

THE LANCET

Digital Health

Supplementary appendix 1

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Thygesen JH, Zhang H, Issa H, et al. Prevalence and demographics of 331 rare diseases and associated COVID-19-related mortality among 58 million individuals: a nationwide retrospective observational study. *Lancet Digit Health* 2025; 7: e144–56.

Supplement for: A nationwide study of 331 rare diseases among 58 million individuals: prevalence, demographics and COVID-19 outcomes

Supplementary methods	3
Data sources and data quality	3
Primary care	3
Hospitalisation	3
Mortality	4
Data Linkage	4
Comparison with population estimates	4
Demographic data	4
Identification of participants with rare-diseases	5
COVID-19 phenotyping	5
Rare disease COVID-19 analysis	6
COVID-19-related mortality risk analysis	6
Statistical Analysis for COVID-related mortality	6
Comparison of risk of COVID-19-related deaths and high-risk labels	7
Supplementary Figures	7
Supplementary figure 1 - Data processing diagram:	7
Supplementary figure 2 - Comparing prevalence between diagnosis systems:	9
Supplementary figure 3 - Comparing adjusted and raw prevalence estimates:	9
Supplementary figure 4 - Forest plots:	10
Supplementary figure 5 - COVID-19 mortality in relation to age at first diagnosis and prevalence	12
Supplementary data	13
Supplementary data 1 - Rare disease demographics:	13
Supplementary data 2 - Gender and ethnicity differences:	13
Supplementary data 3 - COVID-19 mortality for individual rare diseases:	13
Supplementary data 4 - COVID19 mortality analysis for rare disease categories:	13
Supplementary data 5 - All rare diseases - Kaplan-Meier Plot:	13
Supplementary tables	14
Supplementary table 1 - Rare disease information by category:	14
Supplementary table 2 - Overview of Orphanet data sources for prevalence	15
Supplementary table 3 - STROBE statement:	16

Supplementary data 5 figures

19

References

45

Supplementary methods

Data sources and data quality

Primary care

Data in primary care have been sourced from the General Practice Extraction Service (GPES) Data for pandemic planning and research (GDPPR) system which contains SNOMED-CT concepts for patients registered with a primary care physician in the UK. The dataset contains approximately 96% of the English populations and 98% of all English general practices. Patients records were included in GDPPR when they had coded information matching any of the SNOMED-CT concepts in the Code Clusters applicable for COVID-19 planning during primary care consultations ¹. Around 34,000 unique SNOMED-CT concepts are included (>90% of all those currently extracted for a wide range of purposes by NHS Digital's GP Extraction Service), covering a broad range of diagnoses and procedures (from the start of each person's records) along with laboratory results, physical measurements, clinical referrals, and prescriptions. Primary care EHR have been shown to have a high degree of diagnostic accuracy in validation studies ²

Hospitalisation

Hospital Episode Statistics (HES) contains administrative data from English hospitals in the National Health Service (NHS). HES captures a) inpatient episodes (including maternity), b) outpatient episodes, c) accidents and emergency attendance (A&E), d) critical care, and e) adult mental health. The primary purpose of HES is to facilitate hospital reimbursement which is actioned through a framework called "Payments by Results"³.

HES captures the records of all patients, and their interactions, if they are funded by the NHS irrespective of if they are UK residents or if the care was delivered by an NHS provider. HES record-level data are structured into spells (admissions) which in turn are composed of one or more episodes (most admissions have a single episode)⁴. An episode can be defined as a period of continuous care from a single consultant/speciality and HES contains a row per episode per admission. A spell is terminated when a patient is discharged or dies. HES inpatient data are recorded using WHO ICD-10 and procedures using the Office of Population Censuses and Survey's (OPCS) version 4 clinical classification. A spell can have up to 20 primary and secondary diagnoses or procedures recorded. A primary diagnosis in HES is defined as the main condition treated (or investigated) during the episode of care or where no such definitive diagnosis exists, the main symptom or abnormal finding observed.

HES data are published annually but data are collected through a mechanism known as the Secondary Uses Service (SUS) which curates monthly data extracts from healthcare providers. These monthly extracts are then subsequently used to populate the HES database. While variation is likely to exist between healthcare providers, coded data from hospitalisations (through Hospital Episode Statistics Admitted Patient Care and Secondary Uses Service) have been shown to be robust: median diagnostic accuracy 80.3% (IQR: 63.3-94.1%) and median procedure accuracy of 84.2% (IQR: 68.7-88.7%)⁵. HES and SUS undergo robust data quality controls and validation rules which are further described along with the data processing pipeline elsewhere⁶.

Mortality

In England (and Wales), when a patient dies, it is the statutory duty of the doctor who had attended in the last illness to issue the death certificate and the Office of National Statistics (ONS) centrally collects and curates all deaths. ONS mortality statistics are considered to be the gold standard for death ascertainment and are routinely used in EHR studies to ascertain deaths. During the pandemic however, there were a variety of changes to the processes in which deaths were certified and registered. For example, the time taken for deaths to be registered decreased while the numbers of conditions recorded on the death certificate were greater for deaths involving COVID-19 than those not involving COVID-19, suggesting higher rates of comorbidities in these deaths and good quality of the certification.⁷

Data Linkage

Individual data sources are linked by NHS Digital using the Master Person Service in combination with the Personal Demographics Service. A linkage score is calculated by cross-referencing information across different sources with the demographics in the Personal Demographics Service and signifies the overall associated match confidence. This score is not directly made available to researchers but NHS Digital's monthly reports for data quality maturity index indicate that 97-100% of records submitted to NHS Digital each month include information on NHS number and other key personal variables, providing confidence in the accuracy of the matching process⁸.

Comparison with population estimates

The dataset comprises more than 96% coverage of the English population and represents the English population in terms of age, sex and ethnicity when compared with UK government official statistics for England, it includes a representative distribution of general practices according to geographical location and size. The datasets and their underlying characteristics as described in detail elsewhere⁹.

None of the datasets included patients that have explicitly opted out of their EHR being used for medical research. These are referred to as "Type 1 opt-outs" and as of Sept 1 2021 there were 3,264,327 national data opt-outs¹⁰. Lastly, this is a dynamic cohort where new patients can enter (at birth or new registration with a general practice) during the study period. For example, there are approximately 1 million migrations and student visitors in the UK yearly that are likely to register with a general practitioner but not be adequately captured in the ONS population estimates¹¹. As a result, the total number of participants is a cumulative estimate over the study period and does not entirely align with national population estimates¹².

Demographic data

Date of birth, sex and ethnicity were extracted from GDPPR and HES datasets. Age at first RD diagnosis and COVID-19 was calculated as the difference between date of birth and the relevant first RD diagnostic code and/or the first COVID event divided by 365.25. Ethnicity was categorised as per the ONS²⁶: Asian or Asian British, Black or Black British, mixed, white, other ethnicities and unknown, mapping ethnicity across primary and secondary care prioritising information from primary care²⁷.

Identification of participants with rare-diseases

Orphanet¹³, an extensive online resource for RDs, was used to identify and define RDs. Its rare disease alignment (from Orphanet code to ICD-10 or SNOMED-CT codes), which is managed by a dedicated information scientist at Orphanet^{14,15} was downloaded on the 6th of May 2022. We adopted a stepwise approach to identify RDs that could be accurately mapped to the disease codes available in our data sources. First, we extracted all Orphanet diseases with mappings to ICD-10 or SNOMED-CT (all mapping types included; n = 7,697), the clinical terminologies used in our data sources HES-APC and HES-OP, and GDPPR, respectively. For any rare disease which was mapped to ICD-10 code with fewer than four digits, we automatically expanded to include all sub-concepts. New SNOMED-CT codes were created for RDs in Orphanet terminology mapping exercise.¹⁴ Adaptation of newly defined SNOMED-CT codes in clinical settings can take time, and it is therefore, not surprising that many of these codes were not found used in our EHR data. While many Orphanet listed SNOMED-CT codes are not used in our data sources, those we did find were of high quality with good specificity for identifying RDs according to our manual validation.

Second, to select unique clinical RDs (defined as ‘disorders’ in the Orphanet classification)¹⁶, we applied the disease classification filter available on Orphanet to filter diseases to those included in ‘disorder types’: disease, morphological anomaly, malformation syndrome, or clinical syndrome (n = 5,864). The two largest disorder types removed were *Category* (1,719) and *Clinical subtypes* (769). Third, we excluded all diseases for which mapping to ICD10 was classified as “narrow to broad”, since our initial manual validation revealed that most if not all of these codes included a combination of rare and common diseases and hence would result in a high degree of misclassification (n = 3,384 diseases with ICD-10 mappings were removed, note that some of these diseases also had SNOMED-CT code mappings which were still included).

Fourth, we computed the point prevalence for all the eligible diseases in the TRE (n = 3,817 diseases with at least one affected individual), searching across the linked tables for SNOMED-CT codes in GDPPR, and ICD-10 codes in HES-APC and HES-OP. Fifth, manual curation was performed by a clinician (ACPG) to validate the accuracy of the matching of diseases to SNOMED-CT and ICD-10 codes by considering both the code definitions and the code frequencies from the EHR data. These curated results were further checked by another clinician (TH). All SNOMED-CT codes were specific to the diseases but some ICD-10 codes were still too broad and introduced misclassification as they matched to both rare and common diseases (e.g., familial Parkinson’s disease mapped to Parkinson’s disease). This resulted in the inclusion of 331 RDs with highly specific mapping, of which 164 had ICD-10 mappings, 140 had SNOMED-CT mappings and 27 had both ([figure S1](#)). Participants were defined as having a rare disease if they had one or more codes of each disease’s ICD-10 or SNOMED-CT codes.

COVID-19 phenotyping

To determine the impact of COVID-19 on RD patients we used five previously defined COVID-19 phenotypes¹⁷ In brief we identified 1) positive SARS-CoV-2 tests, 2) COVID-19 diagnosis recorded in primary care, 3) hospital admissions with a COVID-19 diagnosis, 4) ventilatory support related to COVID-19, and 5) COVID-19 mortality; including (a) suspected or confirmed COVID-19 diagnosis with ICD-10 term listed anywhere on the death certificate, (b) death within 28 days of the first recorded COVID-19 event, where a COVID-19 diagnosis was not listed anywhere on the death certificate, or (c) a COVID-19

hospital admission with a discharge method or discharge destination denoting death, irrespective of cause and duration after the index event. These COVID-19 phenotypes were identified using clinical codes and data from the following linked tables; SGSS, GDPPR, SUS, HES-APC, HES-CC, CHES, ONS. Full details of the COVID-19 phenotyping descriptions can be found in the supplementary methods and in the paper by Thygesen et al¹⁷.

Vaccination status was determined from the COVID-19 vaccination dataset, including all vaccinations administered after Dec 12, 2020 (when the first official dose was administered in England). Patients were classified as fully vaccinated once 14 days had elapsed after their second dose.

Rare disease COVID-19 analysis

COVID-19-related mortality risk analysis

To enable exploration of less prevalent RDs by grouping we adopted the 25 RD categories defined by Orphanet, that allows for diseases to belong to multiple categories. We performed a retrospective cohort analysis comparing COVID-19 related mortality in people with a certain RD or a RD category, with matched controls from the general population. Differences in COVID-19 related mortality were addressed with a time-to-event analysis.

The study period for assessing COVID-19 related mortality was set from 2020-09-01 to 2021-11-30, spanning what is commonly referred to as Wave 2 and the first half of Wave 3 of the pandemic in the UK, driven by the original strain/Alpha variant and Delta variant, respectively¹⁸. This period reflects a time of the pandemic after which vaccination efforts were rolled out and the COVID-19 testing capacity remained high. A minimum follow-up for each individual was upheld to ensure sufficient time to observe COVID-19 related deaths within 28 days of an infection, in line with the UK Government and Public Health England reported threshold¹⁹.

Individuals whose date of first COVID event fell within the period analysed for COVID-19 mortality were included. Cohorts were formed for each RD, using exact matching on age group, sex, ethnicity and vaccination at the ratio of two controls (not affected by the condition) per individual with a RD.

Analysis by disease categories was done with the identical setting. Analysis by disease categories allows for inclusion of rare-disease individuals affected by a disease too underpowered to be studied on their own. We used the 25 disease categories defined by Orphanet. Cohort formation and matching method was identical with the process for single diseases. An individual could be mapped to multiple disease categories either by being affected by multiple diseases or since one disease can belong to multiple categories.

Statistical Analysis for COVID-related mortality

Survival functions were estimated using Kaplan–Meier estimator. Differences in survival distributions between people with RDs and matched controls were tested with the log-rank test. The hazard ratio and confidence interval was estimated using a univariable Cox proportional hazards model, using time of event as the dependent variable and RD status as the independent variable. Proportional hazards assumption was tested using the Schoenfeld residuals²⁰. Differences with log-rank test p-value < 0.05

and lower-bound of confidence interval of hazard ratios >1 were considered significant increases of risk. To determine which rare-disease/category showed the largest difference in mortality to unaffected individuals, a conservative approach of using the lower-bound confidence interval for the hazard ratio for comparison, was adopted. Comparing the lower-bound confidence interval, rather than the main estimate, ensured the top findings did not have large overlapping confidence intervals between affected and unaffected, due to smaller sample sizes. A tree-plot was created to illustrate the HR and 95%CI for rare diseases and disease groups with significant increase of COVID-19-related mortality.

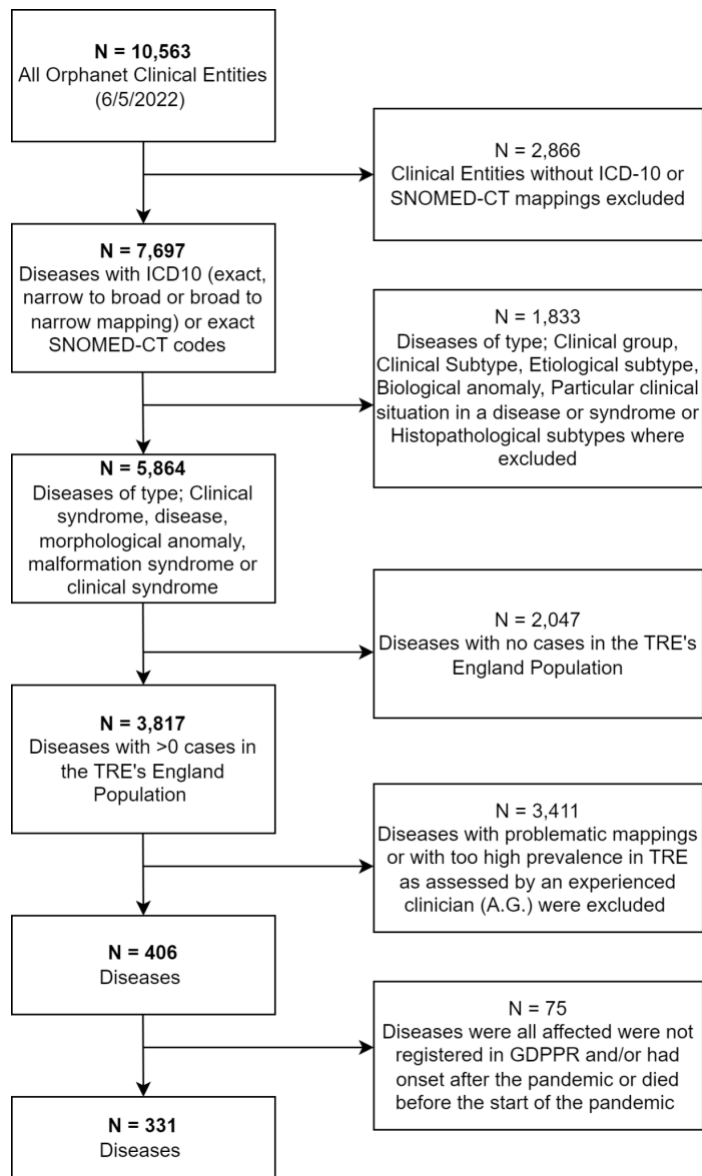
Comparison of risk of COVID-19-related deaths and high-risk labels

Diseases included in the shielding list were obtained from NHS digital²¹. The list of diseases with increased COVID-19 mortality rate in unvaccinated individuals was compared to the shielding list to identify diseases that were missed in the shielding scheme.

Supplementary Figures

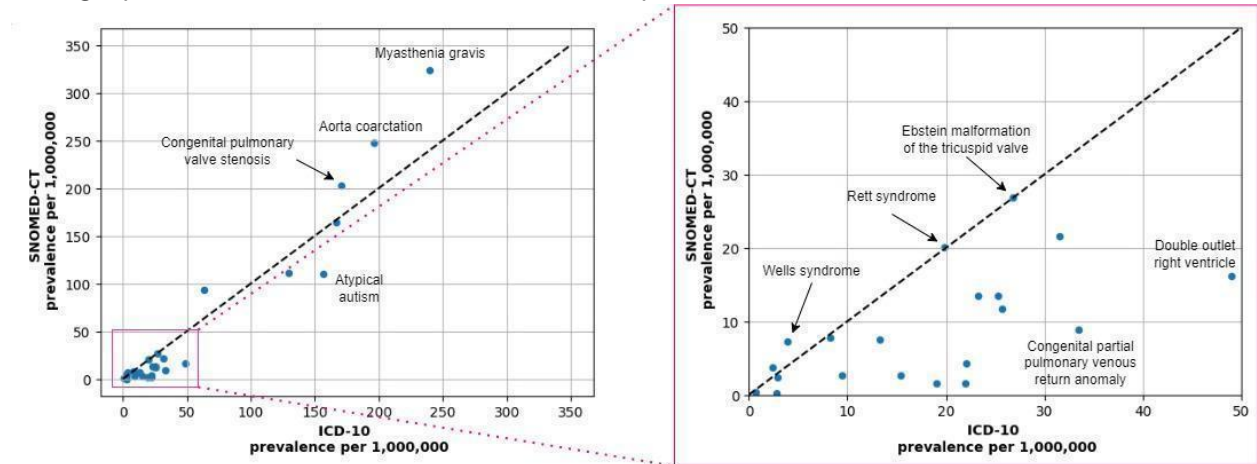
Supplementary figure 1 - Data processing diagram:

The diagram depicts the rare-disease inclusion steps. TRE = Trusted Research Environment), GDPPR = General Practice Extraction Service Extract for Pandemic Planning and Research.



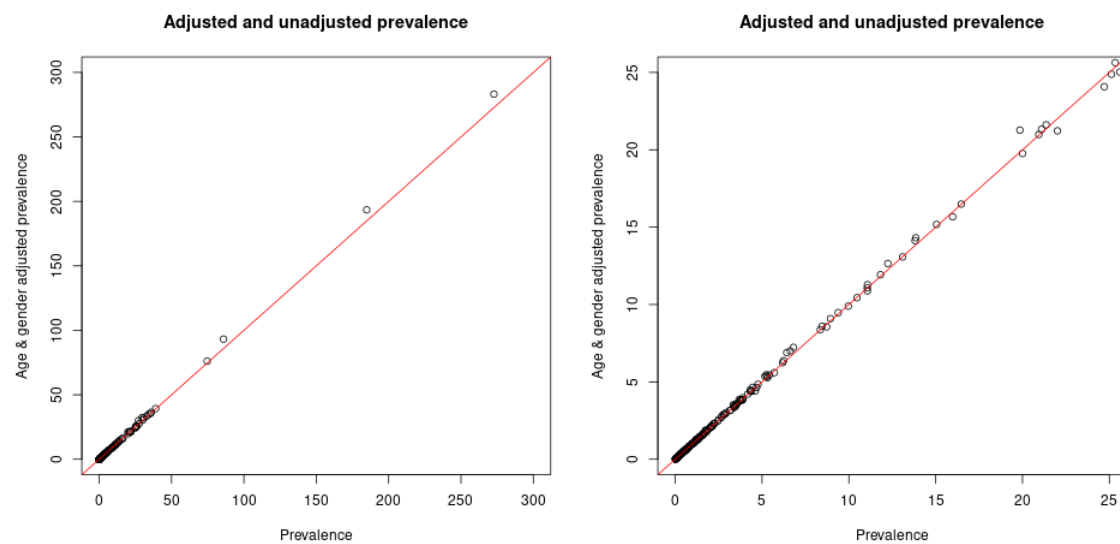
Supplementary figure 2 - Comparing prevalence between diagnosis systems:

Comparison of prevalence for 27 rare-diseases where affected individuals were identified both from primary care sources (SNOMED-CT) and secondary care data (ICD-10). The left panel shows prevalence per 1,000,000 for all 27 conditions, the dashed line indicating similar prevalence between the sources. The right panel is zoomed in and focused on the less prevalent conditions.



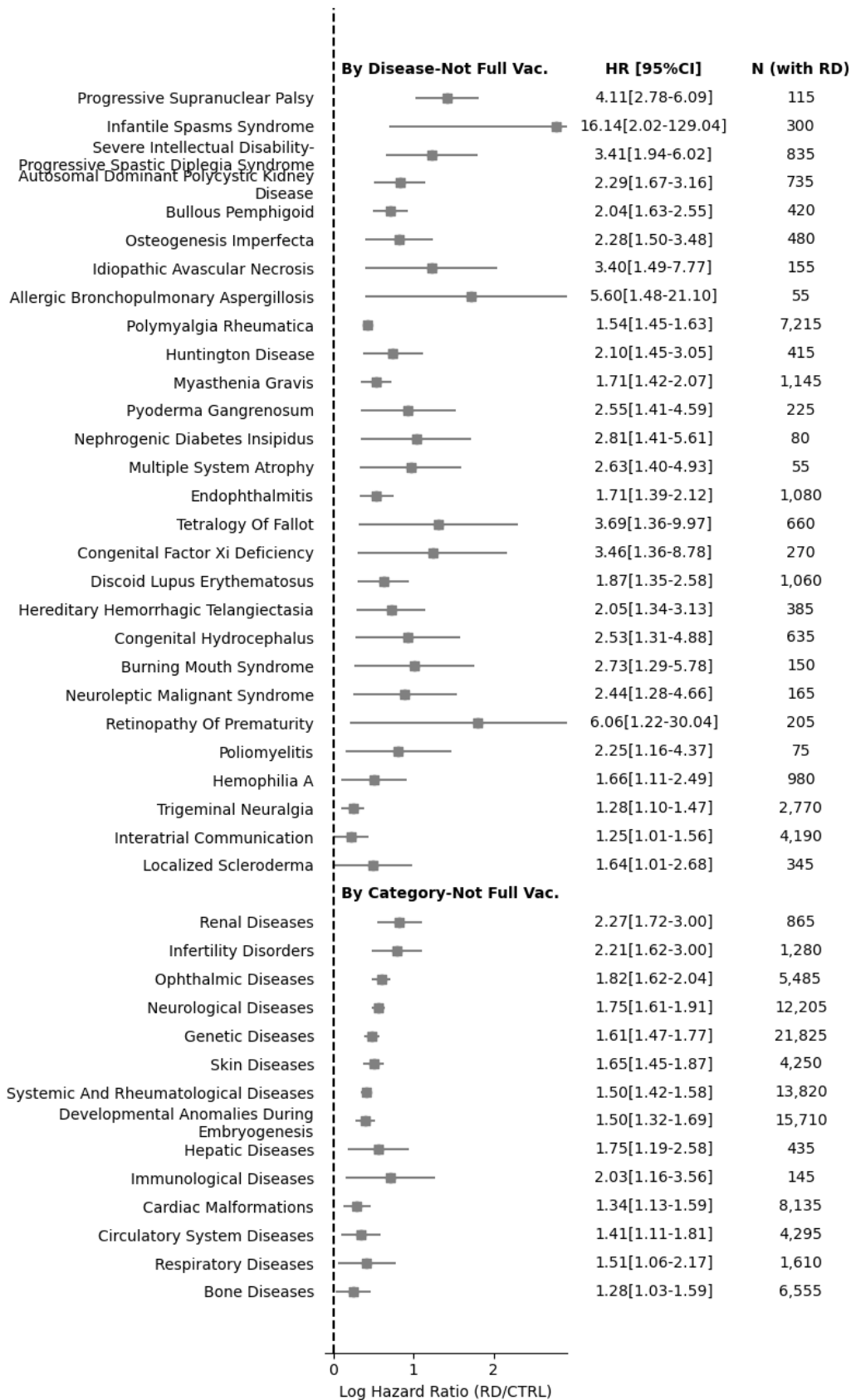
Supplementary figure 3 - Comparing adjusted and raw prevalence estimates:

Correlation between raw point prevalence per 100.000 on the x-axis, and age and gender adjusted point prevalences per 100.000 (using the English 2021 census data as reference population, see methods for detail). Right plot shows all 331 rare-diseases, left plot is zoomed in showing prevalence smaller than 25 cases per 100.000, for both estimates.



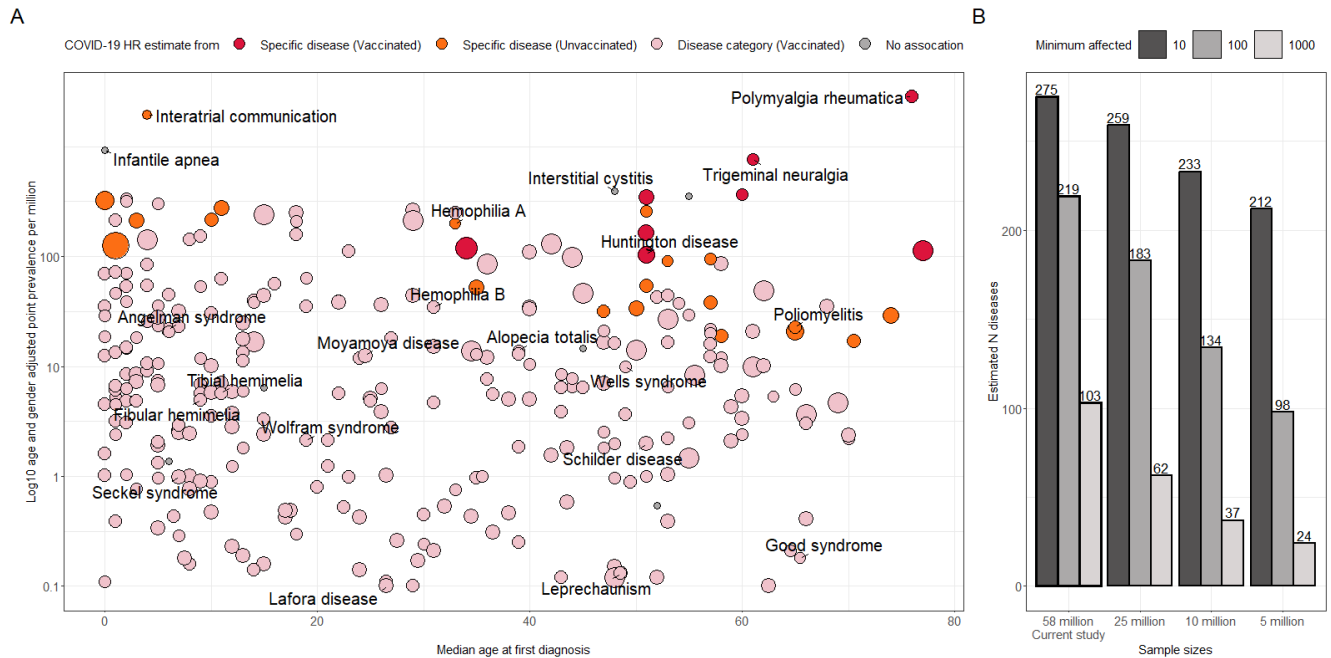
Supplementary figure 4 - Forest plots:

Forest plot showing Log Hazard Ratio of COVID-19-related death for rare diseases and categories with significant increase of risk in non-fully vaccinated individuals, based on lower-bound of the 95% confidence intervals (95% CI) of the hazard ratio (HR) between individuals with rare diseases (RD) and matched controls (CTRL). HRs and 95% CIs are displayed numerically on the right.



Supplementary figure 5 - COVID-19 mortality in relation to age at first diagnosis and prevalence

Hazard ratio for COVID-19 mortality for 290 Rare diseases in relation to median age at first diagnosis and prevalence. Point sizes are relative to the estimated Hazard Ratio (HR) for COVID-19 mortality. Colours represents the method by which the HR was estimated, and are prioritised in the following order based on what significant HR estimate is available for each disease; (Red, n = 8) estimates from fully vaccinated affected individuals of a specific rare disease, (Orange, n = 20) estimates from none fully vaccinated affected individuals of a specific rare diseases, (Light red, n = 296) estimates based on the highest HR from any of the rare disease category the specific disease belongs to, (dark grey, n = 7) no significant association found for the specific disease or disease category. Age and gender adjusted prevalence rate per million is shown on log10 scale. (B) Number of diseases with at least 10, 100 and 1000 identified rare disease patients, observed in our current study sample of ~58 million, and the estimated numbers of patients if we project our rare disease prevalence into sample sizes of 25, 10, 5 million populations. All current study prevalence estimates are available for all diseases as supplementary data 1.



Supplementary data

Supplementary data 1 - Rare disease demographics:

Demographics of all 331 rare diseases. *Adjusted population prevalence is age and gender adjusted prevalence per million individuals. Age is presented as the median age of first diagnosis (25% quantile and 75% quantile) and as numbers and percentages of cases within age groups calculated at study start. Age and sex was available for all diseases.

Supplementary data 2 - Gender and ethnicity differences:

Test statistics for Fishers exact analysis of gender and ethnicity differences of the 219 rare diseases with 100 or more patients identified. Ethnicity analysis compares Asian or Asian British and Black or Black British to the majority White ethnicity.

Supplementary data 3 - COVID-19 mortality for individual rare diseases:

Full results of COVID19 mortality analysis for individual rare diseases. Hazard ratios (HR) were calculated for people with rare diseases to matched controls from the general population. Numbers were rounded to the nearest five and numbers less than ten were suppressed due to data governance requirements.

Supplementary data 4 - COVID19 mortality analysis for rare disease categories:

Full results of COVID19 mortality analysis for rare disease categories. Hazard ratios (HR) were calculated for people with rare diseases to matched controls from the general population. Numbers were rounded to the nearest five and numbers less than ten were suppressed due to data governance requirements.

Supplementary data 5 - All rare diseases - Kaplan-Meier Plot:

Kaplan-Meier Plot for rare diseases or categories stratified by if individuals were fully vaccinated, based on lower-bound of the 95% confidence intervals (95% CI) of the hazard ratio (HR) between individuals with rare diseases (RD) and matched controls (CTRL). Events are defined as COVID-related death.

Please see the figures of all KM plots at the end of this document.

Supplementary tables

Supplementary table 1 - Rare disease information by category:

Percentage of diseases with unknown Orphanet frequency estimates by category. Note that multiple illnesses map to multiple categories, which is why column sums are greater than the total number of examined diseases, see supplementary table 2 for information on which diseases are in each category.

Rare disease category	Individuals			Female	Male	Rare diseases with unknown point prevalence in Orphanet (%)
	331 Rare diseases in category	with at least one disease in the category	Mean age of diagnosis			
Genetic diseases	202	377639	23.58 (24.18)	181736 (48.1)	195884 (51.9)	91 (45)
Developmental anomalies during Embryogenesis	152	314917	19.03 (23.2)	150299 (47.7)	164608 (52.3)	94 (61.8)
Systemic and rheumatological diseases	23	233241	61.7 (24.43)	144955 (62.1)	88285 (37.9)	16 (69.6)
Neurological diseases	121	191779	39.95 (26.12)	104356 (54.4)	87417 (45.6)	47 (38.8)
Cardiac malformations	48	175036	19.39 (24.04)	85419 (48.8)	89609 (51.2)	42 (87.5)
Ophthalmic diseases	47	111176	31.21 (29.92)	53427 (48.1)	57749 (51.9)	17 (36.2)
Bone diseases	38	87126	24.77 (22.64)	39366 (45.2)	47759 (54.8)	25 (65.8)
Circulatory system diseases	22	78097	17.34 (22.54)	36044 (46.2)	42051 (53.8)	20 (90.9)
Skin diseases	56	60795	50.66 (22.08)	39792 (65.5)	21002 (34.5)	22 (39.3)
Respiratory diseases	8	57943	7.1 (19.5)	27410 (47.3)	30531 (52.7)	7 (87.5)
Hematological diseases	9	36310	33.94 (21.95)	20222 (55.7)	16084 (44.3)	0
Urogenital diseases	2	27560	42 (22.46)	20054 (72.8)	7505 (27.2)	1 (50)
Otorhinolaryngological diseases	11	25748	7.37 (16.28)	10816 (42)	14932 (58)	6 (54.5)
Gastroenterological diseases	4	22057	52.67 (20.06)	11296 (51.2)	10760 (48.8)	1 (25)
Infertility disorders	9	17265	34.88 (24.18)	6270 (36.3)	10994 (63.7)	5 (55.6)
Endocrine diseases	25	14107	18.69 (19.84)	4333 (30.7)	9773 (69.3)	10 (40)
Systemic and rheumatological diseases of childhood	6	13764	21.67 (24.13)	5759 (41.8)	8004 (58.2)	3 (50)
Odontological diseases	7	13737	22.03 (15.62)	7885 (57.4)	5850 (42.6)	2 (28.6)
Renal diseases	3	11569	50.59 (16.98)	5836 (50.4)	5733 (49.6)	0
Inborn errors of metabolism	18	7150	21.76 (27.15)	3653 (51.1)	3497 (48.9)	5 (27.8)
Hepatic diseases	2	6368	53.66 (20.28)	3369 (52.9)	2999 (47.1)	0
Cardiac diseases	4	2681	24.38 (21.47)	1528 (57)	1153 (43)	2 (50)
Immunological diseases	14	2288	43.75 (26.64)	1170 (51.1)	1118 (48.9)	11 (78.6)
Gynecological and obstetric diseases	4	2057	16.19 (17.83)	1015 (49.3)	1042 (50.7)	2 (50)
Allergic disease	3	1599	60.09 (18.74)	797 (49.8)	802 (50.2)	3 (100)

Supplementary table 2 - Overview of Orphanet data sources for prevalence

Orphanet prevalence (per million)	<1	1-9	10-99	100-500	All estimates	
Number of rare disease studies (N=145)						
Location						
UK	3	5	10	5		
EU	6	18	28	17		
Worldwide	43	5	2	4		
Source type*						
Review	7	3	7	3		
Article (observational study, survey)	6	9	5	5		
Case series report	16	2	3	2		
Guideline	0	1	0	1		
European Medicine Agency Report	0	2	3	7		
Expert	24	10	22	7		
RCT	0	0	0	1		
Unclear**	0	3	0	0		

*The categories below are not mutually exclusive, as some rare diseases have multiple sources listed.

**Little information from these 3 as there is no reference data. Otherwise we can merge them with the 'article' category.

Supplementary table 3 - STROBE statement:

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation	Page No
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract <u>Retrospective cohort study</u>	1
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	3-4
Objectives	3	State specific objectives, including any prespecified hypotheses	6
Methods			
Study design	4	Present key elements of study design early in the paper	5-6, S3-4
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5-6, S2-4
Participants		(a) <i>Cohort study</i> —Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up <i>Case-control study</i> —Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls <i>Cross-sectional study</i> —Give the eligibility criteria, and the sources and methods of selection of participants	5-6, S2-4
	6	(b) <i>Cohort study</i> —For matched studies, give matching criteria and number of exposed and unexposed <i>Case-control study</i> —For matched studies, give matching criteria and the number of controls per case Exact match with ratio Matching criteria: Age group Sex Ethnicity Deprivation Smoking Comorbidities (Might not be able to match) Medications (Might not be able to match)	5-6, S2-4

Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	S3
Data sources/measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	5, S2-3
Bias	9	Describe any efforts to address potential sources of bias	
Study size	10	Explain how the study size was arrived at	5, S2
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	S3
Statistical methods		(a) Describe all statistical methods, including those used to control for confounding	6, S3
		(b) Describe any methods used to examine subgroups and interactions	
		(c) Explain how missing data were addressed	
	12	(d) <i>Cohort study</i> —If applicable, explain how loss to follow-up was addressed <i>Case-control study</i> —If applicable, explain how matching of cases and controls was addressed <i>Cross-sectional study</i> —If applicable, describe analytical methods taking account of sampling strategy	S3
		(e) Describe any sensitivity analyses	
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	8

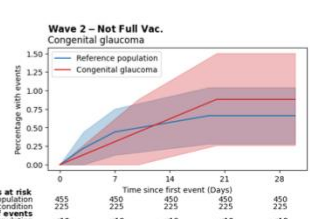
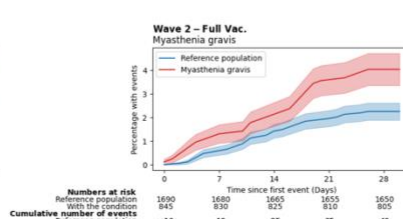
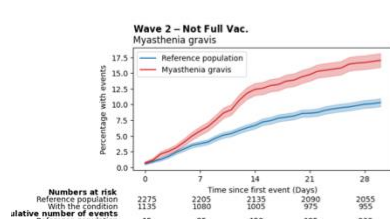
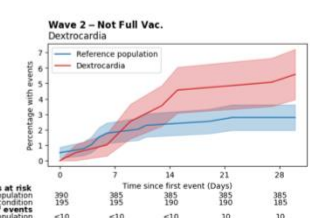
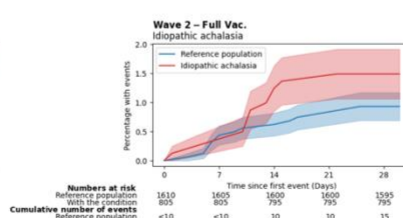
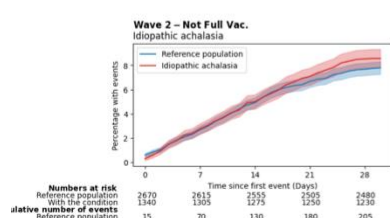
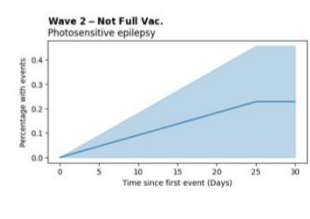
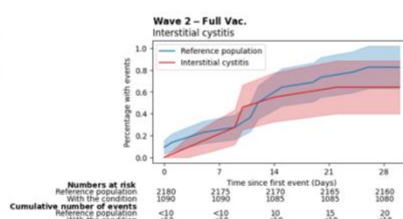
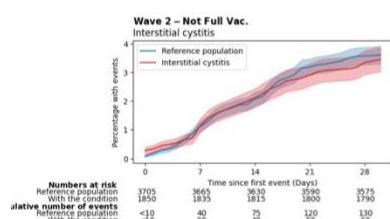
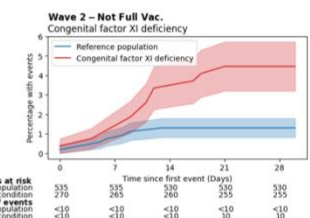
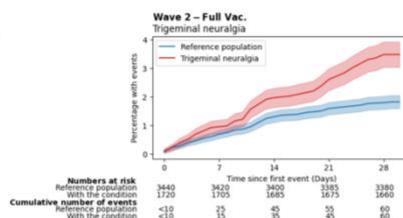
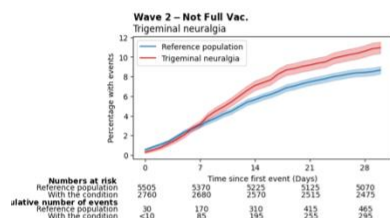
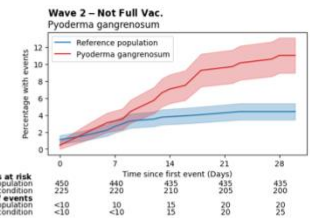
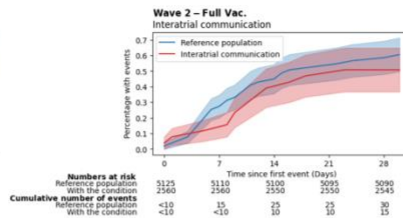
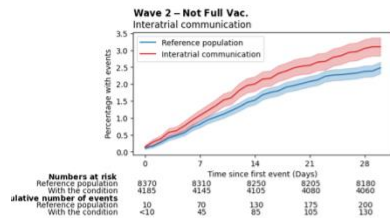
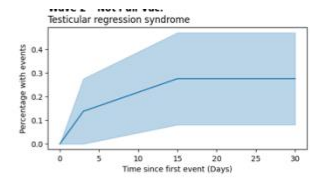
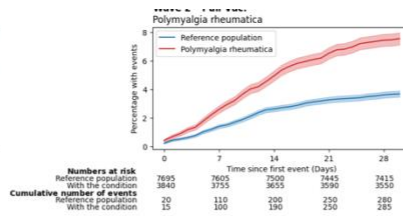
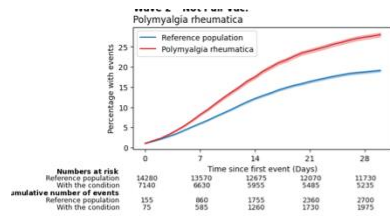
		(b) Indicate number of participants with missing data for each variable of interest	
		(c) <i>Cohort study</i> —Summarise follow-up time (eg, average and total amount)	
Outcome data	15*	<i>Cohort study</i> —Report numbers of outcome events or summary measures over time	8-14
		<i>Case-control study</i> —Report numbers in each exposure category, or summary measures of exposure	
		<i>Cross-sectional study</i> —Report numbers of outcome events or summary measures	
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	12
		(b) Report category boundaries when continuous variables were categorized	
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	
Discussion			
Key results	18	Summarise key results with reference to study objectives	15
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	16

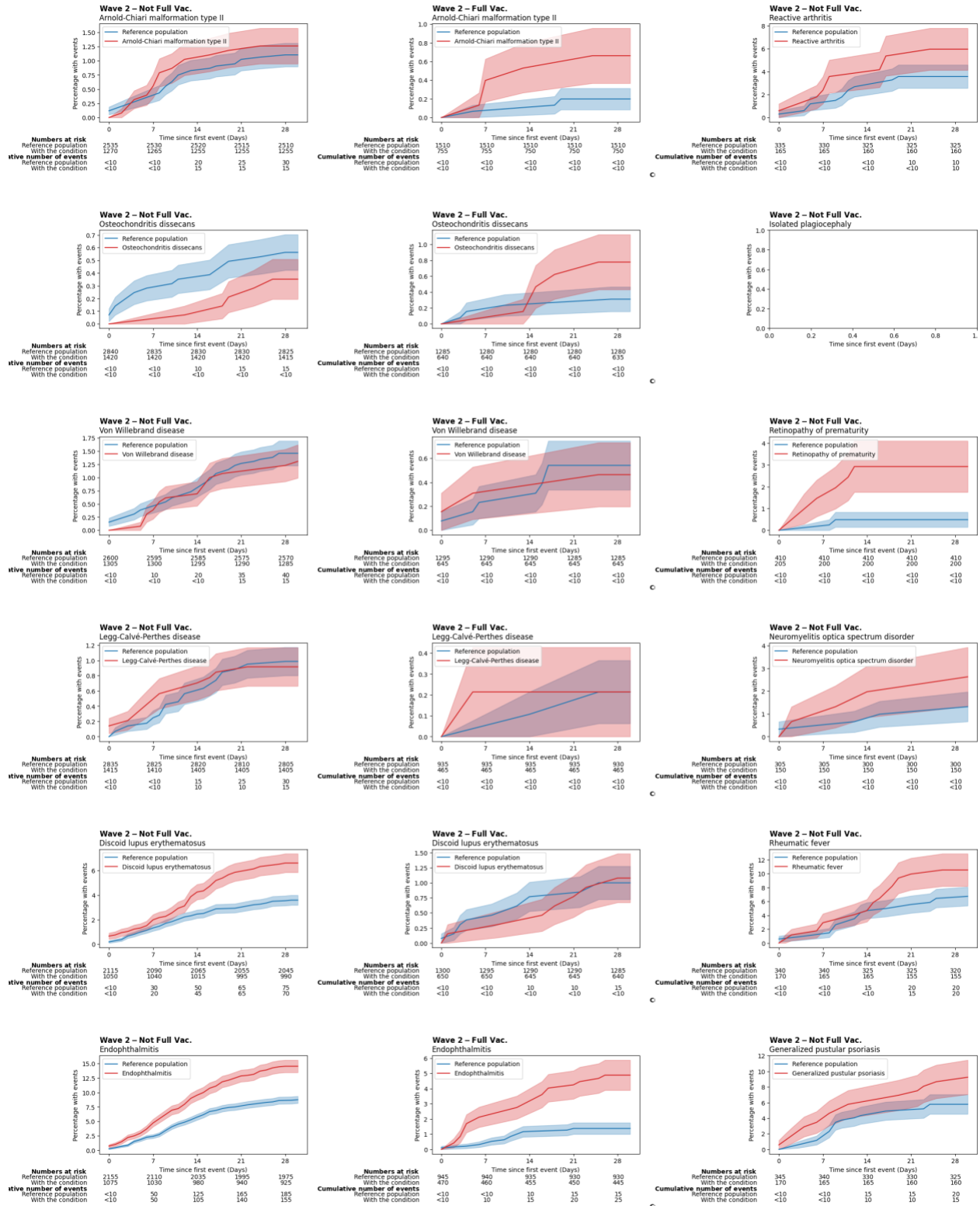
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	16
Generalisability	21	Discuss the generalisability (external validity) of the study results	16
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	16-17

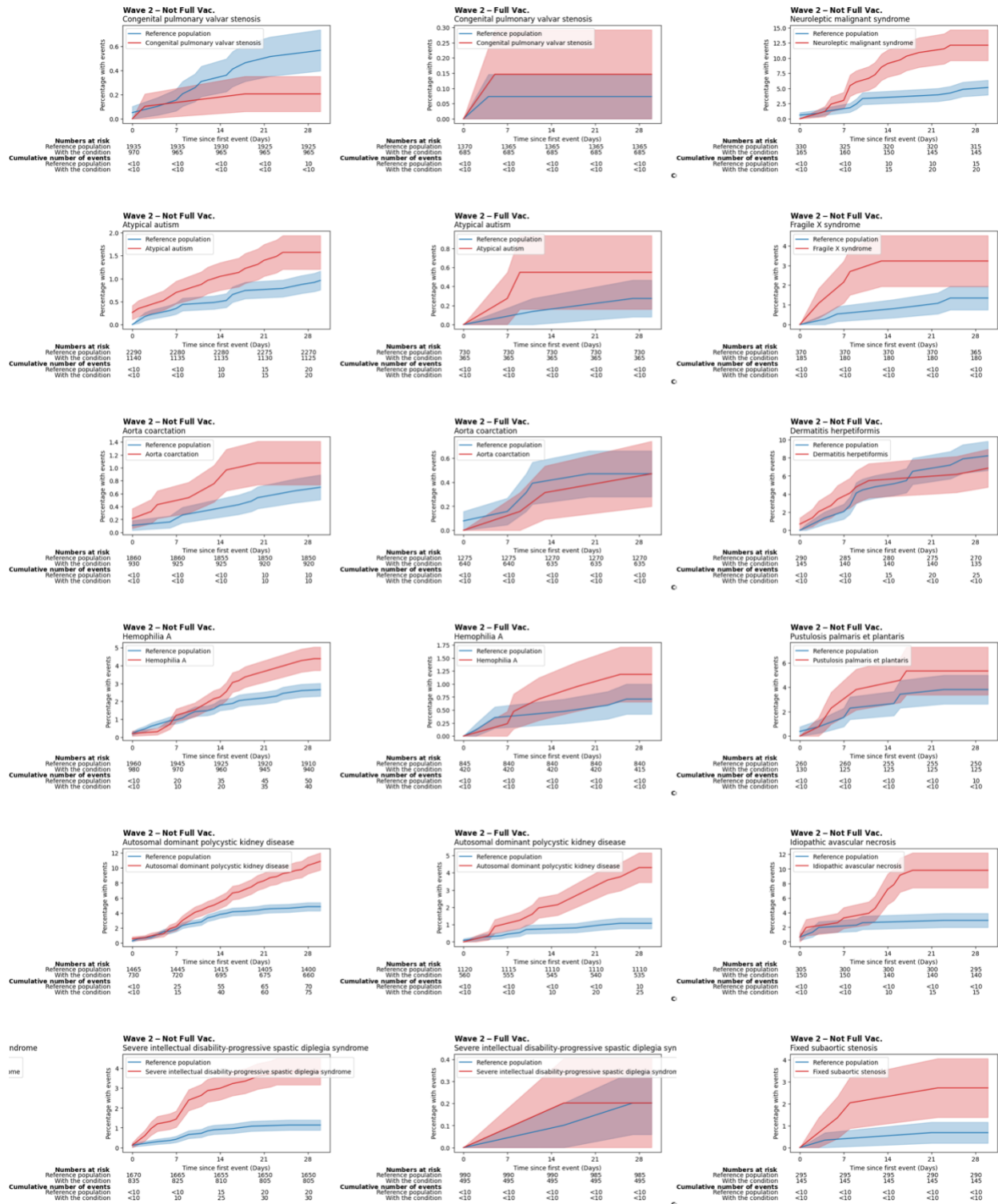
*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

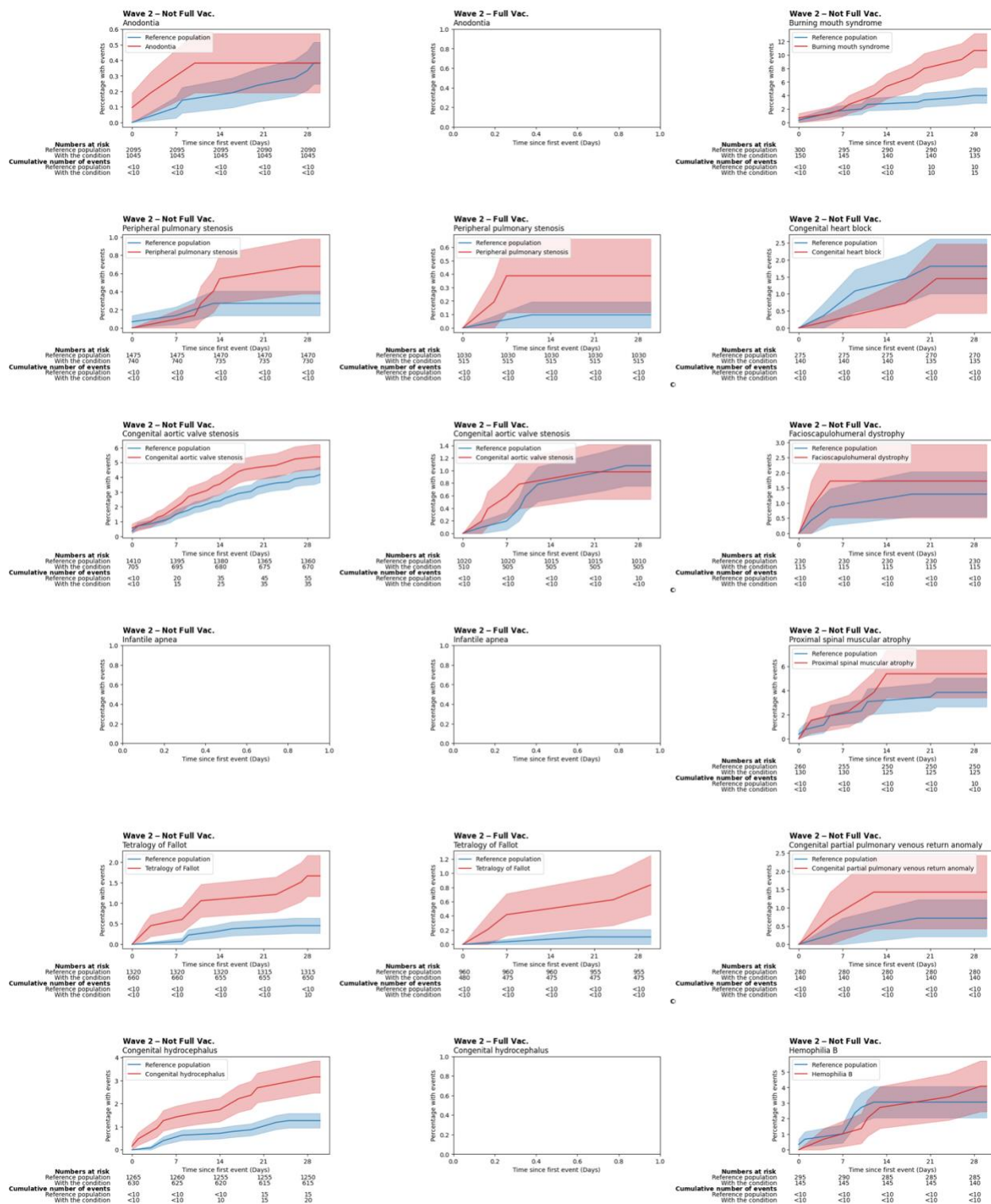
Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is available at www.strobe-statement.org.

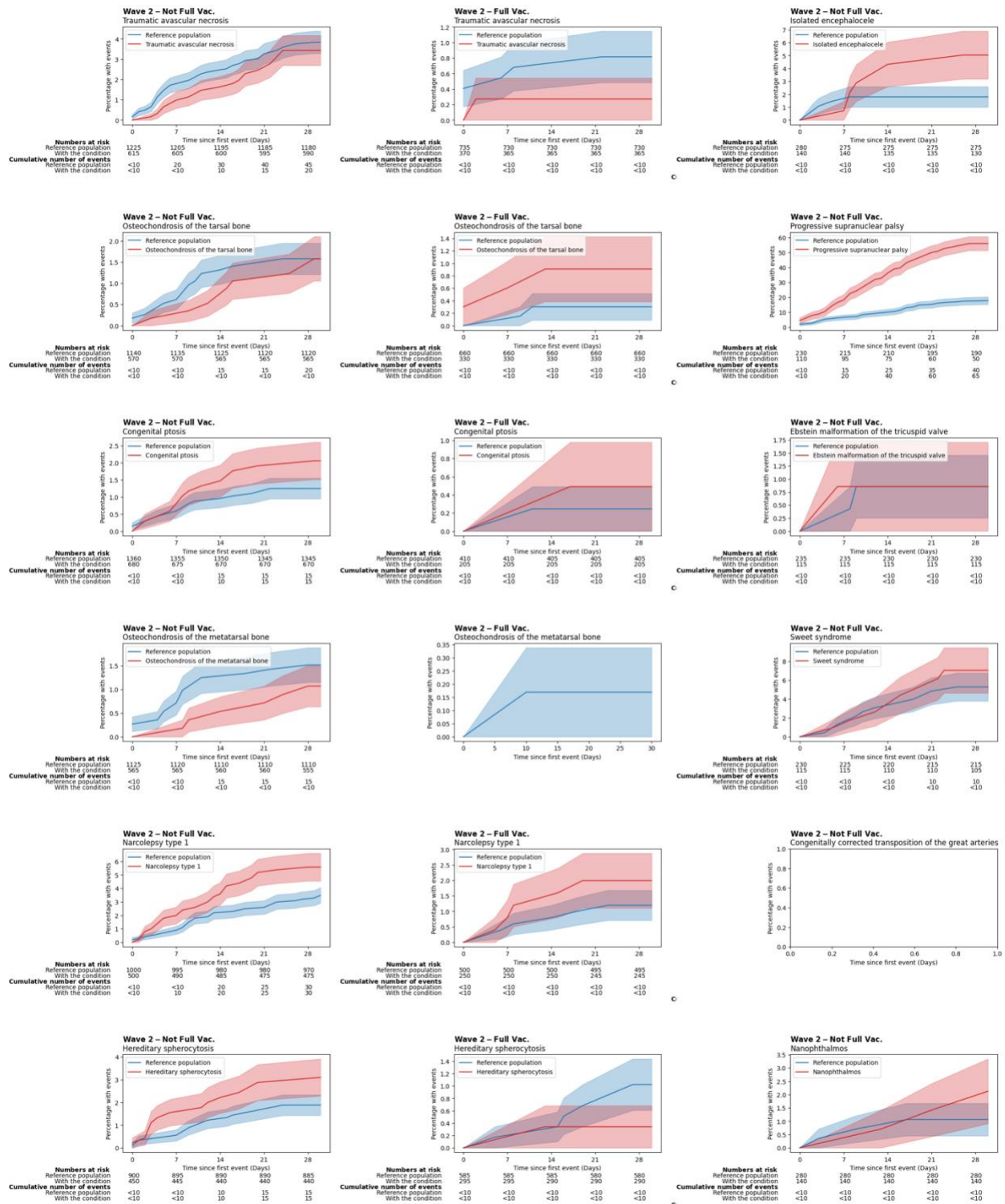
Supplementary data 5 figures

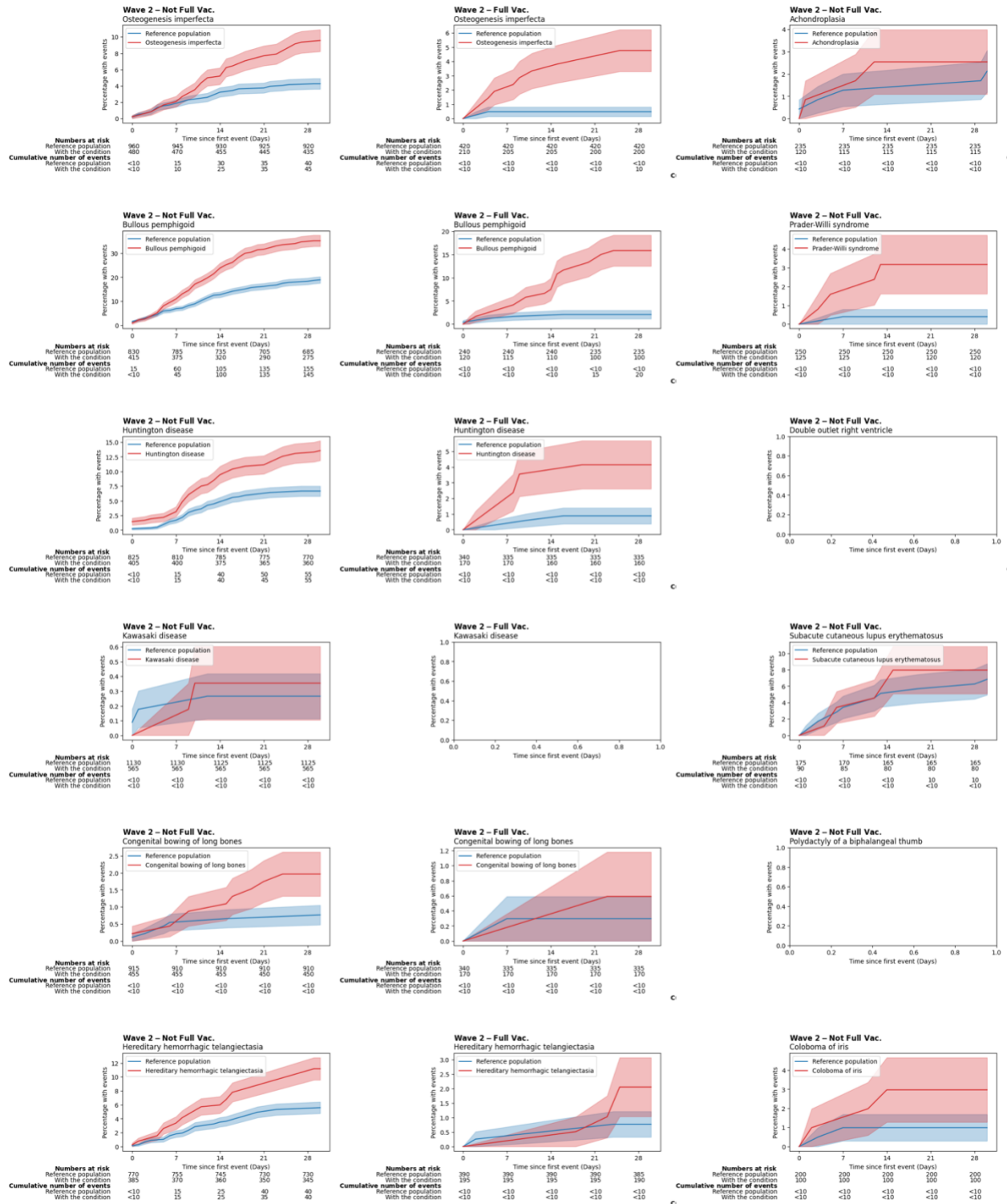


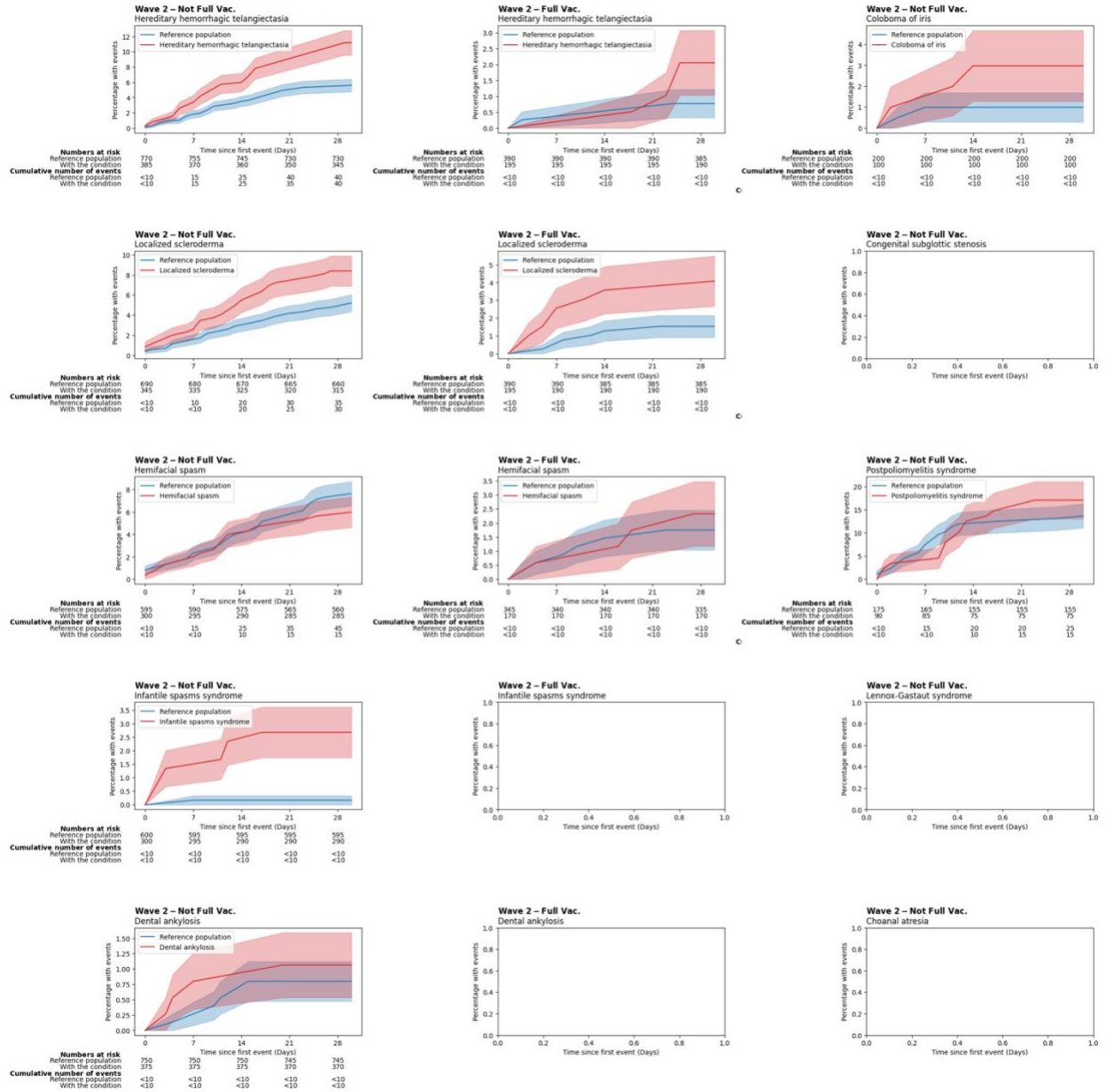


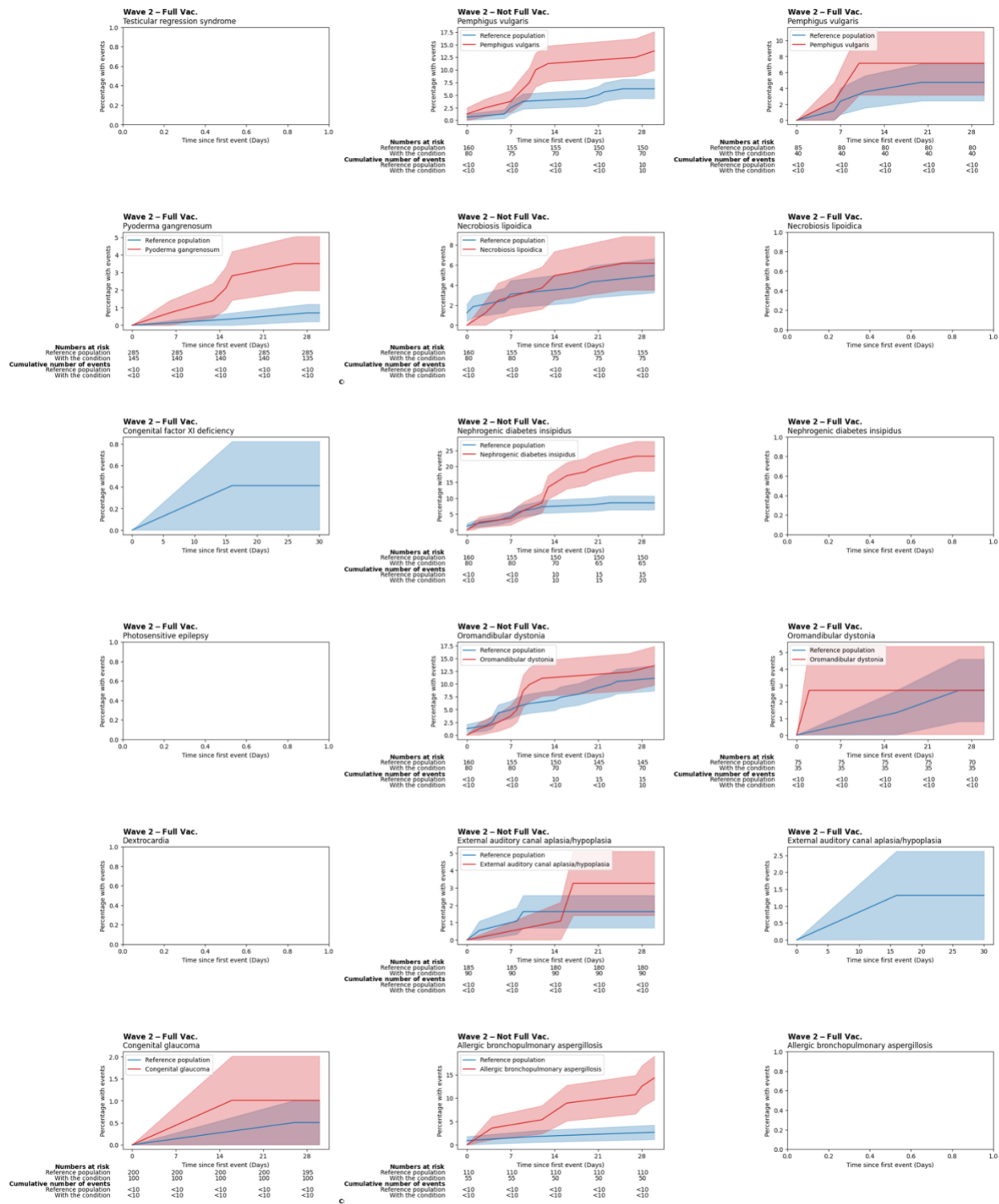


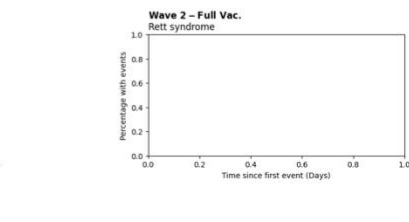
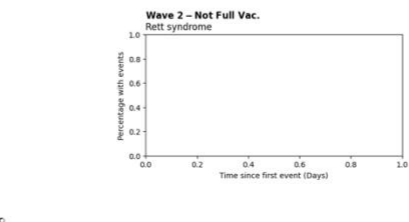
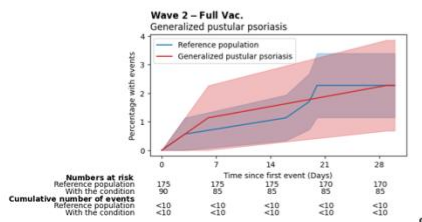
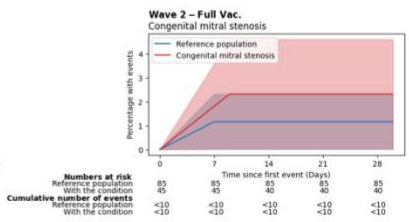
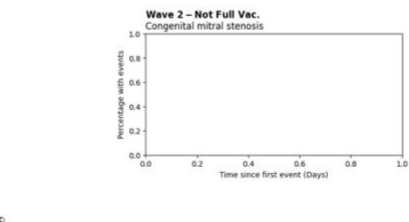
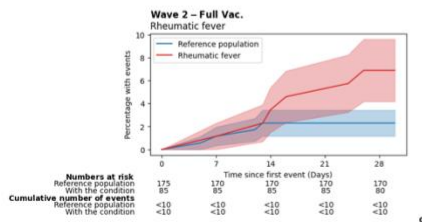
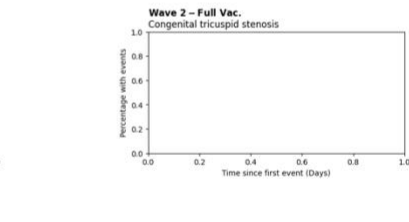
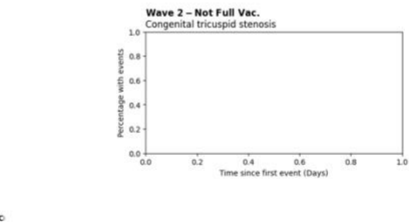
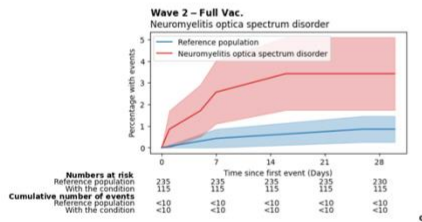
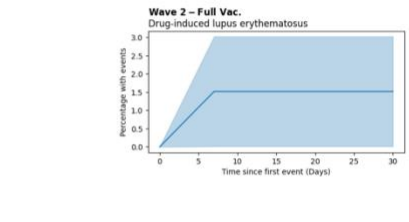
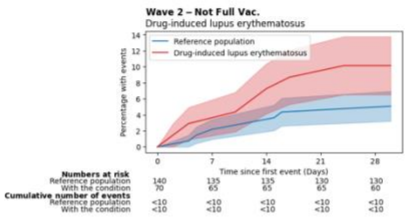
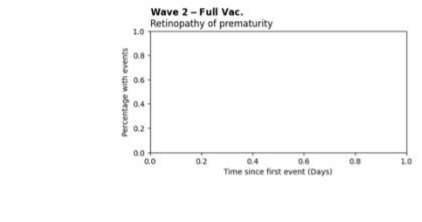
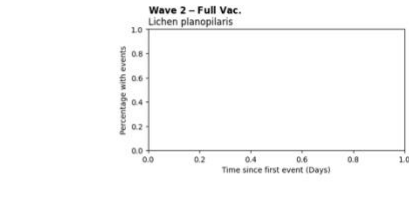
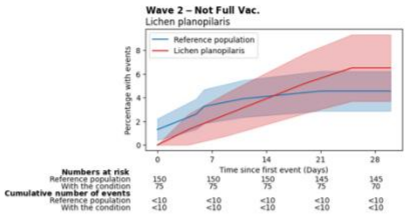
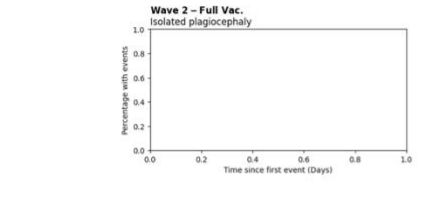
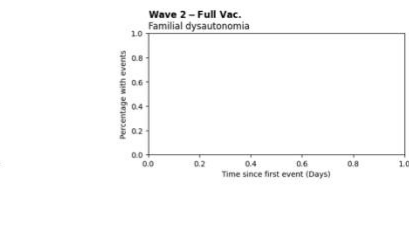
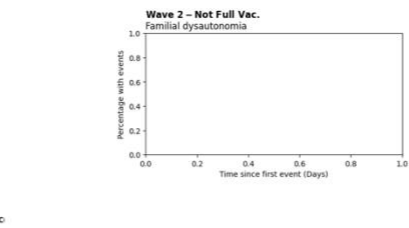
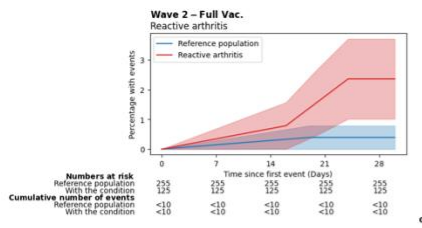


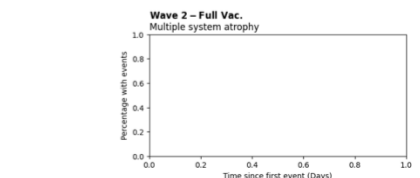
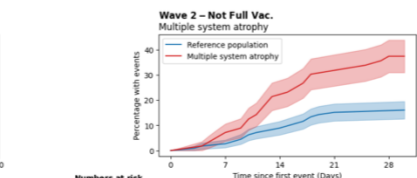
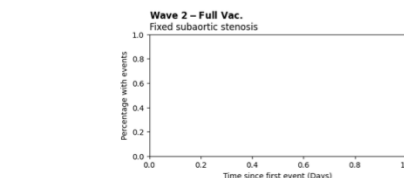
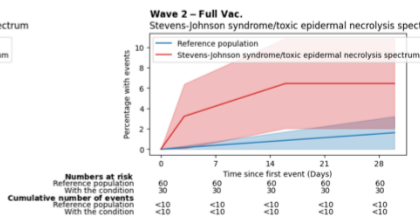
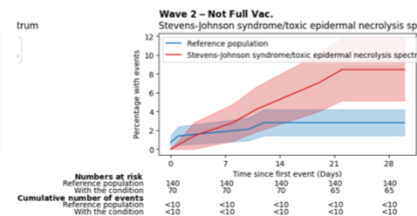
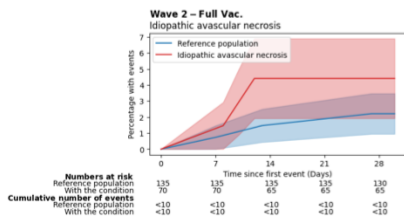
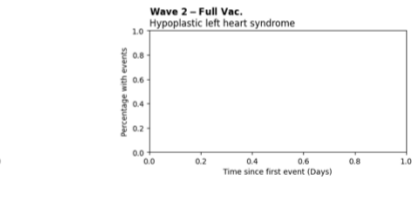
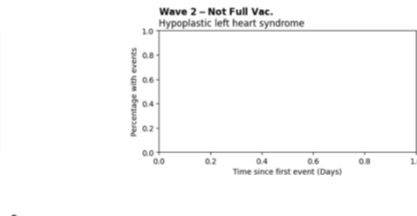
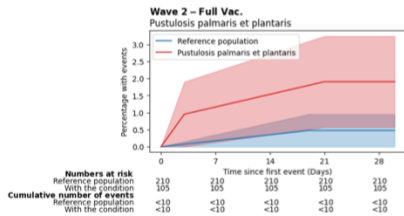
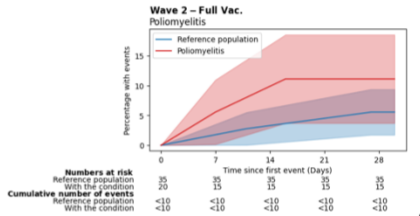
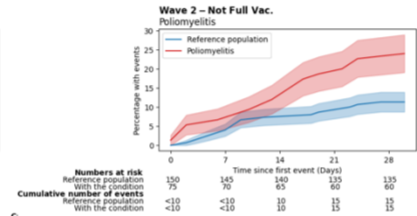
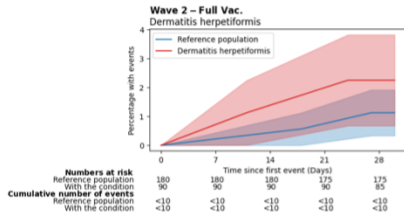
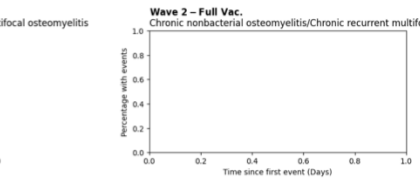
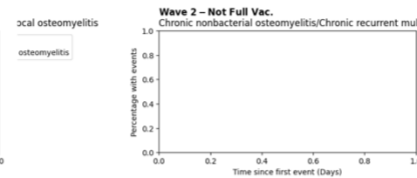
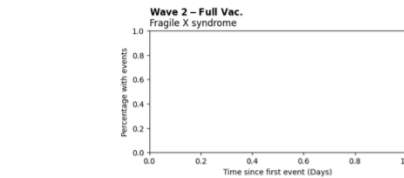
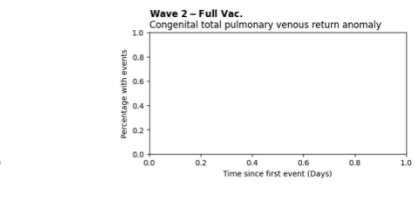
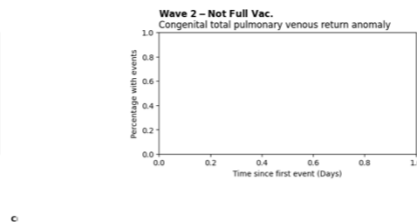
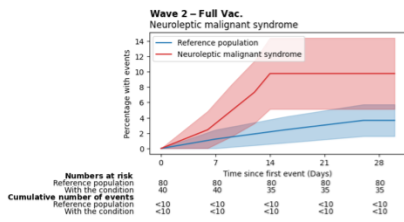


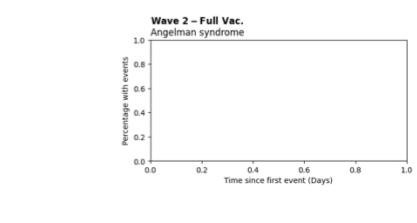
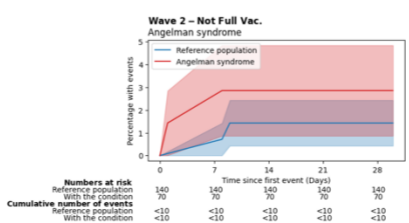
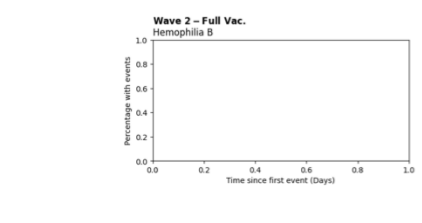
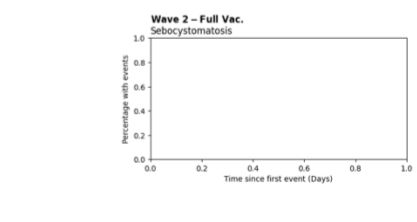
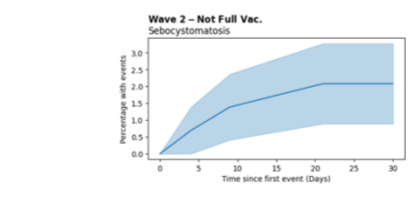
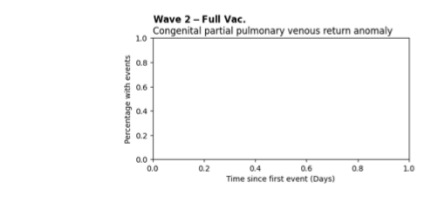
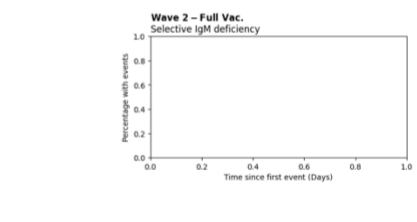
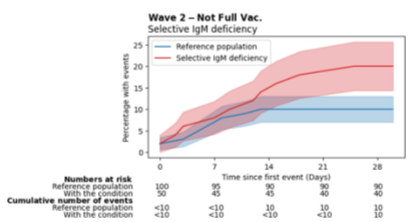
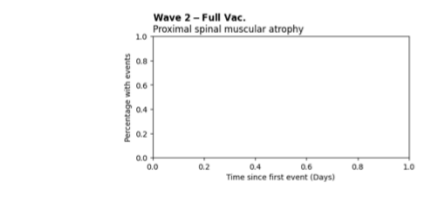
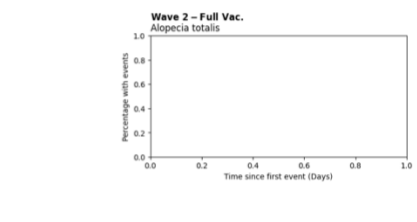
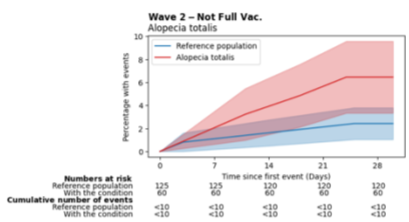
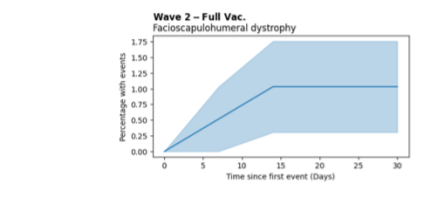
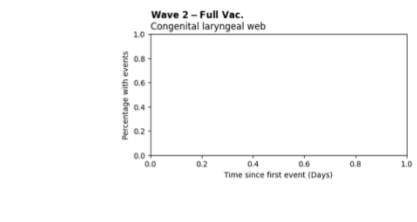
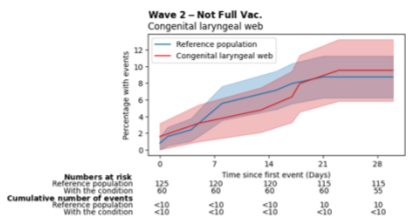
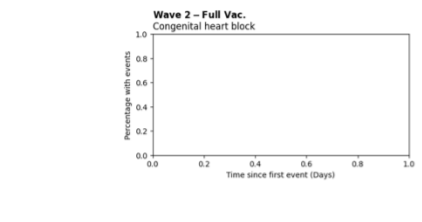
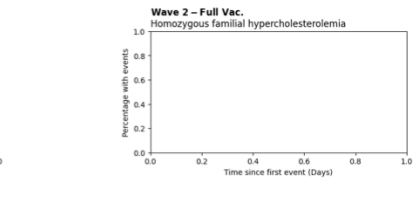
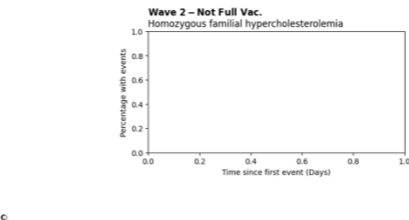
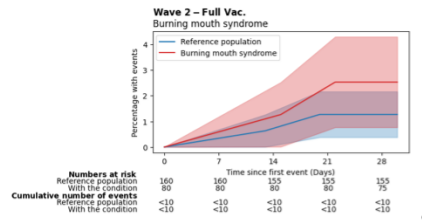


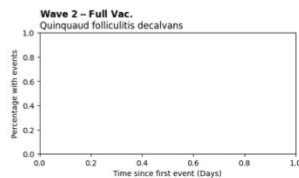
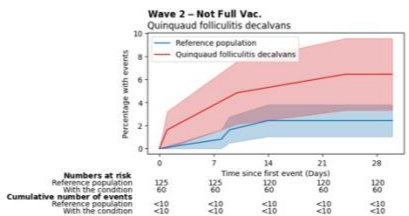
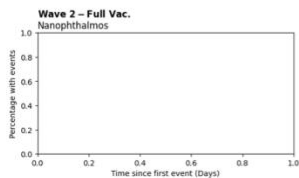
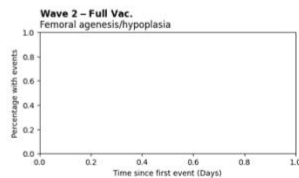
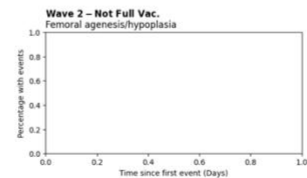
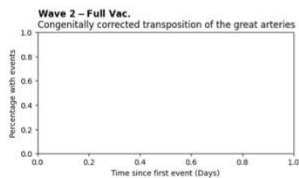
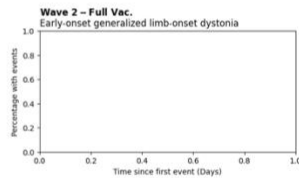
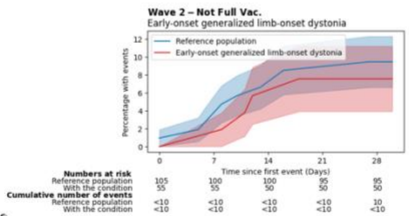
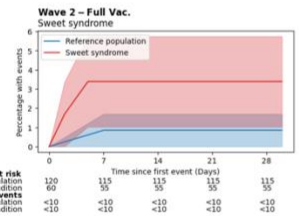
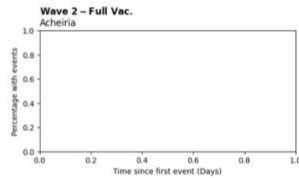
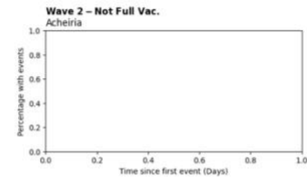
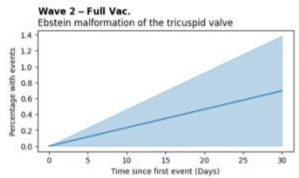
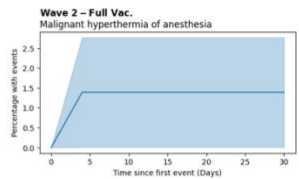
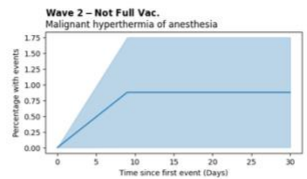
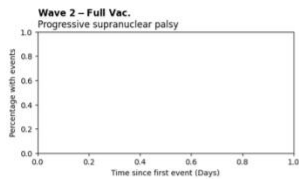
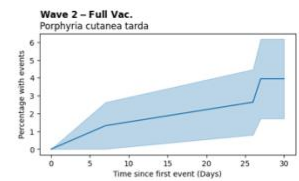
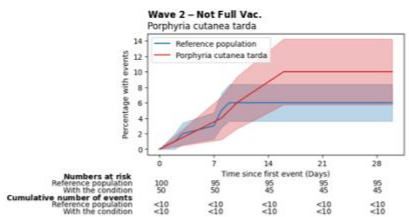
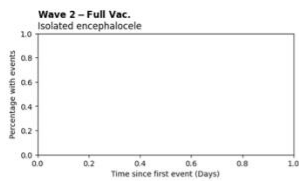


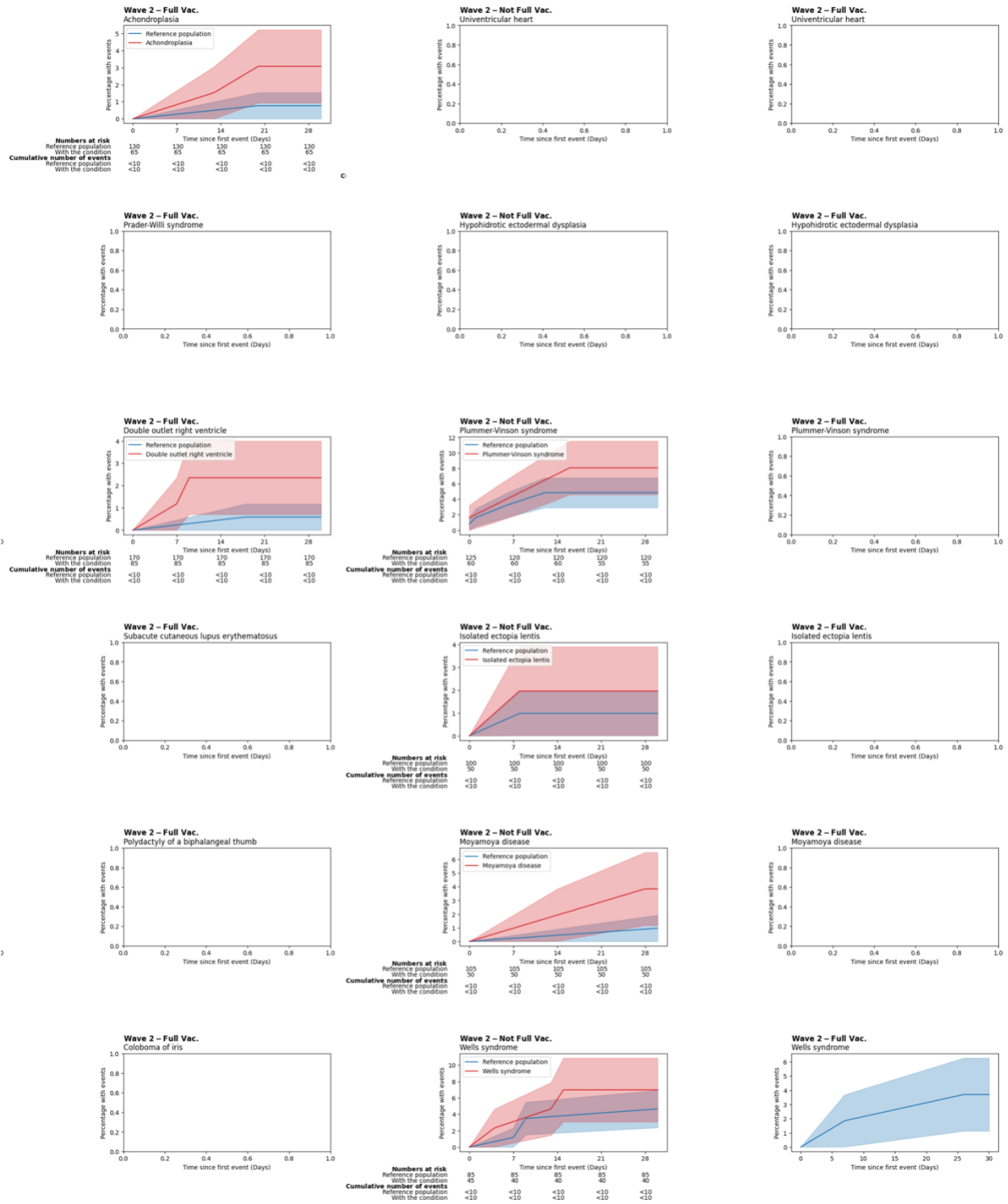


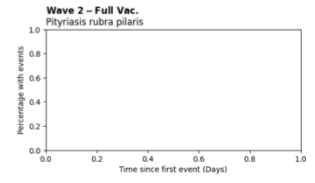
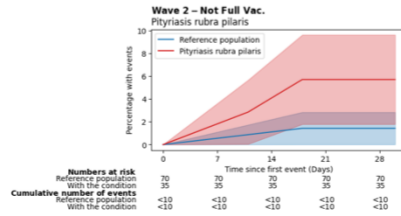
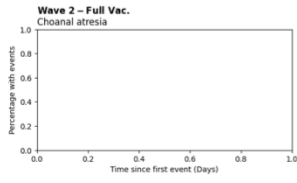
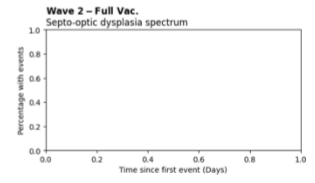
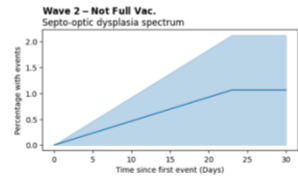
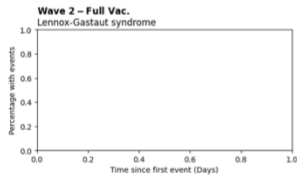
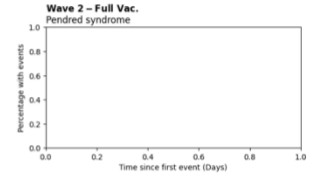
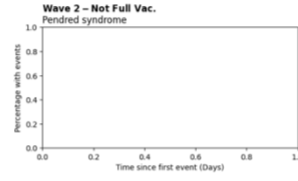
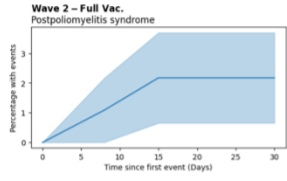
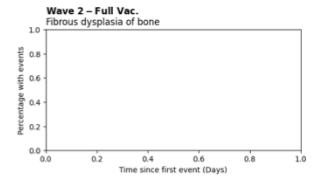
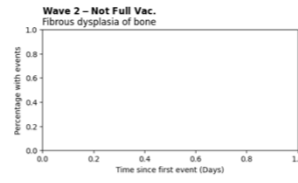
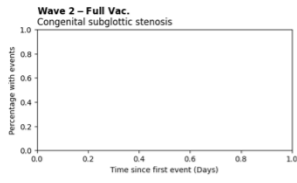


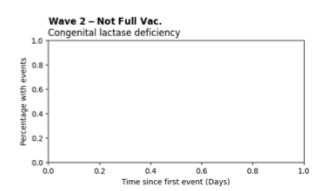
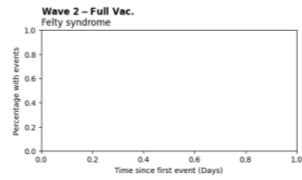
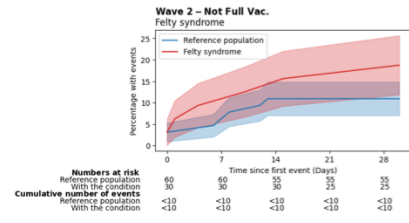
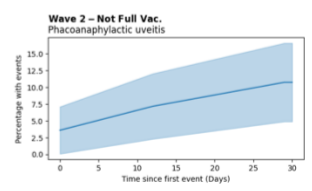
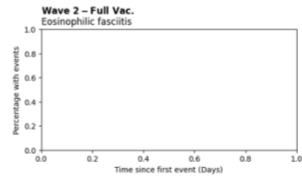
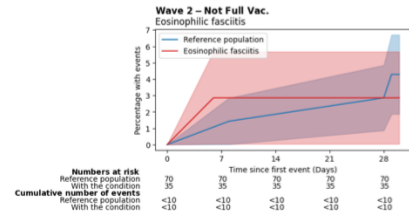
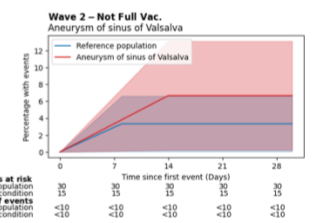
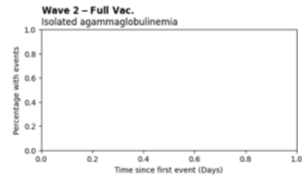
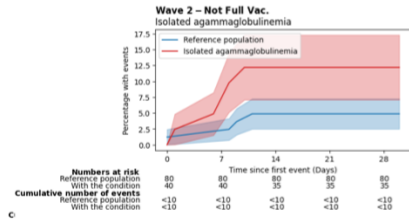
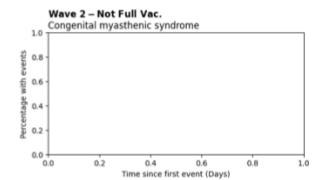
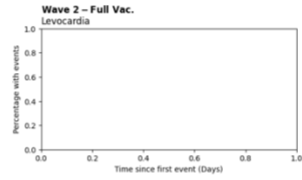
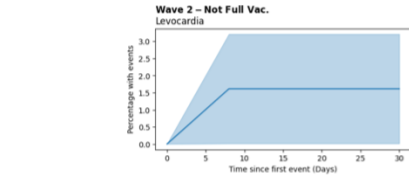
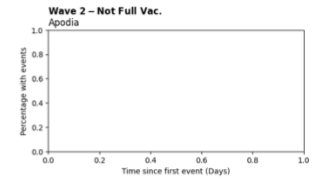
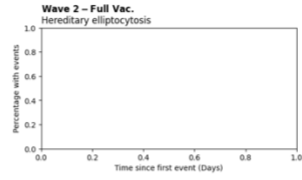
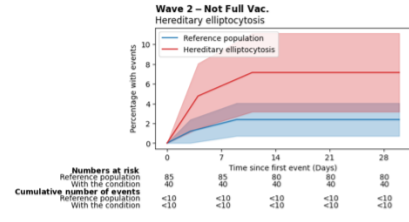
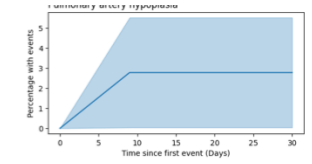
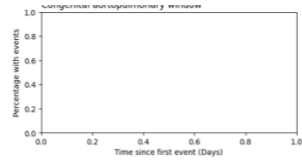
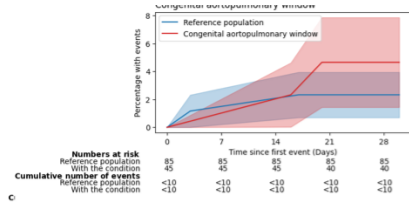


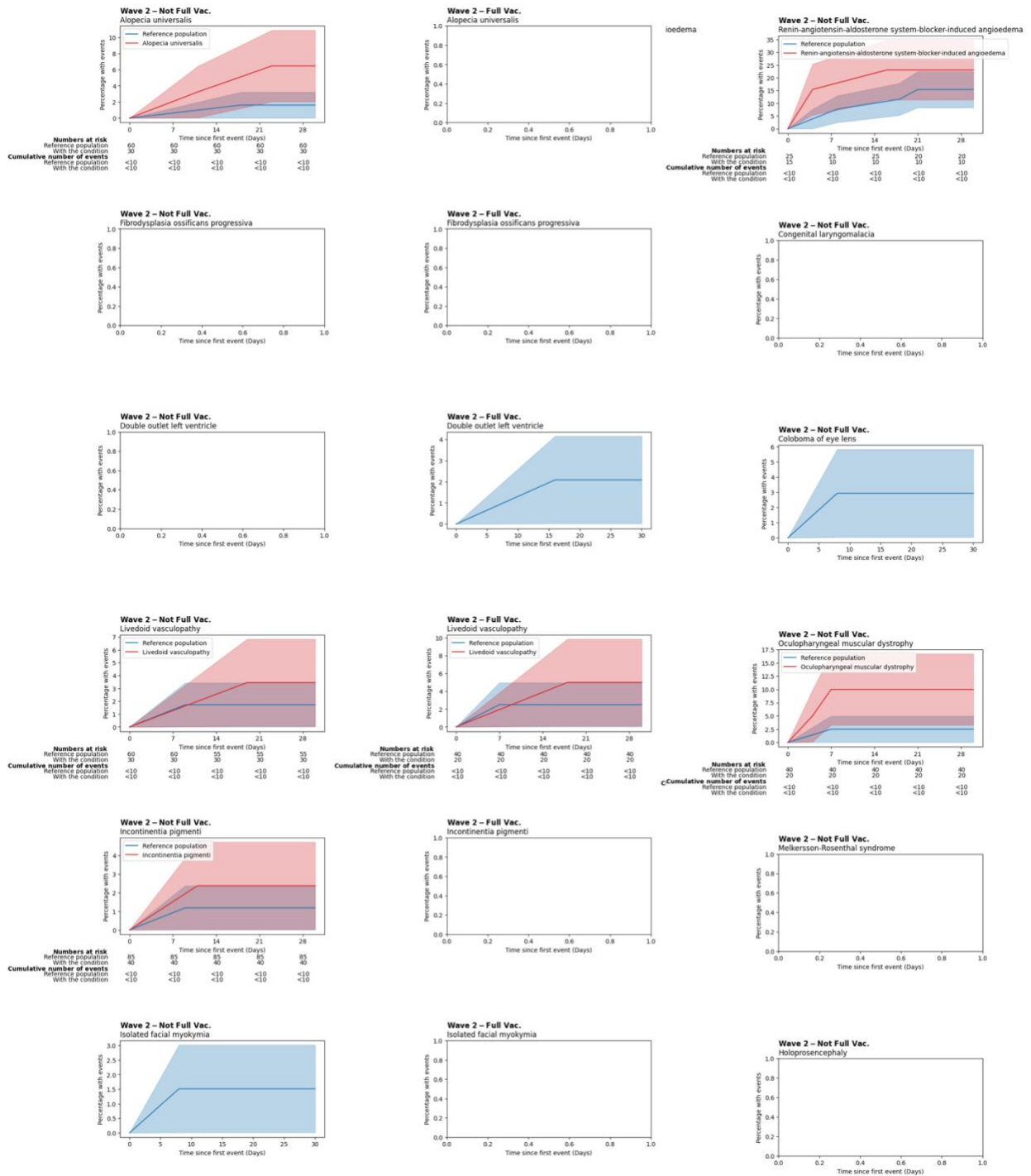


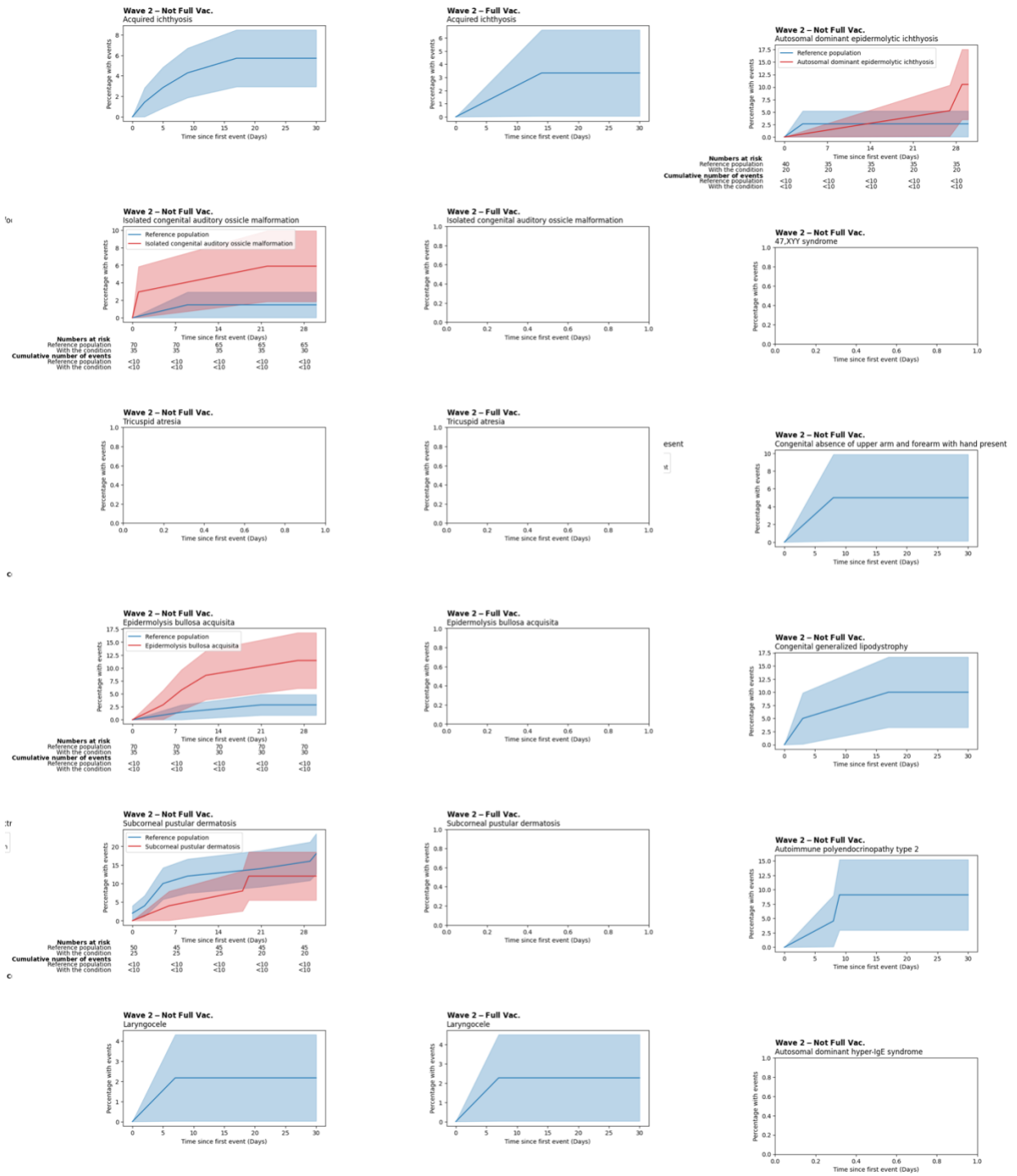


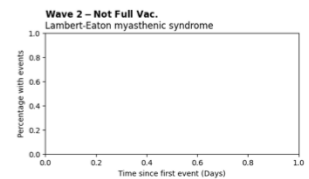
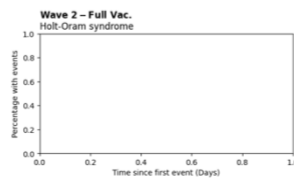
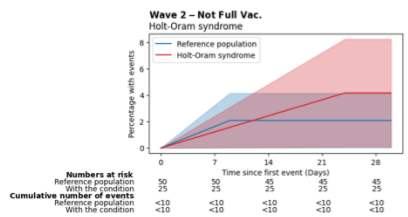
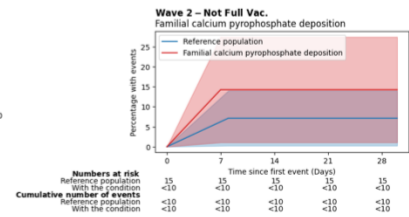
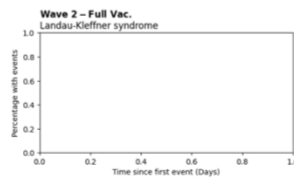
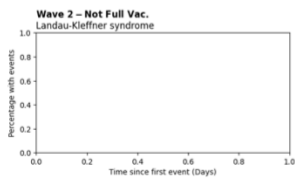
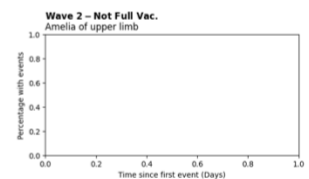
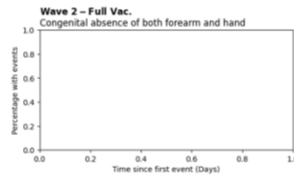
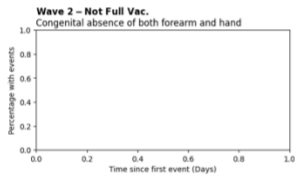
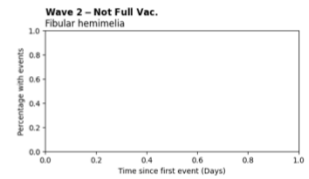
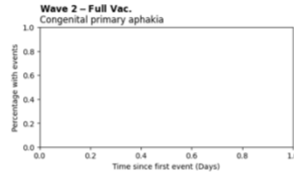
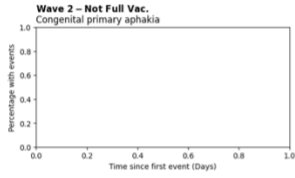
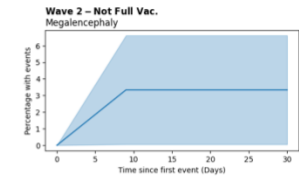
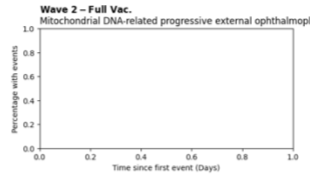
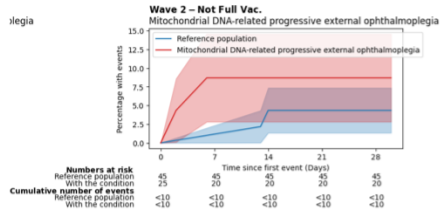
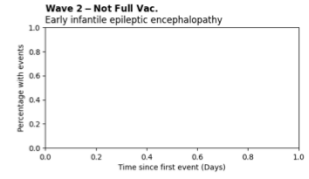
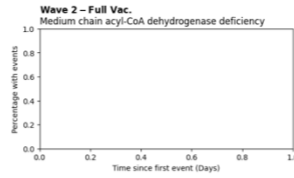
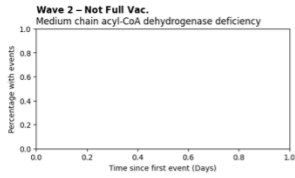


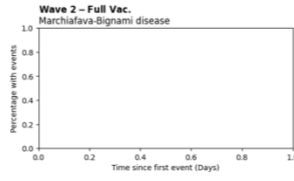
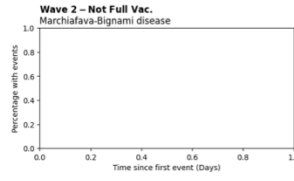
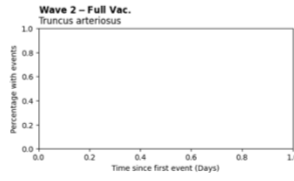
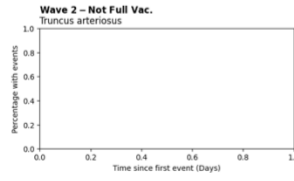
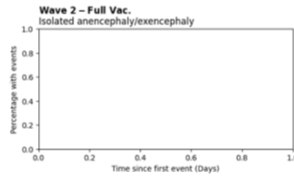
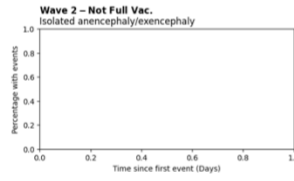
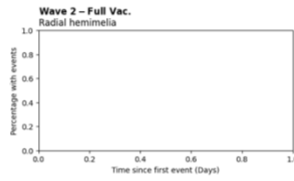
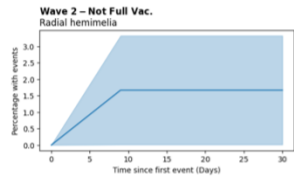
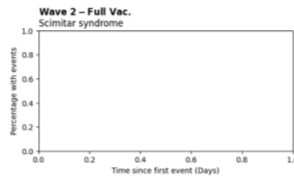
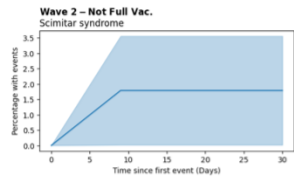
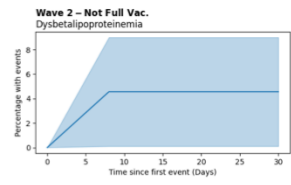
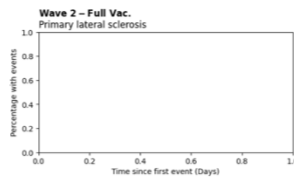
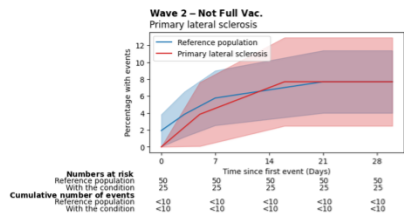


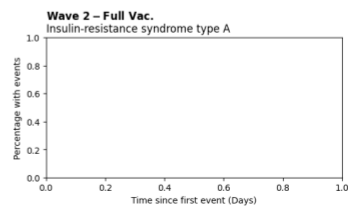
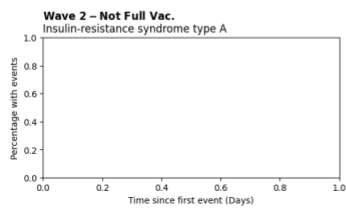
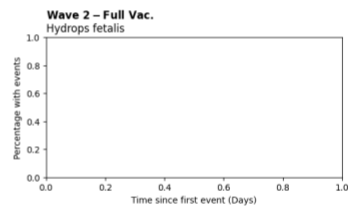
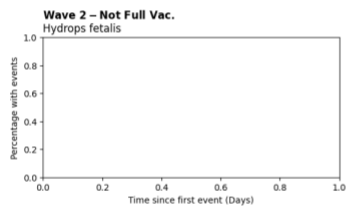
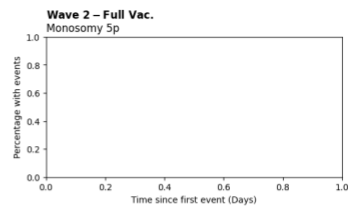
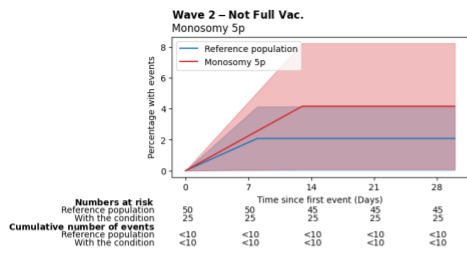
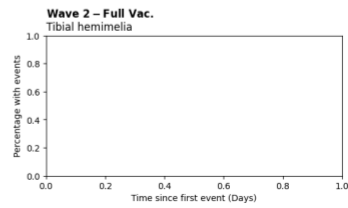
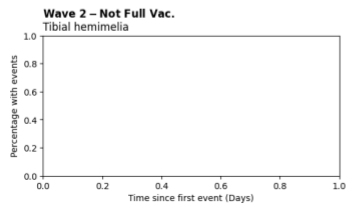
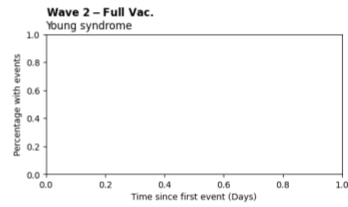
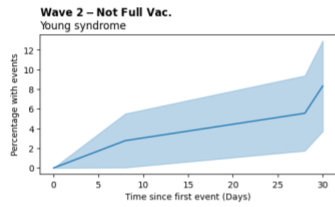
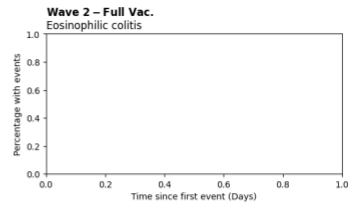
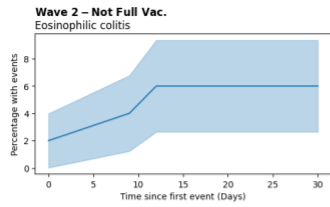


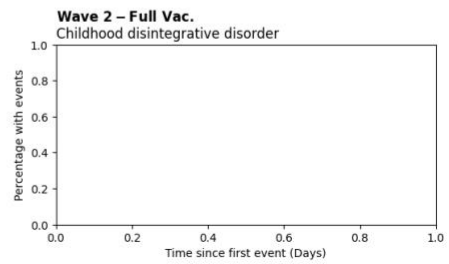
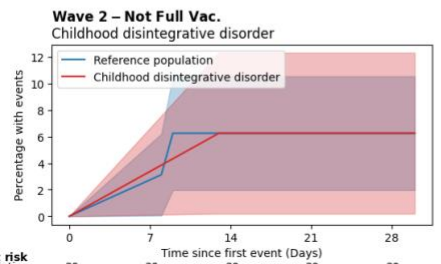
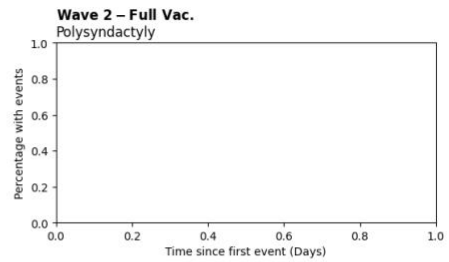
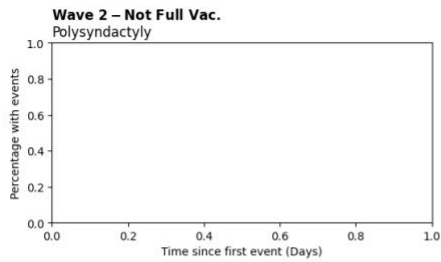
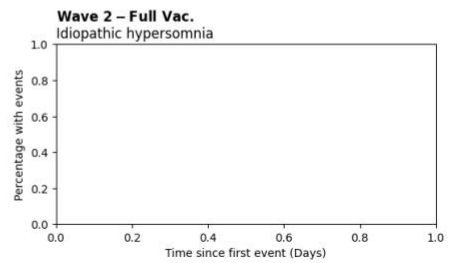
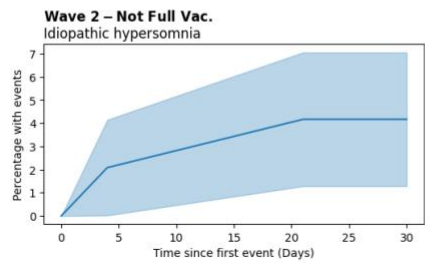
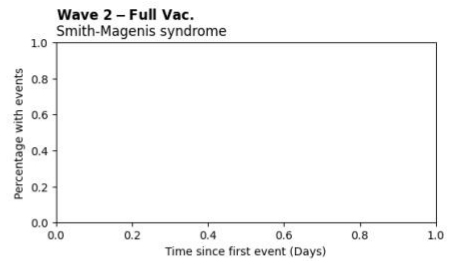
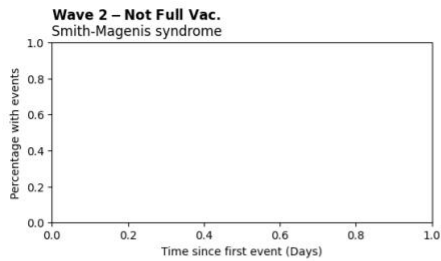










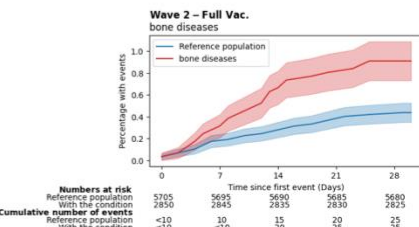
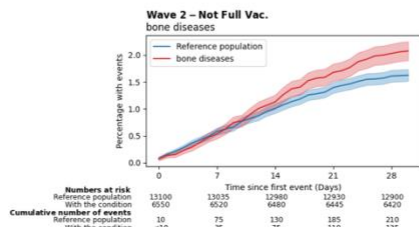
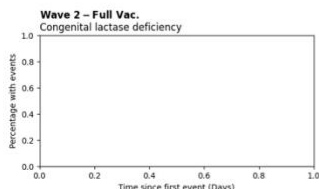
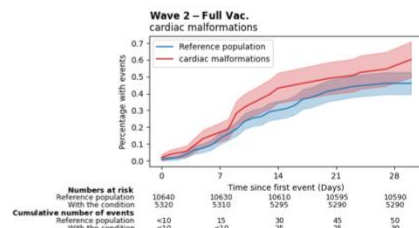
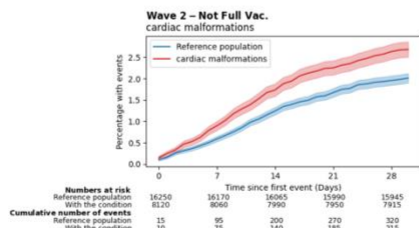
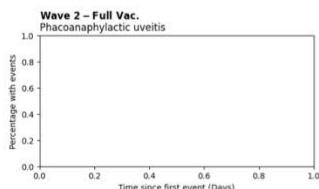
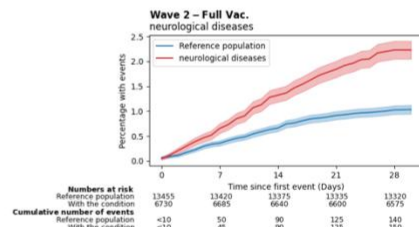
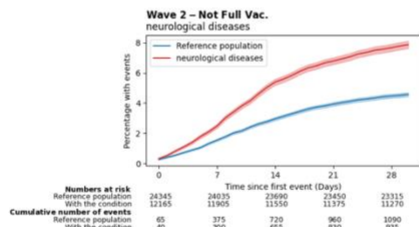
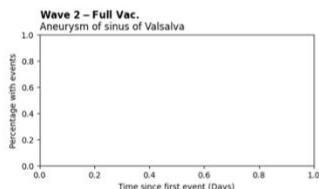
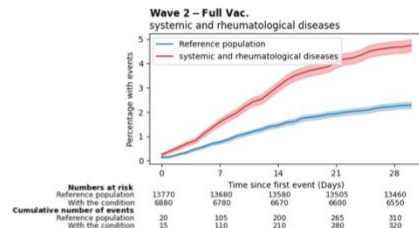
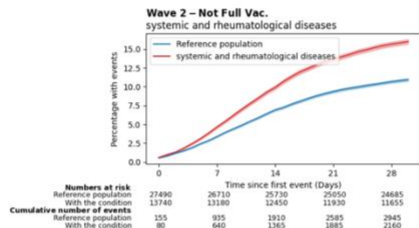
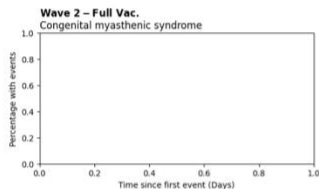
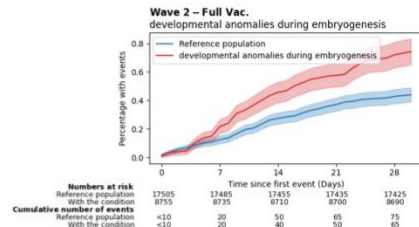
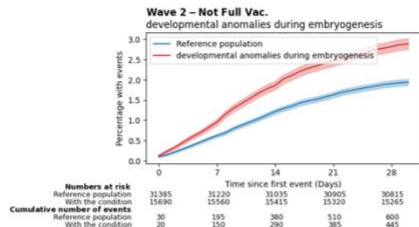
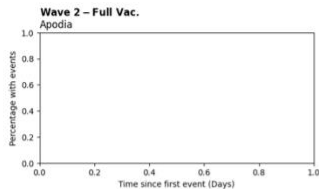
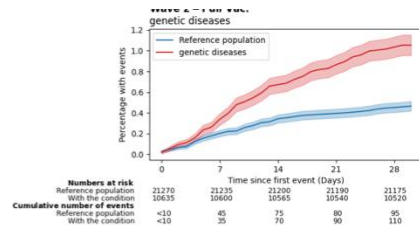
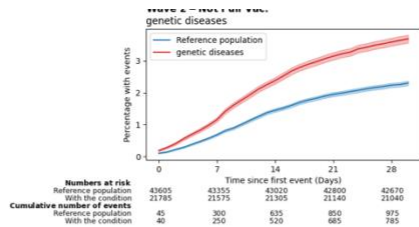
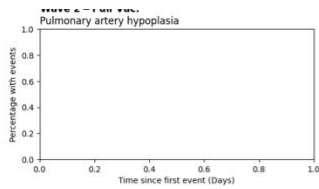


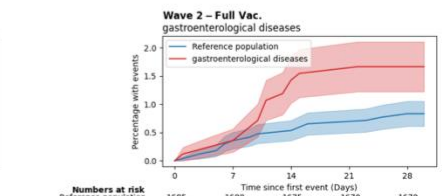
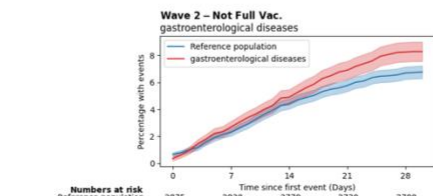
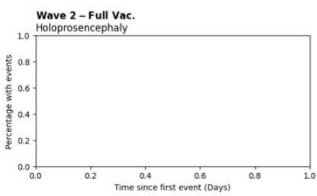
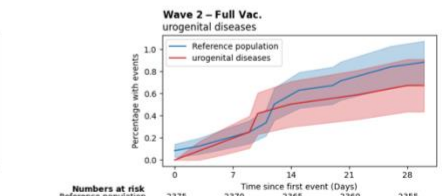
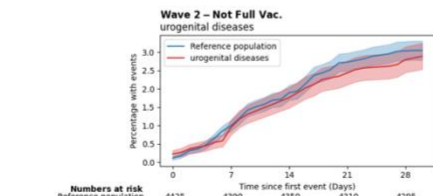
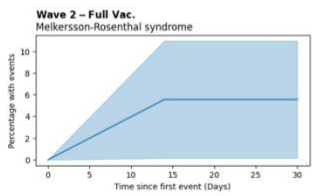
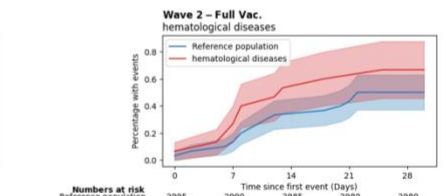
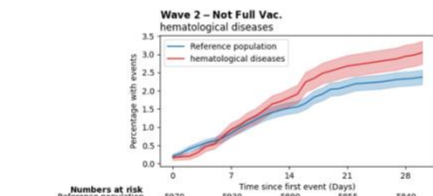
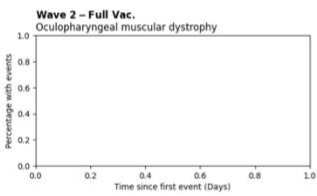
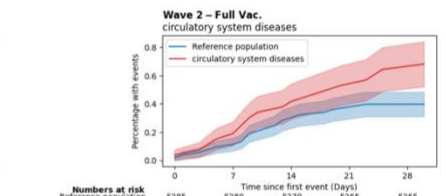
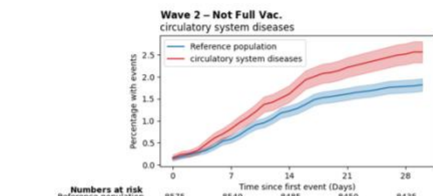
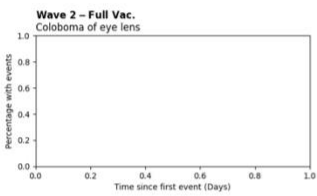
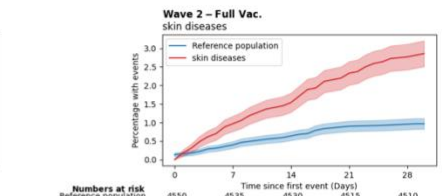
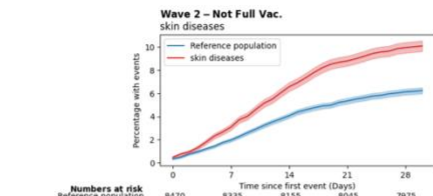
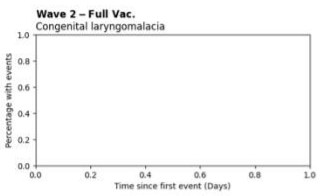
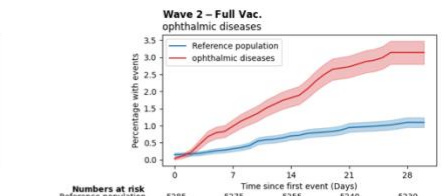
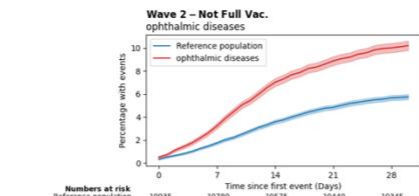
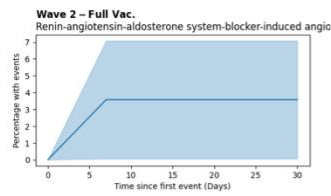
Numbers at risk

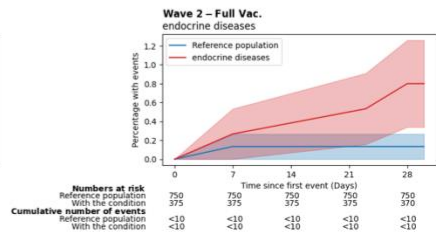
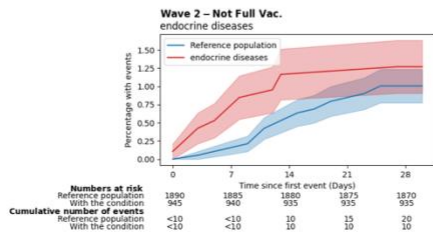
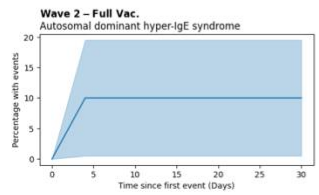
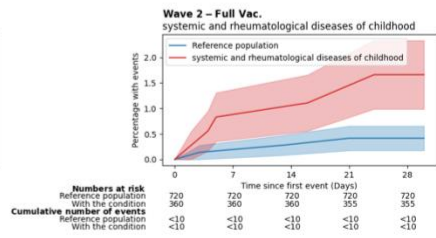
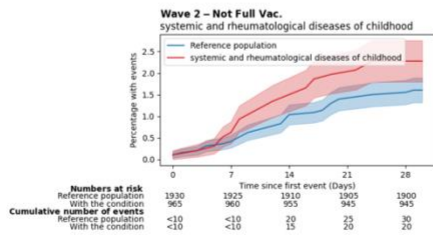
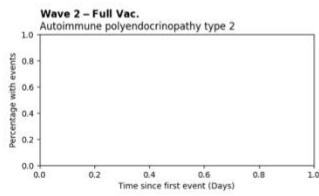
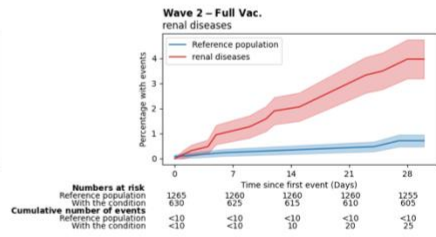
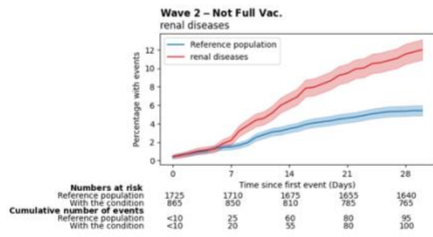
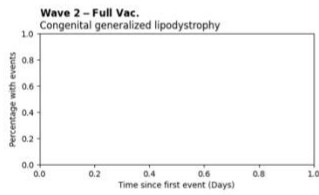
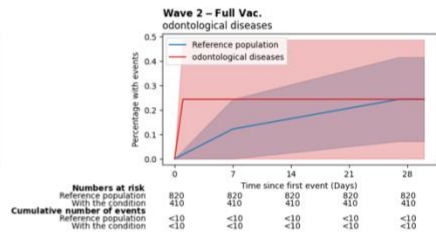
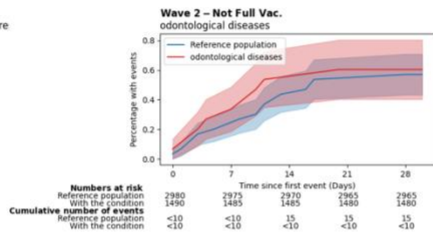
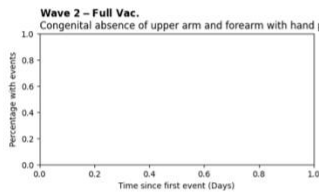
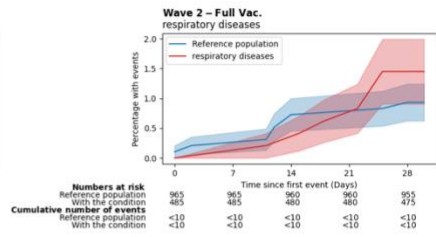
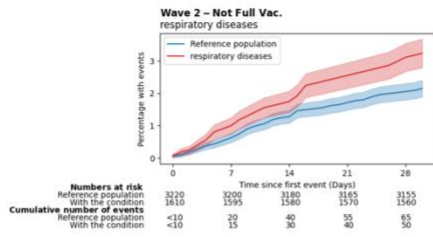
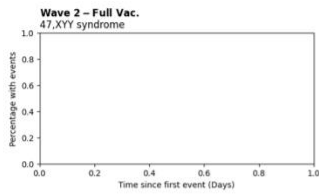
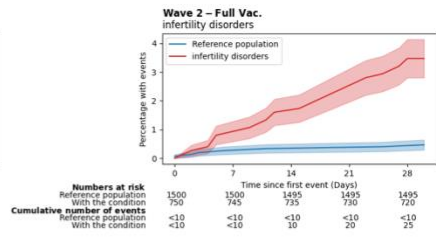
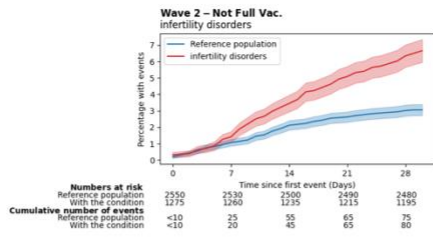
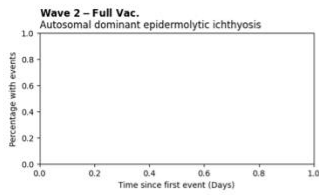
Time since first event (Days)	Reference population	With the condition
0	30	15
7	30	15
14	30	15
21	30	15
28	30	15

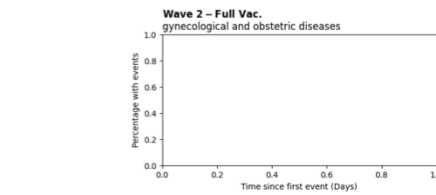
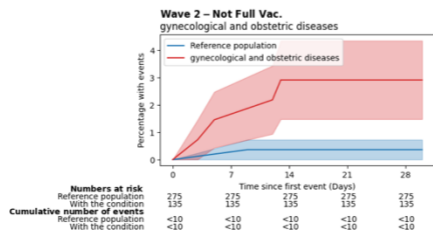
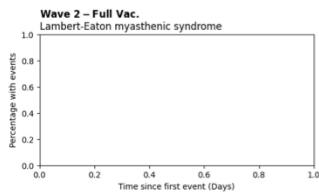
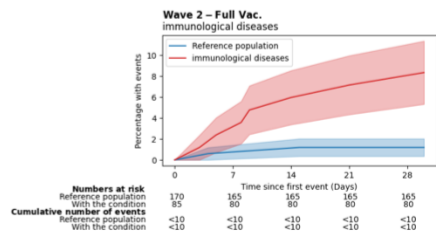
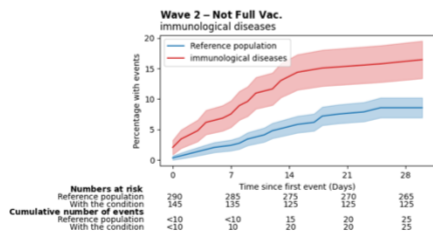
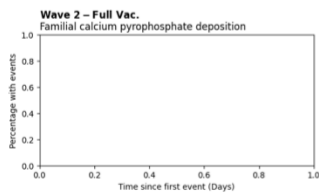
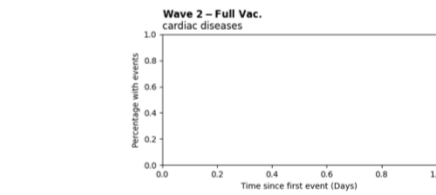
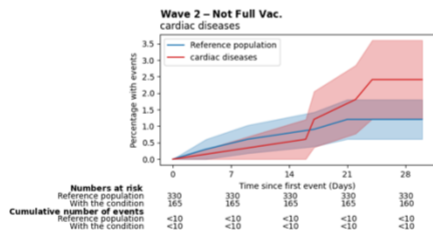
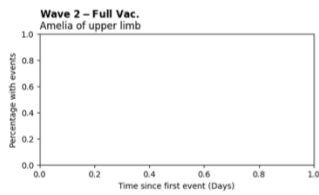
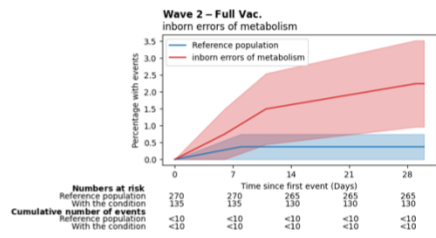
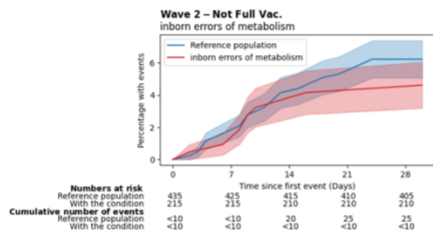
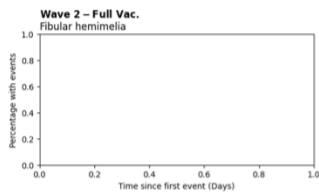
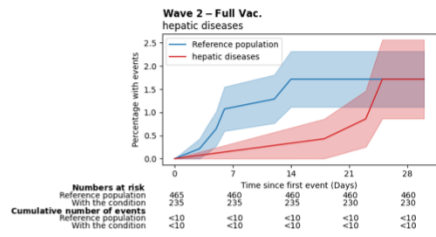
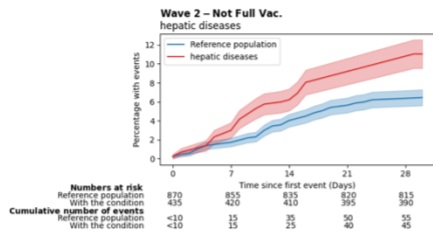
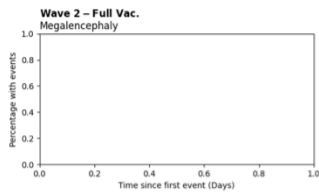
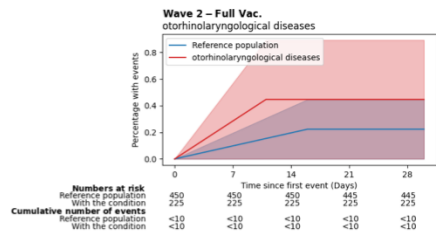
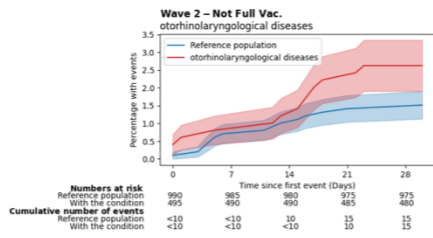
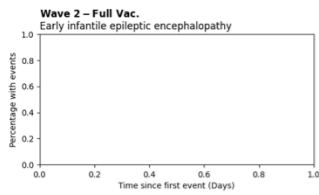
Cumulative number of events

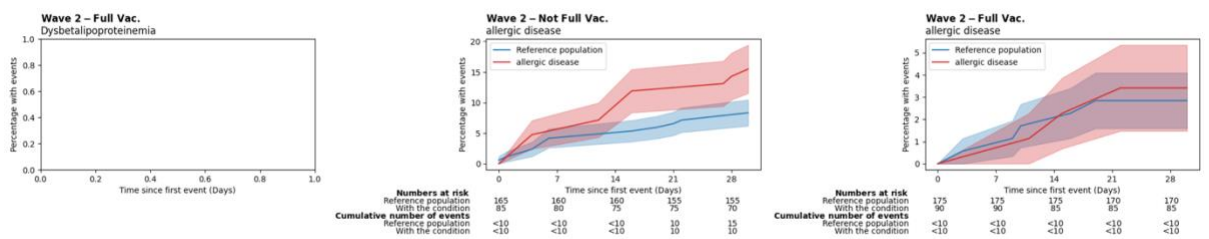
0	0	0
7	0	0
14	0	0
21	0	0
28	0	0











References

- 1 General Practice Extraction Service (GPES) Data for pandemic planning and research: a guide for analysts and users of the data. NHS Digital. <https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data> (accessed Jan 20, 2022).
- 2 Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010; **69**: 4–14.
- 3 Department of Health and Social Care. Payment by Results in the NHS: a simple guide. GOV.UK. 2013; published online March 25. <https://www.gov.uk/government/publications/simple-guide-to-payment-by-results> (accessed Jan 25, 2022).
- 4 Boyd A, Cornish R, Johnson L, Simmonds S, Syddall H, Westbury L. Understanding Hospital episode statistics (HES). London, UK: CLOSER 2017. <https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf>.
- 5 Burns EM, Rigby E, Mamidanna R, *et al.* Systematic review of discharge coding accuracy. *J Public Health* 2012; **34**: 138–48.
- 6 The processing cycle and HES data quality. NHS Digital. <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/the-processing-cycle-and-hes-data-quality> (accessed Jan 25, 2022).
- 7 Campbell A. Quality of mortality data during the coronavirus pandemic, England and Wales - Office for National Statistics. 2020; published online Dec 3. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/qualityofmortalitydataduringthecoronaviruspandemicenglandandwales/2020> (accessed Jan 26, 2022).
- 8 NHS Data Quality Maturity Index. NHS Digital. <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality> (accessed March 13, 2019).

-
- 9 Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373**: n826.
 - 10 [MI] National Data Opt-out, September 2021. NHS Digital. <https://digital.nhs.uk/data-and-information/publications/statistical/national-data-opt-out/september-2021> (accessed Jan 20, 2022).
 - 11 Baker C. Population estimates & GP registers: why the difference? 2016; published online Dec 12. <https://commonslibrary.parliament.uk/population-estimates-gp-registers-why-the-difference/> (accessed Jan 20, 2022).
 - 12 Summary of latest statistics. GOV.UK. <https://www.gov.uk/government/statistics/immigration-statistics-year-ending-september-2021/summary-of-latest-statistics> (accessed Jan 20, 2022).
 - 13 Orphanet. Orphanet: an online rare disease and orphan drug data base. Orphanet. <http://www.orpha.net> (accessed May 5, 2022).
 - 14 Alignments – orphadata. <https://www.orphadata.com/alignments/> (accessed March 2, 2023).
 - 15 Orphanet ICD-10 Coding Rules for Rare diseases. Orphanet. https://www.orpha.net/orphacom/cahiers/docs/GB/Orphanet_ICD10_coding_rules_R1_Nom_ICD_EP_06.pdf (accessed April 21, 2023).
 - 16 Nguengang Wakap S, Lambert DM, Olry A, *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020; **28**: 165–73.
 - 17 Thygesen JH, Tomlinson C, Hollings S, *et al.* COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. *Lancet Digit Health* 2022; **4**: e542–57.
 - 18 COG-UK/Mutation Explorer. <https://sars2.cvr.gla.ac.uk/cog-uk/> (accessed Feb 2, 2023).
 - 19 Official UK Coronavirus Dashboard. <https://coronavirus.data.gov.uk/> (accessed May 14, 2021).
 - 20 Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**: 515–26.
 - 21 NHS digital; COVID-shielding list documentation - Additions and subtractions. NHS digital (archive). <https://webarchive.nationalarchives.gov.uk/ukgwa/20220610000338/https://digital.nhs.uk/coronavirus/shielded-patient-list/methodology/additions-and-subtractions?key=> (accessed Feb 2, 2023).