Reviewers' Choice

Quantifying Uncertainty and Variability of Recessive Disease Prevalence Using Monte Carlo Estimates

American Society of Human Genetics 2021
2021-10-18

Peter Komar, Kaanan Shah, Connor Barnhill, Marcus Soliai, Jin Ju, Alex Petukhov, Morgan Paull, Michael Pettigrew and Sun-Gou Ji

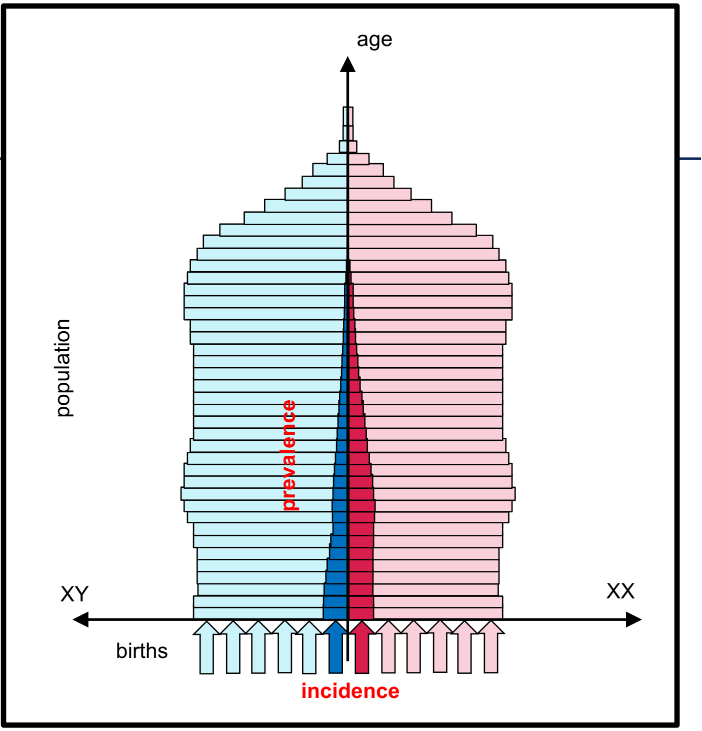
Content

1. Background
2. Challenges
3. Method
4. Results
5. Discussion

Content

1. **Background**
2. Challenges
3. Method
4. Results
5. Discussion

1. Background



Prevalence and Incidence

Incidence = fraction of newborns affected by the disease

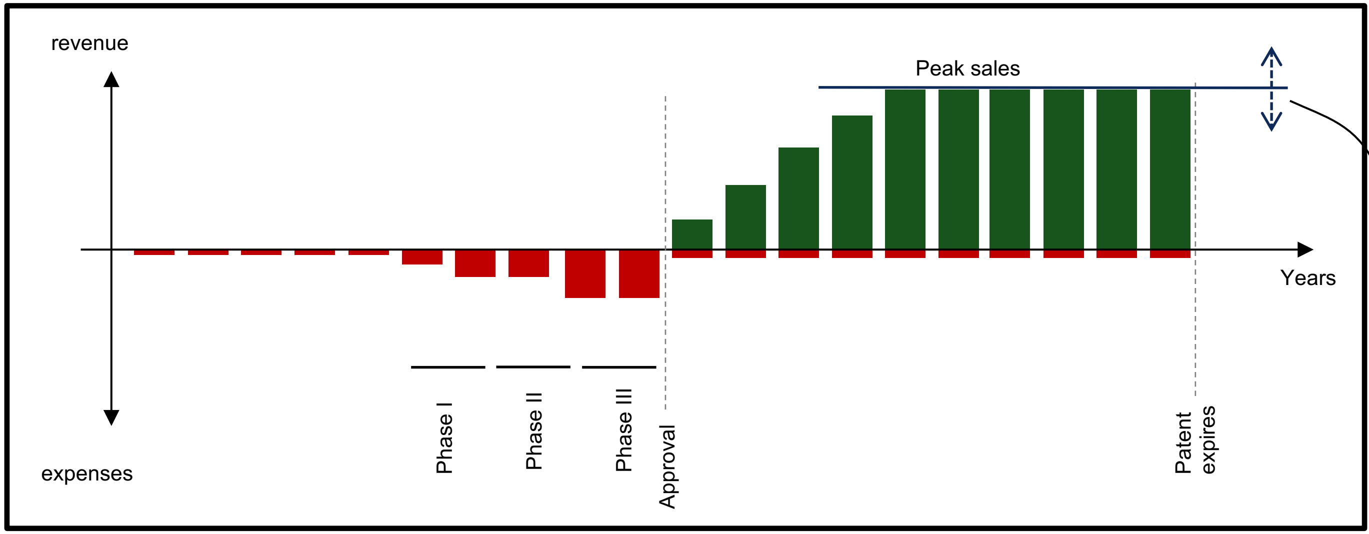
Prevalence = fraction of people affected by the disease

Net present value modeling

Prevalence ∝ peak sales ∝ revenue

Risk-adjusted net present value:

$$\text{rNPV} = \sum_{\text{time}} \text{Prob(success)} \times (\text{revenue} - \text{expenses}) \times \text{discount}$$



Quantifying the uncertainty in prevalence translates to lower and upper bounds on rNPV.

$$\text{rNPV}^{+\Delta_{\text{up}}}_{-\Delta_{\text{down}}}$$

Content

1. Background

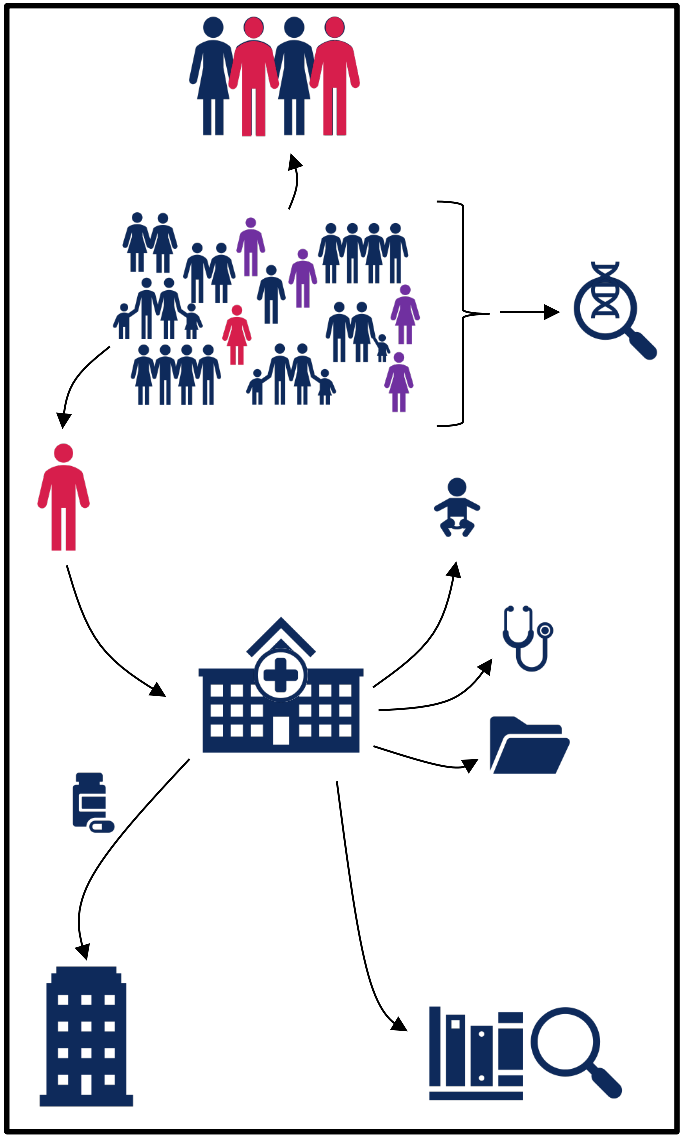
2. Challenges

3. Method

4. Results

5. Discussion

2. Challenges



Source	Advantage	Disadvantage
Patient societies	Few false positives	incomplete
Population genetics	High sensitivity	Works only for hereditary diseases
Newborn screens	Most accurate	Covers only few diseases, missing new genes
Physician interviews	Accurate patient count	Unknown size of ”covered population”
Health records	Detailed phenotypes	Incomplete, missing new diseases
Claims data	Large population	Incomplete, focused on billing (not diagnosis)
Published epidemiology studies	Carefully curated	High effort, usually for specific disease within certain geographic region and time

2.1 Challenges of population genetics approach

Allele frequencies of the top causative variants can differ widely between people from different ethnicity. To account for this bias, we need to estimate ethnicity-specific allele frequencies. To deal with the low sample size, we employ a random-effect model that benefits from partial pooling of different ethnicities.

Basic idea: **incidence**
= **variant pathogenicity** x **variant frequency**

Pathogenicity:

1. Large number of variants with uncertain significance

2. New genes have many fewer publications, therefore seemingly less evidence

3. Quantifying penetrance is impossible for rare variants. Discrete pathogenicity classes don't translate to penetrance.

4. Pathogenicity is different for different isoforms, and often it is not the canonical isoform that is most expressed in the relevant tissue.

variants	ClinVar	HGMD	VarSome
v1	Pathogenic	DM?	Uncertain Significance
v2	Likely Pathogenic	DM	Likely Benign
v3	Likely Benign	(missing)	Uncertain Significance
v4	Pathogenic	(missing)	Likely Pathogenic

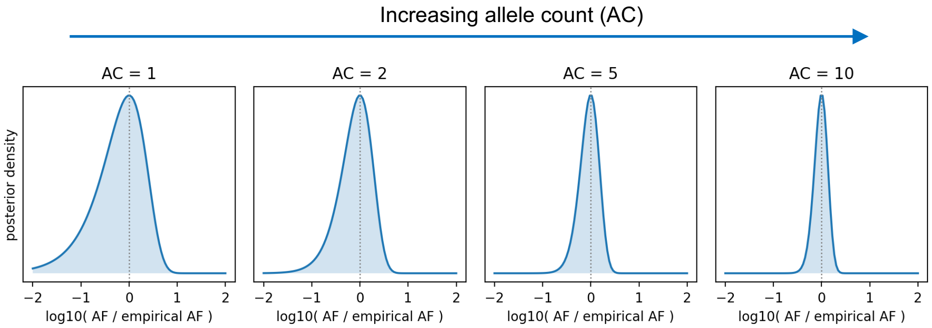
Frequency:

1. Allele frequencies (and prevalence) can vary widely between different ethnic groups (e.g. Sickle-cell anemia)

2. Population genetics databases may
   - Oversample the majority ethnicity, or
   - Oversample large minority ethnicities (if striving for genetic diversity)
   - Even in the best case, can be representative of at most one of many geographical regions with different genetic ancestry

-> Using the sample "global" allele frequency is biased.

3. Certain ethnic groups may be under-sampled
-> Empirical allele frequencies are noisy.



AF 95%CI: [–97%, +270%]

AF 95%CI: [–87%, +180%]

AF 95%CI: [–68%, +105%]

AF 95%CI: [–52%, +71%]

Content

1. Background
2. Challenges
3. **Method**
4. Results
5. Discussion

3. Method

First, we limit the variants to those with confidently pathogenic minor alleles and with associations with the disease in question. To do this, we filter based on functional annotation from multiple sources and tune the method to be accurate for diseases with known incidence.

gnomAD, TOPMed, UKB

HGMD, ClinVar, VarSome, VEP

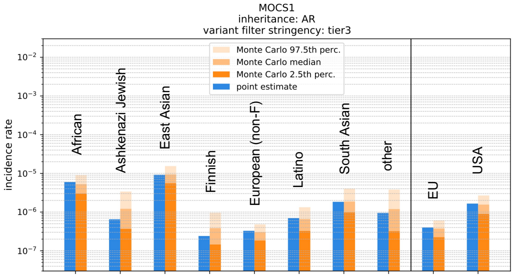
USA, EU demographics

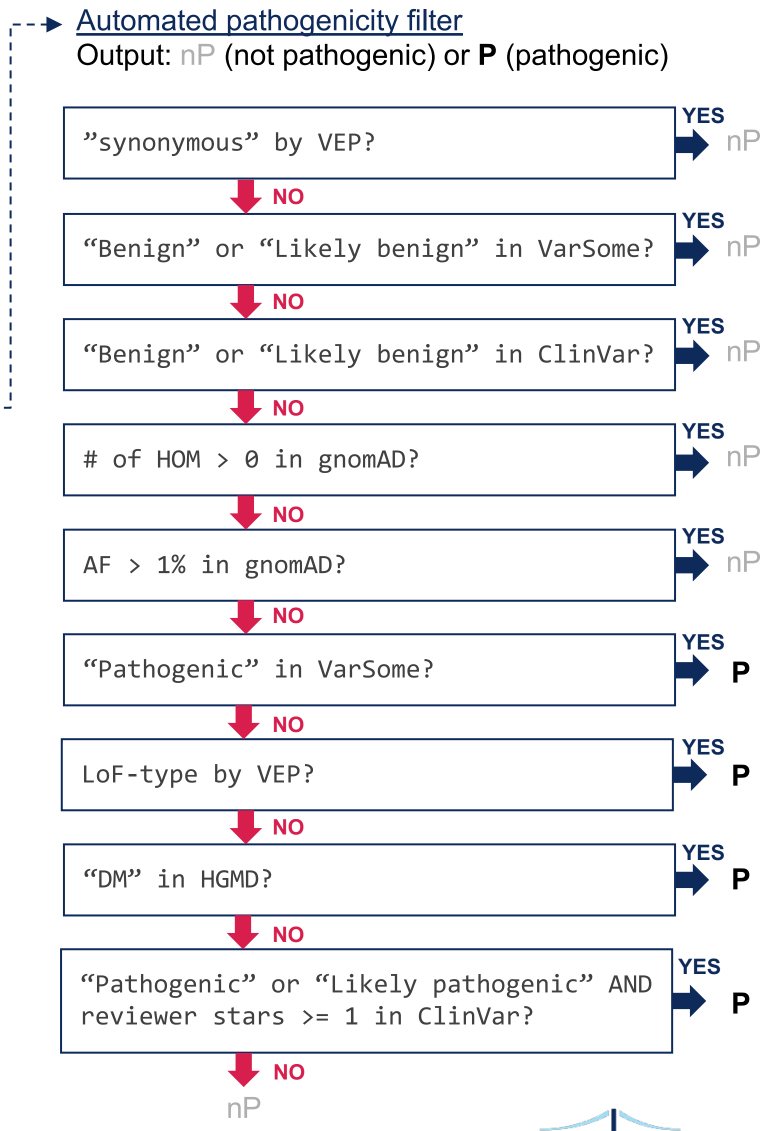
1. Join all variants in the gene region (gnomAD, TOPMed, UKB) with pathogenicity data (ClinVar, HGMD, VarSome, VEP)
2. Select only variants with the relevant phenotypes
3. **Select pathogenic variants using pathogenicity filters**
4. Estimate ancestry-specific carrier frequencies with Monte Carlo algorithm
5. Compute incidence based on inheritance pattern
6. Weight ancestry-specific number with regional ethnicity fractions to get regional incidence (USA and EU numbers)

- incidence rates and
- # of new annual cases

for

- regions: USA, EU
- broken down by ethnicities





Automated pathogenicity filter

Output: nP (not pathogenic) or **P** (pathogenic)

3.1 Incidence formulas

Then, we quantify the ethnicity-specific allele frequencies of each variant with their Monte Carlo Markov chain samples drawn from a non-linear random-effect model with binomial outcome and logit-normal prior shared between ethnic groups. Finally, allele frequency samples are combined across variants to produce the posterior of ethnicity-specific frequencies of carriers, assumed to be unaffected, and aggregated to yield region-specific numbers.

🟦 = input data

Population-specific allele frequency (*v:* variant, *p:* population)

$$\mathrm{AF}_{p,v} \leftarrow \text{MCMC samples from binomial model: } \mathrm{AC}_{p,v} \sim \mathrm{Binomial}(\mathrm{AF}_{p,v}, \mathrm{AN}_{p,v})$$

Incidence for each population *p*

$$\mathrm{Incidence}_p = \left[\sum_{v \,\in\, \mathrm{variants}} \mathrm{AF}_{p,v} \right]^2$$

Assuming
1. recessive inheritance
2. complete genetic mixing
3. low carrier frequency

Incidence for geographical region

$$\mathrm{Incidence} = \sum_{p \,\in\, \mathrm{populations}} \mathrm{Incidence}_p \times \mathrm{fraction}_p$$

Assuming
4. No mixing between populations

Prevalence for geographical region

$$\mathrm{Prevalence} = \mathrm{Incidence} \times \text{Life expectancy, if affacted}$$

Assuming
5. constant birth rate
6. constant ethnic composition

3.2 Quantifying the uncertainty

Directly computing with Monte Carlo samples allows one to characterize the uncertainty with straightforward summary statistics and enables propagating it to downstream analyses without losing information.

Source of uncertainty

Pathogenicity:

- Variants often labeled as "Uncertain significance" or "Conflicting interpretation" by one or more databases
- Different levels of evidence of pathogenicity
- Incomplete penetrance

High-frequency variant with uncertain pathogenicity causes large uncertainty in the final incidence estimate.

Frequency:

- Under-sampled ethnicities
- Zero allele counts with low sample size

Incidence among minorities is underestimated and has higher uncertainty.

How we quantify it

1. Use two pathogenic filters

   "Conservative"

   - only SNPs,
   - VEP "LoF" annotation is not enough for pathogenic verdict

   "Liberal"

   - SNPs, MNPs, short InDels
   - VEP "LoF" annotation is enough for pathogenic verdict

2. Construct [low, high] range from their results

1. Fit a Bayesian, non-Gaussian random effect model to the allele count data across ethnicities
2. Quantify credible intervals (CI) of allele frequencies using Markov chain Monte Carlo (MCMC) sampling
3. Quantify CI for the final incidence estimate by directly computing on the MCMC samples of allele frequencies

3.3 Bayesian random effect model for allele counts

For each variant, the ethnicity-specific allele frequencies are estimated using a non-Gaussian random effect model. We use the Markov Chain Monte Carlo engine STAN to sample the Bayesian posterior.

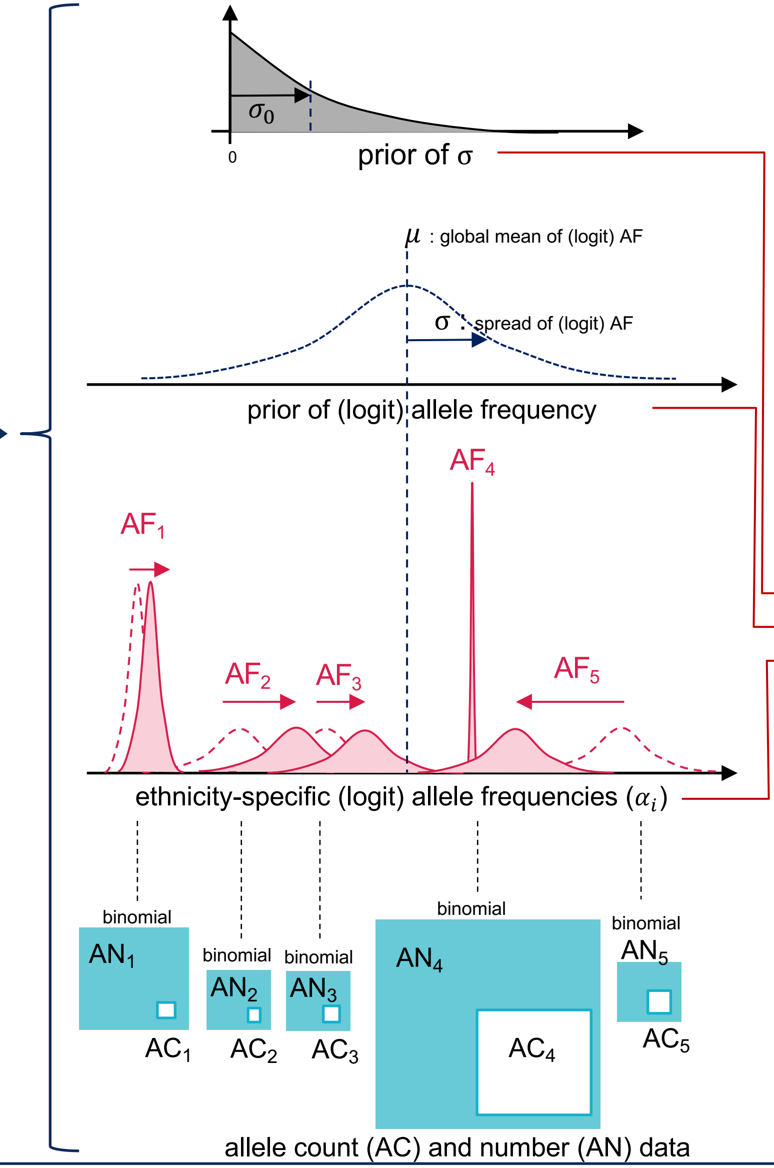
Challenge: Some ethnicities have low allele counts

Assumption: Allele frequencies are similar between ethnicities

Solution: Estimate AF in all ethnicities simultaneously, adaptively borrowing information from estimate for the global AF (Bayesian random effect model for binary data)

Consequence:

- AF > 0 even if allele count (AC) = 0 for any one ethnicity
- Low allele numbers (AN) lead to AF estimates close to the global AF, with high uncertainty
- Ethnicities with high AN are affected marginally, as if they were estimated independently
- Global AF is estimated adaptively, taking the uncertainty of ethnicity-specific AFs into account



MCMC engine (STAN) code

```
//logit-normal-binomial.stan


data {
    int M;
    int AC[M];
    int AN[M];
    real sigma0;
}

parameters {
    real alpha[M];
    real mu;
    real<lower=0> sigma;
}

model {
    sigma ~ exponential(1 / sigma0);
    alpha ~ normal(mu, sigma);
    AC   ~ binomial_logit(AN, alpha);
}

generated quantities {
    real AF[M];
    AF = inv_logit(alpha);
}
```

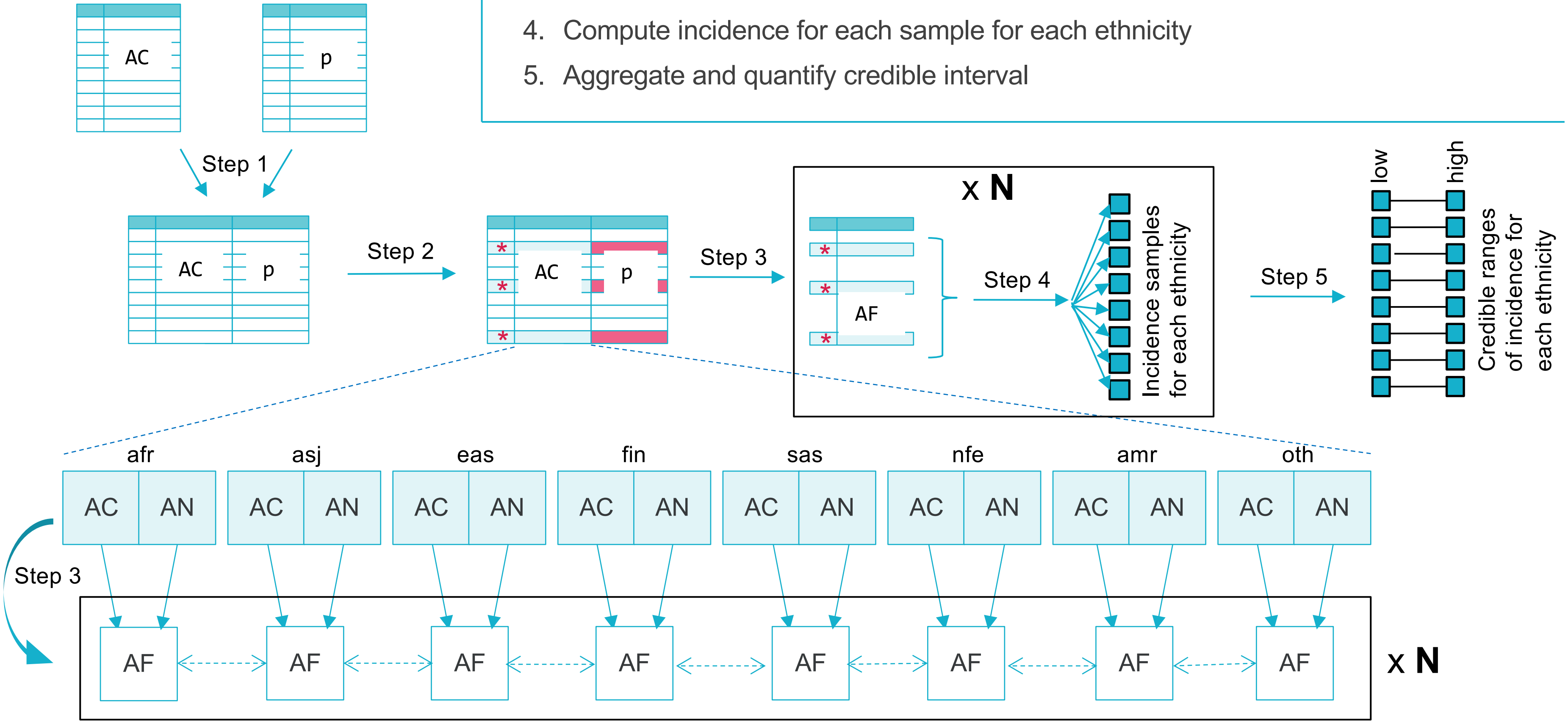
Logit function

transforms (0,1) to (-inf, + inf)

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

3.4. Overview of incidence estimation - Illustration of algorithm

1. Merge allele count (AC) and pathogenicity (p) data
2. Select pathogenic variants
3. Fit the logit-normal-binomial random effect model to the allele counts (AC) and numbers (AN) of each variant separately, and obtain MCMC sample
4. Compute incidence for each sample for each ethnicity
5. Aggregate and quantify credible interval



Content

1. Background

2. Challenges

3. Method

4. Results

4.1. Validation

4.2. RDEB

5. Discussion

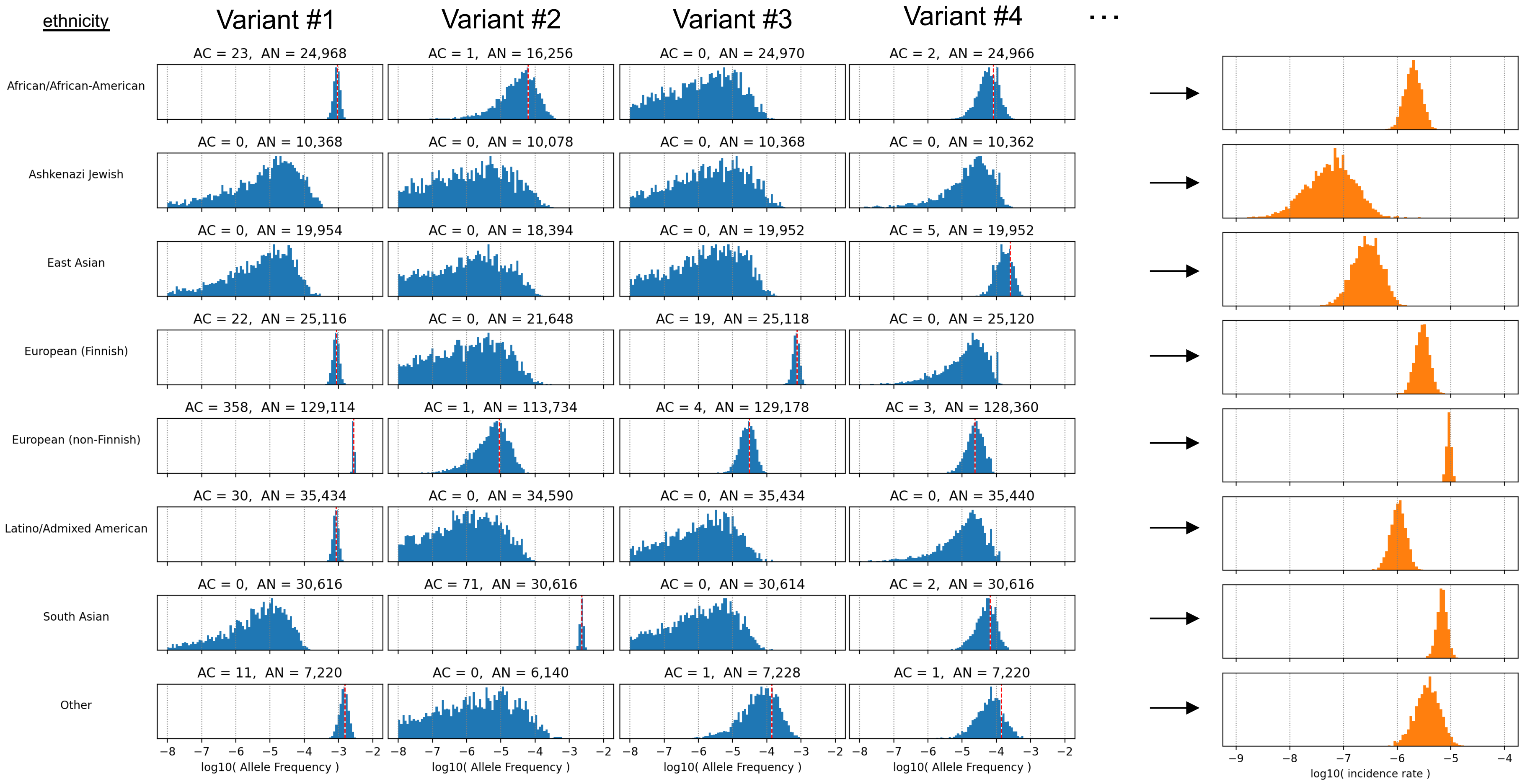
4.1.1. Validation: Example output

MCMC samples of allele frequencies (AF) of the top 4 pathogenic variants of a sample gene

🟦 MCMC samples of AF

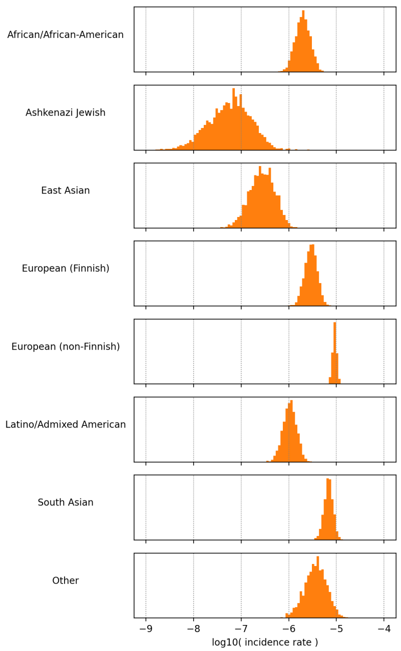
- - - - - empirical AF

🟧 MCMC samples of incidence



4.1.2. Validation: Integrating across ethnic groups

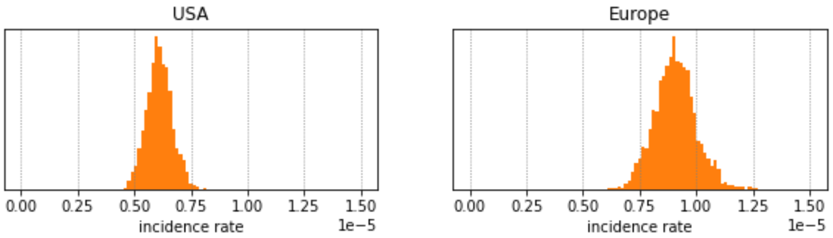
Ethnicity-specific incidence



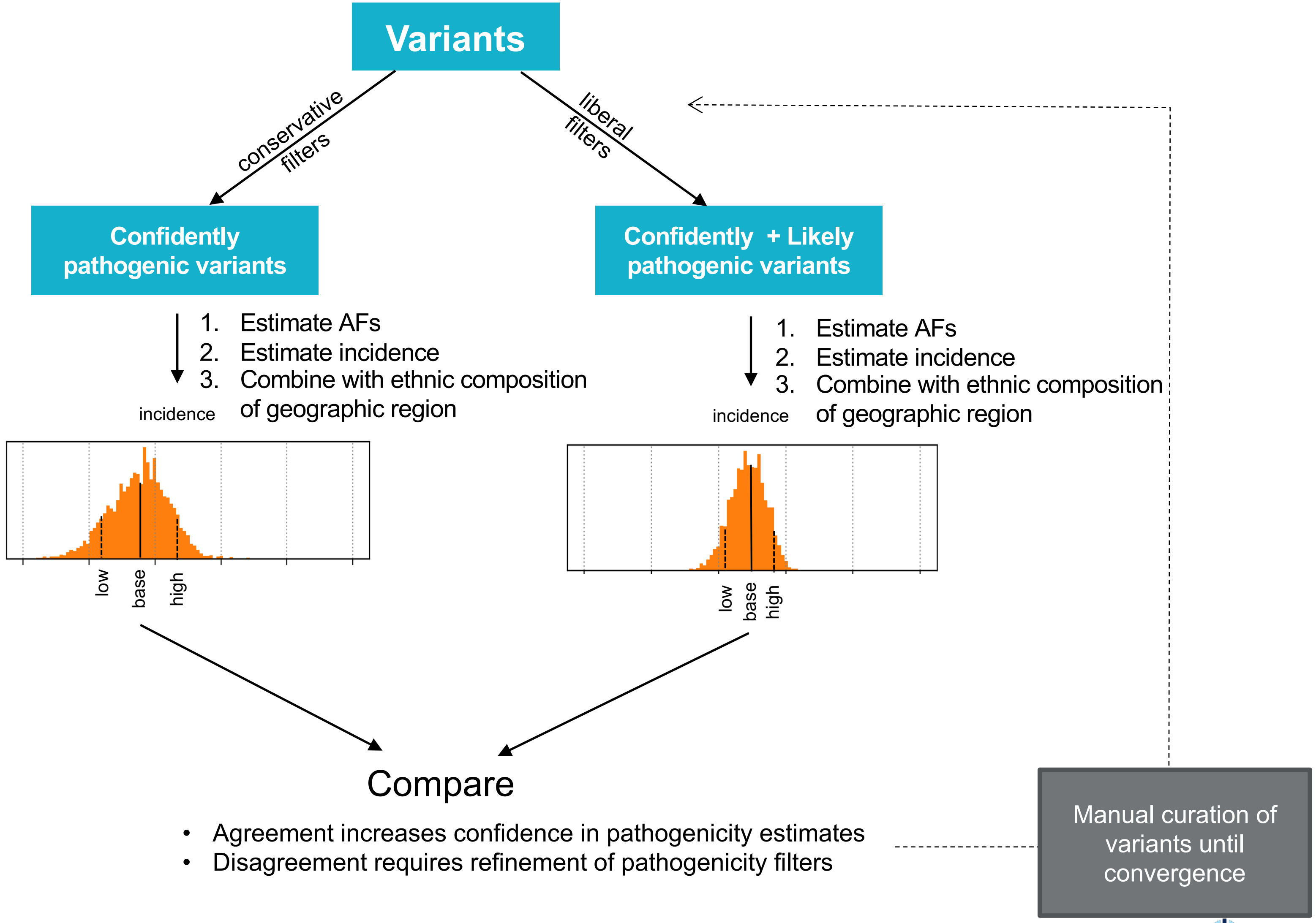
Region-specific ethnicity composition

	USA	EU
x	13.2%	0.7%
x	1.8%	0.2%
x	5.3%	0.2%
x	0%	0.7%
x	59.4%	97.3%
x	18.3%	0.3%
x	2.0%	0.5%
x	0%	0%

↓ Σ ↓ Σ



4.1.3 Validation: Different pathogenicity stringency levels



Variants

Confidently pathogenic variants

1. Estimate AFs
2. Estimate incidence
3. Combine with ethnic composition of geographic region

Confidently + Likely pathogenic variants

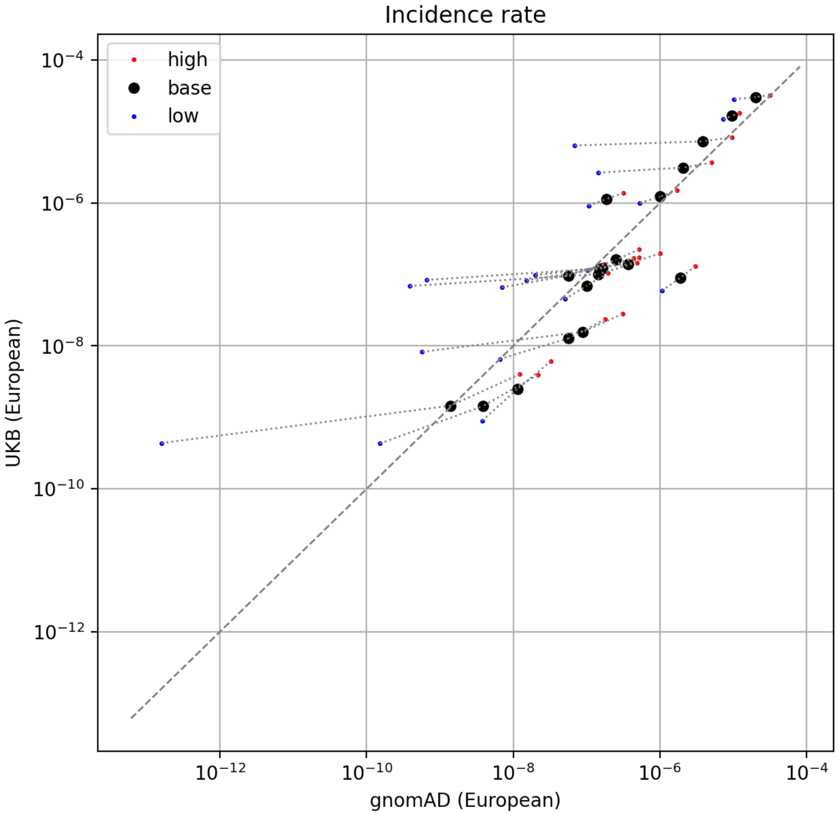
1. Estimate AFs
2. Estimate incidence
3. Combine with ethnic composition of geographic region

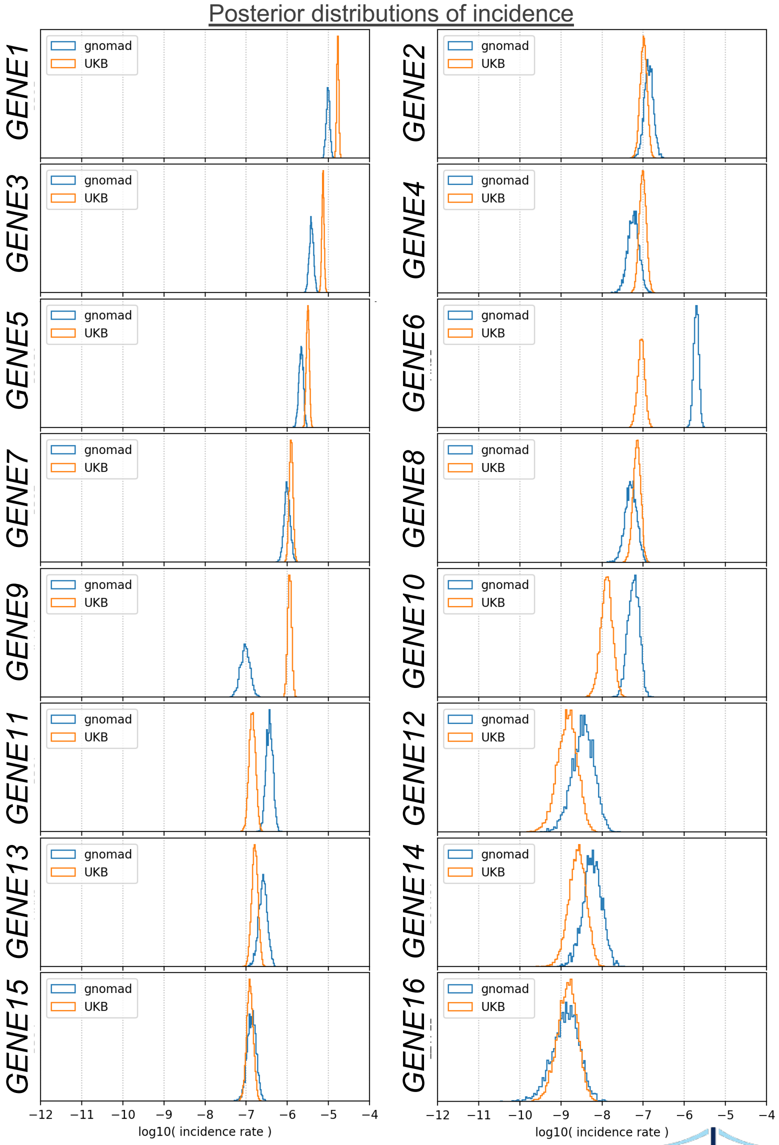
Compare

- Agreement increases confidence in pathogenicity estimates
- Disagreement requires refinement of pathogenicity filters

Manual curation of variants until convergence

4.1.4. Validation: Comparison between different population databases (UK Biobank vs gnomAD)





- The results on 16 genes were largely concordant between gnomAD and UK Biobank.
- However, this comparison also highlighted genes that suggested orders of magnitude differences that are worth further calibration.

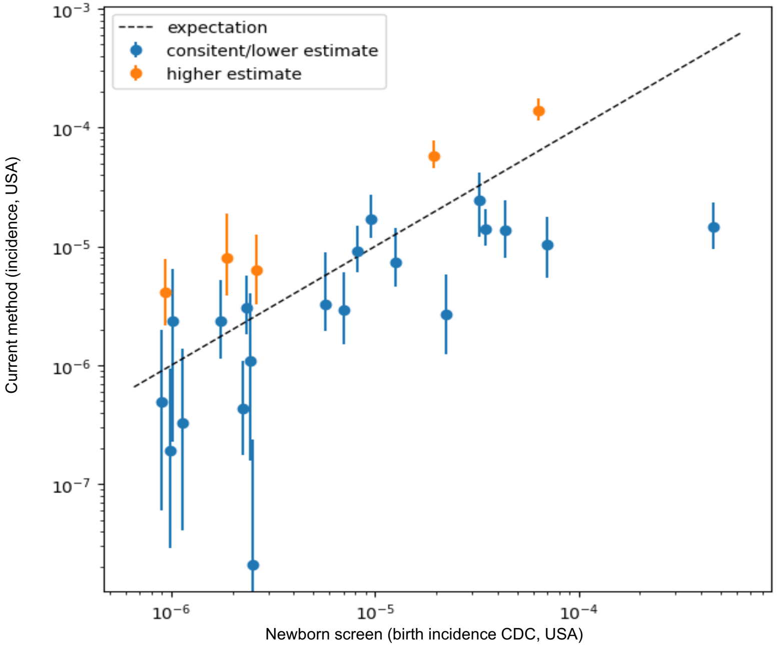
4.1.5. Validation: Incidence estimation

High-confidence newborn screens

- Results from automatic estimation (no phenotype filtering) match results from new-born screen qualitatively

- 14 out of 24 genes were quantitatively discrepant between the two methods
  - for 5 genes, CompGen > NBS
  - Differences in incidence estimates could mostly be explained by manual curation of variants and phenotypes included in the algorithm
  - One such example was phenylketonuria (PKU) where our method included both mild and classical PKU

- These results suggest that although a default run can provide reliable estimates for some diseases (58%), manual curation of the input data is critical



Content

1. Background

2. Challenges

3. Method

4. Results

4.1. Validation

4.2. RDEB

5. Discussion

4.2.1 Background on Recessive Epidermolysis Bullosa (RDEB)

- Dystrophic epidermolysis bullosa (DEB) is a rare genodermatosis due to mutations in the *COL7A1* gene encoding the alpha-chain of collagen 7 (C7). C7 deficiency results in dermalepidermal junction separation with severe, painful blistering and scarring. Both dominant (DDEB) and recessive (RDEB) forms occur with RDEB being typically more severe.
- Multiple subtypes of RDEB exists that may be treated using a protein replacement therapy:
  - Generalized, severe RDEB is the most severe form of RDEB
  - Other RDEB, which includes
    - Intermediate-form RDEB, RDEB inversa, RDEB pruriginosa, Localized RDEB, Acral RDEB, Nails only RDEB, Pretibial RDEB, Self-improving RDEB
- Epidemiology estimates, from Fine et al. (2016):
  - Prevalence:
    - RDEB, generalized severe: 0.36 / 1M people (26%)
    - RDEB, other: 0.99 / 1 M people (73%)
  - Incidence:
    - RDEB, generalized severe: 0.57 / 1M births (18.7%)
    - RDEB, other: 2.48 / 1M births (81.3%)
- We tested the incidence estimation algorithm to triangulate potential treatable population of a protein replacement therapy.

4.2.2 Phenotype curation for RDEB

Manual curation of phenotypes in ClinVar and HGMD was conducted to select only the most relevant ones. This was an iterative efforts with multiple feedback from clinical scientists and experts in RDEB.

```
selected_phenotypes = [
   "Abnormal blistering of the skin",
   "Bullous lesions",
   "Dominant dystrophic epidermolysis bullosa with absence of skin",
   "Dystrophic epidermolysis bullosa",
   "Epidermolysis Bullosa Distrophica Autosomal Recessive (RDEB)",
   "Epidermolysis bullosa",
   "Epidermolysis bullosa dystrophica",
   "Epidermolysis bullosa dystrophica inversa, autosomal recessive",
   "Epidermolysis bullosa dystrophica with amniotic band syndrome",
   "Epidermolysis bullosa dystrophica, Pasini type",
   "Epidermolysis bullosa dystrophica, autosomal recessive, localisata variant",
   "Epidermolysis bullosa dystrophica, intermediate",
   "Epidermolysis bullosa dystrophica, inversus type",
   "Epidermolysis bullosa dystrophica, nails only",
   "Epidermolysis bullosa dystrophica, pretibial",
   "Epidermolysis bullosa dystrophica, recessive",
   "Epidermolysis bullosa dystrophica, recessive, intermediate",
   "Epidermolysis bullosa dystrophica, recessive, localised",
   "Epidermolysis bullosa dystrophica, recessive, pruriginosa",
   "Epidermolysis bullosa dystrophica, recessive, self-improving",
   “Epidermolysis bullosa pruriginosa",
   "Epidermolysis bullosa pruriginosa, autosomal dominant",
   "Epidermolysis bullosa pruriginosa, autosomal recessive",
   "Epidermolysis bullosa, pretibial, autosomal recessive",
   "Generalized dominant dystrophic epidermolysis bullosa",
   "Recessive dystrophic epidermolysis bullosa",
]
```

```
excluded_phenotypes = [
   "Abnormality of the skin",
   "Abnormality of the thyroid gland",
   "Anonychia",
   "Autism spectrum disorder",
   "Bart syndrome",
   "Barts syndrome",
   "Bladder Urothelial Carcinoma",
   "Brain Lower Grade Glioma",
   "Breast cancer",
   "Bullous dermolysis of the newborn",
   "Bullous ichthyosiform erythroderma",
   "COL7A1-related disorders",
   "COL7A1-related epidermolysis bullosa",
   "Cerebral palsy, modifier of",
   "Conotruncal heart defects",
   "Ductal breast carcinoma",
   "Epidermal nevus",
   "Epidermolysis bullosa dystrophica, self-improving",
   "Finger syndactyly",
   "Ichthyosis (disease)",
   "Inborn genetic diseases",
   "Inflammatory bowel disease",
   "Liver hepatocellular carcinoma",
   "Lung squamous cell carcinoma",
   "Nail disorder, nonsyndromic congenital, 8",
   "Nail dystrophy",
   "Palmoplantar blistering",
   "Pancreatic adenocarcinoma",
   "Persistent cloaca",
   "Pretibial epidermolysis bullosa",
   "Short stature",
   "Skin erosion",
   "Skin fragility with non-scarring blistering",
   "Toe syndactyly",
   "Transient bullous dermolysis",
   "Transient bullous dermolysis of the newborn",
   "Uveitis",
]
```

4.2.3 RDEB birth incidence estimate

- Only "Pathogenic" or "Likely Pathogenic” variants in all databases considered were included in the calculation
- pLoF variants were included in the ”liberal” estimates when computing the higher bound of the interval range
- Results

population	new annual cases (low)	new annual cases (base)	new annual cases (high)
Europe	13.8	23.25	36.2
USA	10.6	20.4	36.5

population	incidence rate (low)	incidence rate (base)	incidence rate (high)
Europe	3.26e-06	5.48e-06	8.53e-06
USA	2.79e-06	5.38e-06	9.63e-06

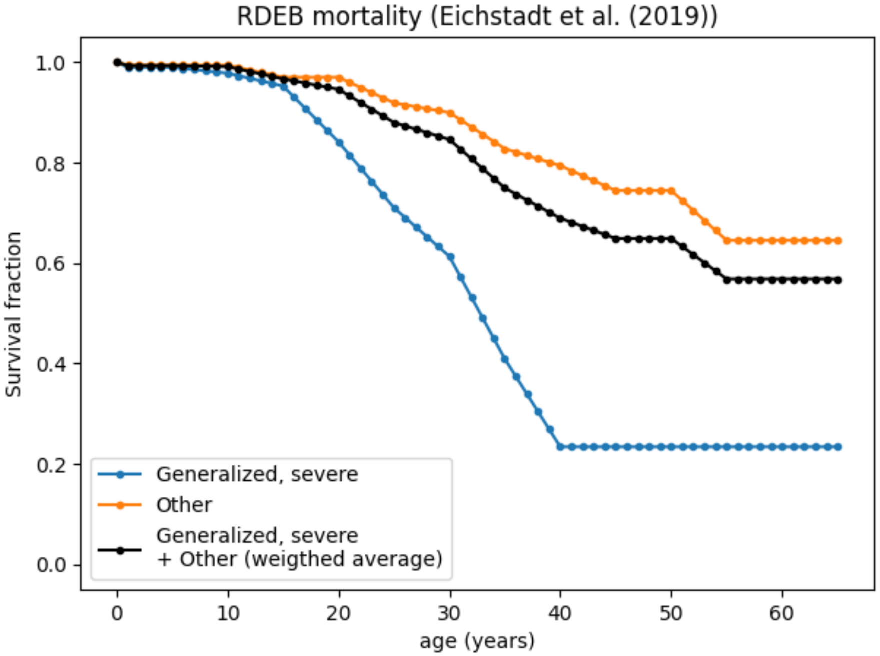
- Conclusion: Overall, the estimate is higher than literature estimate (0.57 + 2.48) / 1M = 3.05 / 1M, but the 95%-credible interval [2.79, 9.63] / 1M contains the literature estimate.

Source of literature estimate: Fine et al. (2016)

4.2.4 RDEB overall and pediatric prevalence estimates

Eichstadt et al. (2019), Table 2. Interpolated for each year. Life expectancy:

- RDEB Generalized Severe: 37.0 years
- RDEB Other: 55.7 years
- RDEB (GS + Other, weighted average, with weights 18.7%, 81.3%): 52.2 years



interpolated for every year

Total number of pediatric patients

- Life-expectancy times incidence rate yields the total number of patients. Multiplied by the pediatric fraction yields the number of patients under 19.

RDEB (in US)

USA	RDEB base [low, high]
Incidence (per million births)	5.38 [2.79, 9.63]
Newborns per year	20.4 [10.6, 36.5]
Prevalence (per million)	3.6 [1.8, 6.4]
Patients	1065 [553, 1905]
Patients (age <= 18 years)	383 [200, 685]

Validation against data from US claims:

- RDEB + DDEB US patients: 1554 [1042, 2394] (Prognos: 1214)
- RDEB + DDEB US patients under 19 years old: 525 [342, 827] (Prognos: 560)

Content

1. Background
2. Challenges
3. Method
4. Results
5. **Discussion**

Discussion

Key assumptions of the model

- Well-defined genetic cause
- Population of each ethnicity is well-mixed
- Pathogenic variants are inherited
- Carriers are found in general population
  with high enough frequency (> 1 in 200,000)
- HET genotype have little effect
- Pathogenic variants have high penetrance

Limitations of the model

- Variants with incomplete, low penetrance
- Modifiers
- Polygenic or environmental effects
- AD and X-linked diseases
- De-novo mutations
- Consanguinity
- Founder effects