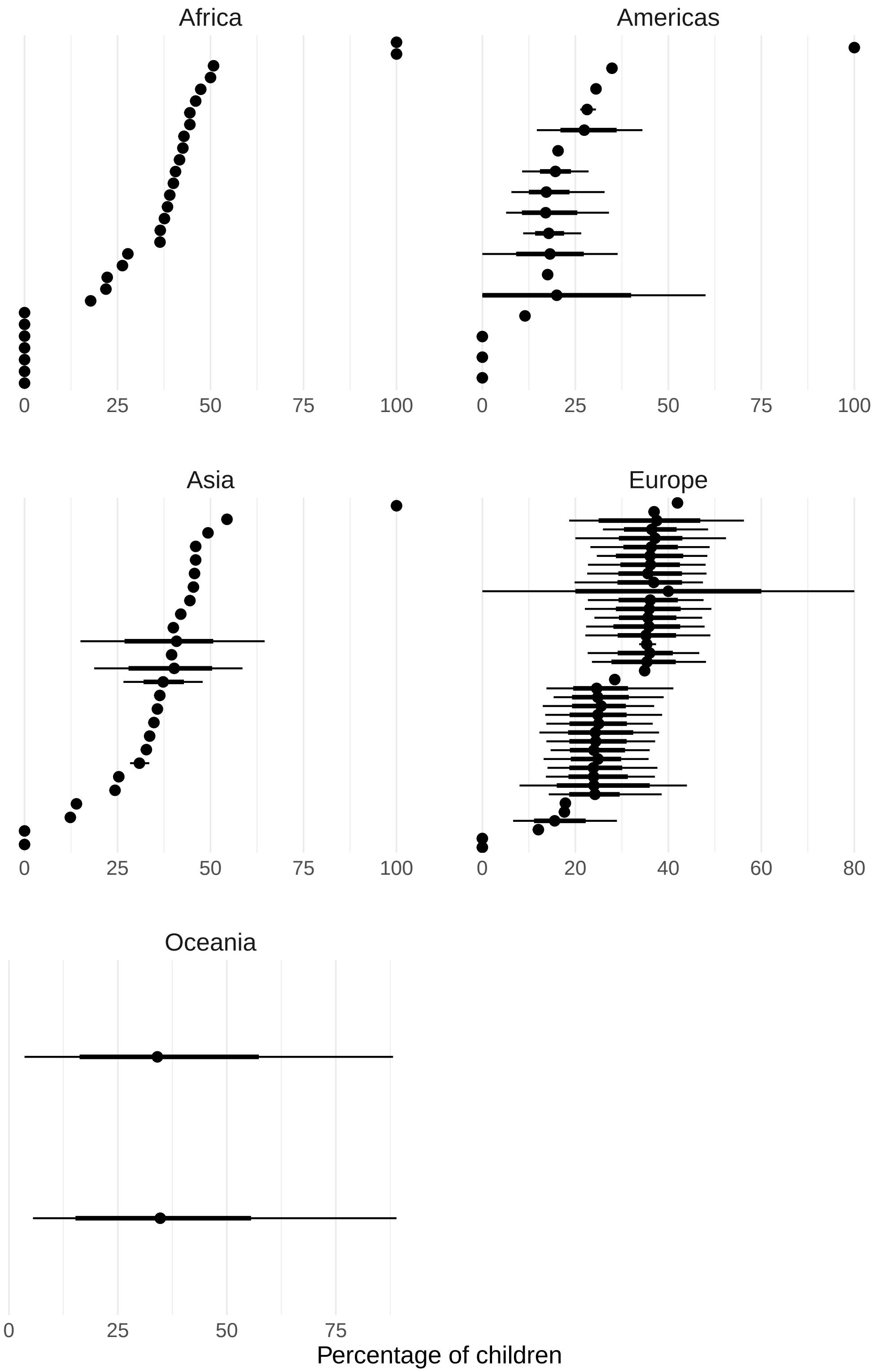

Figure 7: Uncertainty in posterior predictive distribution of the proportion of children among Syrian refugees across regions

Limitations

While the imputed dataset is useful in estimating regional and global estimates of the age/sex distribution of refugees, the modelling approach is top-down and uses only geographic proximity and regional groupings as a means to improve available-data figures. It cannot substitute data collected and compiled at the national level. If the main goal is to obtain reliable estimates of the demographic characteristics of refugees on the level of country of asylum, strengthening the production process of official statistics on refugees in said country should be the primary course of action. Recommendations and implementation examples developed by countries in the framework of the Expert Group on Refugee, IDP and Statelessness Statistics (see EGRISS 2018) can provide helpful guidance to national statistical systems. Projecting populations from census, survey or administrative data on age/sex structure, fertility, mortality and displacement are another way to obtain more reliable estimates of the age/sex distribution at country level, although this requires much more effort for a single country and is therefore better suited if estimates for a specific country are required. Estimates from the imputed dataset at any geographical level should always be disseminated and communicated with uncertainty intervals or by showing the posterior predictive distribution over imputations. While care has been taken to include and test prior assumptions and hierarchical geographic structures to model uncertainty from parameters and predictions, we made a blanket assumption that all available age/sex data was measured without uncertainty and without measurement error. This is an over-simplification. For example, counts for age/sex brackets on Afghan refugees in Iran reported by UNHCR for end-2021 are outputs of a demographic projection model with its own uncertainty, however these counts are included without variance in the current version of the imputation model. While it would be possible to include estimates of uncertainty for the available part of the age/sex data, this would be technically more challenging and lead to more complex models. It might also be institutionally more difficult since, at least to date, most global datasets such as the one on refugee demographics have been disseminated without measures of uncertainty.

Discussion

With the revised estimate for the proportion of children among refugees of 42 per cent being significantly lower than the available-data figure of 47 per cent and in light of the bias in the available data shown in the descriptive analysis, we recommend use of the imputed dataset where the estimation of regional and global figures of age/sex brackets is the goal of an analysis of UNHCR's official refugee data. The imputed dataset is also suitable for analysis on the level of country of asylum or country of origin, however, in these cases particular care should be taken to analyse and communicate uncertainty over the imputations, for example by disseminating interval- instead of point estimates. The imputed dataset contains iso-3 codes for countries of origin and asylum, allowing for straightforward aggregation over custom regions and country groupings as required by an analyst. Since UNHCR's official population dataset has the same format year-on-year, fitting the same model and creating an imputed dataset from its posterior predictive distribution can be repeated for future publications of the ASR data.

While it would seem logical to use the same approach to estimate proportions of children (or other demographic groups) among asylum-seekers and IDPs, the models do not lend themselves naturally to these datasets: The proportion of missing demographic data is much higher among these groups, and the uncertainty in estimating parameters and predicting new values from such sparse datasets would lead to high uncertainty over imputations. For example, age/sex data was only available in 13 out of 35 countries reporting IDPs at end-2021. For IDPs, an additional challenge is the lack of a geographical country of origin/asylum structure as for refugees. Country-by-country demographic projections using additional data such as from censuses and household surveys seem therefore a more promising modelling approach to estimate the age/sex structure for IDPs.

References

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Expert Group on Refugee, IDP and Statelessness Statistics (EGRISS). 2018. "International Recommendations on Refugee Statistics." *Online Edition.* https://egrisstats.org/wp-content/uploads/2021/12/International-Recommendations-on-Refugee-Statistics.pdf.

UNHCR. 2022. "Refugee Data Finder." *Online Edition.* https://www.unhcr.org/refugee-statistics/download/.

United Nations, Department of Economic and Social Affairs, Population Division (UN DESA). 2022. "World Population Prospects 2022." *Online Edition.*

Annex I: Model output

Table 5: Model - fixed effects

Coef	Estimate	Est.Error	l-95% CI	u-95% CI
mufemale04_Intercept	−2.69	0.12	−2.95	−2.45
mufemale511_Intercept	−1.89	0.11	−2.11	−1.67
mufemale1217_Intercept	−2.31	0.12	−2.56	−2.07
mufemale60_Intercept	−3.06	0.14	−3.35	−2.80
mumale04_Intercept	−2.74	0.14	−3.03	−2.48
mumale511_Intercept	−1.77	0.08	−1.94	−1.61
mumale1217_Intercept	−2.12	0.10	−2.33	−1.94
mumale1859_Intercept	−0.04	0.08	−0.19	0.11
mumale60_Intercept	−3.07	0.14	−3.34	−2.80
mufemale04_neighborNo	−0.67	0.05	−0.78	−0.58
mufemale511_neighborNo	−0.39	0.03	−0.46	−0.32
mufemale1217_neighborNo	−0.31	0.03	−0.37	−0.25
mufemale60_neighborNo	−0.55	0.04	−0.64	−0.46
mumale04_neighborNo	−0.61	0.05	−0.71	−0.51
mumale511_neighborNo	−0.40	0.03	−0.47	−0.33
mumale1217_neighborNo	−0.31	0.03	−0.38	−0.24
mumale1859_neighborNo	0.44	0.04	0.36	0.51
mumale60_neighborNo	−0.38	0.05	−0.46	−0.28

Table 6: Model - origin country group level standard deviation

Coef	Estimate	Est.Error	l-95% CI	u-95% CI
sd(mufemale04_Intercept)	0.43	0.20	0.03	0.81
sd(mufemale511_Intercept)	0.69	0.13	0.45	0.95
sd(mufemale1217_Intercept)	0.73	0.14	0.46	1.01
sd(mufemale60_Intercept)	0.81	0.16	0.51	1.12
sd(mumale04_Intercept)	0.66	0.16	0.36	1.00
sd(mumale511_Intercept)	0.46	0.10	0.25	0.66
sd(mumale1217_Intercept)	0.49	0.12	0.25	0.73

sd(mumale1859_Intercept)	0.47	0.08	0.34	0.63
sd(mumale60_Intercept)	0.88	0.19	0.50	1.28

Table 7: Model - asylum region group level standard deviation

Coef	Estimate	Est.Error	l-95% CI	u-95% CI
sd(mufemale04_Intercept)	1.63	0.10	1.44	1.84
sd(mufemale511_Intercept)	1.43	0.10	1.25	1.64
sd(mufemale1217_Intercept)	1.48	0.09	1.31	1.67
sd(mufemale60_Intercept)	1.43	0.11	1.23	1.67
sd(mumale04_Intercept)	1.56	0.10	1.38	1.78
sd(mumale511_Intercept)	1.10	0.07	0.96	1.25
sd(mumale1217_Intercept)	1.27	0.08	1.11	1.44
sd(mumale1859_Intercept)	1.07	0.06	0.96	1.19
sd(mumale60_Intercept)	1.67	0.12	1.45	1.91

Table 8: Model - asylum country group level standard deviation

Coef	Estimate	Est.Error	l-95% CI	u-95% CI
sd(mufemale04_Intercept)	0.31	0.01	0.30	0.32
sd(mufemale511_Intercept)	0.22	0.01	0.21	0.23
sd(mufemale1217_Intercept)	0.19	0.01	0.18	0.20
sd(mufemale60_Intercept)	0.22	0.01	0.21	0.24
sd(mumale04_Intercept)	0.31	0.01	0.30	0.32
sd(mumale511_Intercept)	0.22	0.01	0.20	0.23
sd(mumale1217_Intercept)	0.20	0.01	0.19	0.21
sd(mumale1859_Intercept)	0.31	0.01	0.30	0.32
sd(mumale60_Intercept)	0.26	0.01	0.25	0.27