Government of **Western Australia**
Department of **Health**

Bayesian modelling methods used for Public Health Atlas indicators

Epidemiology Directorate
Public and Aboriginal Health Division
Department of Health WA

December 2025

Table of Contents

List of Figures

List of Tables

Abbreviations

ASPR	Age Specific Rate
ASR	Age Standardised Rate
AUC	Area under the ROC Curve
BOD	Burden of Disease
CI	Credible Interval
DALY	Disability Adjusted Life Years
DIC	Deviance Information Criterion
EP	Exceedance Probability
HD	Health District
HDI	Highest Density Interval
HMDC	Hospital Morbidity Data Collection
HR	Health Region
HWSS	Health and Wellbeing Surveillance System
INLA	Integrated Nested Laplace Approximation
LGA	Local Government Area
PHA	Public Health Atlas
PHP	Public Health Profile
RSE	Relative Standard Error
ROC	Receiver-Operating Characteristic Curve
SRR	Standardised Rate Ratio
ST	Spatio-temporal
WA	Western Australia
WAIC	Watanabe–Akaike Information Criterion
YLD	Years Lived with Disability
YLL	Years of Life Lost

1. Introduction

This document describes the Bayesian modelling methods used to estimate epidemiological measures for small areas such as Local Government Areas (LGAs) and Health Districts (HDs) for the Western Australian (WA) Public Health Atlas (PHA) and local government area Health and Wellbeing Profiles.

1.1 What is Bayesian modelling?

A Bayesian model is a statistical model that uses probability to represent uncertainty within the model, both the uncertainty regarding the output (e.g., estimated disease/risk factor prevalence, counts, age standardised rates, age specific rates and standardised rate ratios) and the uncertainty regarding the input (e.g., raw data and parameters such as socioeconomic status, remoteness, and service accessibility) to the model.

Compared to conventional small area analysis methods, Bayesian methods have the following advantages:

- ✓ With the inclusion of prior distributions (i.e., existing evidence), researchers can include structured assumptions about spatial and temporal relationships to improve estimation where there are not enough cases to derive reliable estimates. Traditionally if there is insufficient data, estimates for the area would not be reported.
- ✓ Gaps in data can be filled where the issue of reliability using conventional methods to derive epidemiological measures has not been solved.
- ✓ Level of uncertainty can be reported for the measures/indicators for an area (this usually cannot be reported using conventional methods).
- ✓ Different models can be fitted based on different measure characteristics.

1.2 Levels of geography modelled using Bayesian methods

In areas with small population sizes and disease/condition counts, reliable epidemiological measures cannot be derived. This is the case for some LGAs and HDs within WA. To produce small area estimates with increased stability and certainty at the LGA and HD levels, Bayesian spatio-temporal modelling was used to obtain estimated disease/condition counts (modelled/fitted counts, not observed values) to derive associated epidemiological measures.

Due to the ready availability and high reliability of data and measures at Health Region (HR) and State geographical levels, measures presented at these levels were not derived via

Bayesian modelling but computed directly using observed disease/condition counts to calculate associated measures.

2. Data and model types

Three types of data (administrative data, survey data, and burden of disease (BOD) data) were modelled using Bayesian methods and are presented in the PHA. For a full list of indicators presented in the PHA for each data type, please refer to the PHA Data Dictionary.

Please see Section 4 (Interpreting model output) for further information on (i) understanding the difference between a credible interval (CI) and a confidence interval, and (ii) interpreting the 'Comparison to State' measure, both of which are mentioned in the following section.

Additionally, please refer to the Appendix 1 for mathematical notations used in this document.

2.1 Administrative data - Counts

Administrative data includes all indicators other than those mentioned in the HWSS and BOD sections of the PHA Data Dictionary. It includes indicators such as potentially preventable hospitalisations, tobacco-related hospitalisations, and hospitalisations due to injury and poisoning, among others, sourced from the WA Hospital Morbidity Data Collection (HMDC) as well as aetiological fractions derived by the Epidemiology Directorate, Department of Health WA. It also includes death data from the WA Mortality Dataset, Western Australian Notifiable Infectious Disease Database, and cancer incidence data identified through the WA Cancer Registry.

Key measures reported in the PHA for administrative data include modelled counts (with associated 95% CIs for modelled LGA and HD level data), Age Standardised Rates (ASRs) and associated 95% CIs for LGA and HD data and 95% confidence intervals for HR data, and Age Specific Rates (ASPRs). Another key measure reported in the PHA is a 'Comparison to State' of disease/condition rates. For LGA and HD level data the comparison to State is based on the exceedance probabilities (EPs) of the posterior draws (i.e., samples), to identify whether the disease/condition ASR is higher, lower, or similar between the specific LGA/HD compared to the State ASR. For HR level data, the comparison to State is based on the Standardized Rate Ratio (SRR) values which were calculated using raw unmodelled counts for the specific HR. For more information on interpreting the comparison to State variable, please refer to Section 4.2. Raw unmodelled

data at the HR level was suppressed in the following cases: (i) if the count was less than 6, the count was suppressed to protect privacy and data confidentiality, (ii) if the count was less than 20, the ASRs were suppressed because the derived rates were unreliable, and (iii) if the count was less than 6 the 'Comparison to State' variable was suppressed because the derived estimate was unreliable.

Modelled data was supressed in injury and poisoning related deaths for the intermediate category "intentional self-harm" to protect individual privacy.

Administrative data at the HR level also included data by Aboriginality, data combined for 10 years, and two 5-year combined sets (for example, data for 2015-2024 will be combined as 2015-2019, 2020-2024 and 2015-2024). Table 1 below outlines what measures are displayed in the PHA by geographical level.

Table 1. Administrative data measures presented in the PHA by geographical level

Geography Level	Measures (by area, year, and sex)
LGA	Modelled count and 95% CI
	ASR and 95% credible interval
	ASPR
	Comparison to State (based on posterior draw EPs)
HD	Modelled count and 95% credible interval
	ASR and 95% CI
	ASPR
	Comparison to State (based on posterior draw EPs)
HR	**Measures (by area, year/combined years, sex, and Aboriginality)**
	Raw count (counts <6, were suppressed)
	ASR and 95% confidence interval (ASRs where count <20, were suppressed)
	ASPR
	Comparison to State (based on SRR 95% confidence interval) (Comparison to State where count <6, were suppressed)

Bayesian modelled measures for LGA and HD level data were derived from the modelled counts using a mixture of forward and backward stepwise regression. ASRs and ASPRs were not estimated directly via the modelling process. Counts were estimated via a Poisson modelling process and those counts were then used to calculate ASRs, ASPRs and the associated credible intervals for these measures.

The input data required for this model included both population estimates and raw counts by area, year, sex, and age. The model was based on a spatio-temporal model with spatial effects and a space-time interaction (Bernardinelli et al., 1995). Combinations of fixed and random effects and interaction terms were fitted for each model. Other random effects tested included first order autoregressive (ar1) and independent and identically distributed (iid) effects. R Integrated Nested Laplace Approximation (INLA) code for the full model is included in Appendix 2. The process to select the best fitting model is outlined in Section 3 (Model fit checks). The base equation and full model (with all variables and interaction terms) used for administrative data is given in Table 2, and notations explained in Appendix 1. For the final regression equations for conditions please contact epi@health.wa.gov.au.

Table 2. Null and full administrative data model equations.

Model	Equation
Null model	y ~ intercept + # fixed effects age group + year + sex + area + offset (log(population)) y ~ Poisson(μ)
Full model	y ~ intercept + # random effects age group (iid) + sex (iid) + area (iid) + area (besag) + year (ar1) + # interaction terms area:sex + area:age group + area:year + age group:sex + sex:year + Age group:year + offset (log(population)) y ~ Poisson(μ)

2.2 Administrative data – Aetiologic fractions

Aetiologic fractions (AFs) were applied to hospital and death data to estimate for particular conditions that are attributable to tobacco use, alcohol use or illicit drug use (Ridolfo & Coles & Sun, 2021; Stevenson, 2001; Van Diemen et al, 2017). An AF is a proportion; its value is dependent on condition, age, gender, Aboriginality and remoteness of residence. For some conditions, such as alcoholic liver disease, the alcohol attributable AF value is 1.00, i.e. its occurrence is wholly attributable to alcohol use. That is, one alcohol related hospitalisation/death due to this condition was counted as one alcohol-attributable hospitalisation/death. Alcohol, tobacco and illicit drug-attributable conditions were identified using ICD-10-AM codes for principal diagnosis and/or external cause codes, for hospitalisations, and using cause of death code and/or multiple causes of death codes, for deaths. A list of conditions can be found the relevant publications (References):

- Tobacco Attributable Fractions – Coles & Sun, 2021
- Alcohol Attributable Fractions – Van Diemen et al, 2017
- Illicit Drug Attributable Fractions – Ridolfo & Stevenson, 2011

For tobacco-related hospitalisations and deaths, and drug-related hospitalisations and deaths, Poisson count models, as described for administrative models, were run for each condition, then aetiologic fractions were applied to estimate the attributable fraction for that condition. The condition level attributable fractions were then summed to give the total estimated attributable counts.

Alcohol-attributable aetiological fractions required data disaggregated by identification as an Aboriginal person. For these, a Poisson model was used to estimate the counts (posterior counts, *PC*), and then a binomial model, as described in the survey section, was used to estimate the Aboriginal person to non-Aboriginal person proportion of counts (posterior proportions, *PP*). These proportions were then applied to the count model estimates to estimate Aboriginal person counts and non-Aboriginal person counts:

$$PP * PCs = PD_A$$

$$(1 - PP) * PC = PD_B$$

The aetiologic fractions for Aboriginal people (AF_A) and non-Aboriginal people (AF_B) were applied to these counts to give the posterior draws of the attributable fraction counts for Aboriginal people (PD_{AFA}) and non-Aboriginal people (PD_{AFB}):

$$PD_A * AF_A = PD_{AFA}$$

$$PD_B * AF_B = PD_{AFB}$$

The two posterior draw estimates are then summed to give the total attributable fraction posterior draws, PD_{AF}:

$$PD_{AFA} + PD_{AFB} = PD_{AF}$$

The attributable fraction posterior draws for each condition (PS_{AFi}, PD_{AFii}, …. PD_{AFn}) were summed to give the total alcohol-attributable fractions, PD_{AF} as a whole:

$$PD_{AF} = \sum_1^n PD_{AFi}$$, for *i* in 1 to *n*.

2.3 Survey data

Survey data consists of data obtained from the WA Health and Wellbeing Surveillance System (HWSS). It includes indicators related to lifestyle behaviours such as current smoking, and alcohol consumption at levels considered high risk for long-term and short-term alcohol related harm, among others. A full list of survey data indicators included in the PHA can be viewed in the PHA Data Dictionary in the HWSS section.

Key measures reported in the PHA for survey data include prevalence by area, year, and sex, and prevalence by age group with associated 95% credible intervals for modelled LGA and HD level data and associated 95% confidence intervals for HR level data. Prevalence measures with a Relative Standard Error (RSE) greater than 50% or a prevalence of zero (for HR level data) were excluded and not presented in the PHA due to privacy policies, or to withhold an unreliable prevalence value (see Hogg & Cramb (2022) for further information on modelled RSE).

Additionally, a 'Comparison to State' measure is reported. For LGA and HD level modelled data the comparison to State is based on the EPs of the posterior draws, to identify whether the prevalence is higher, lower, or similar between the specific LGA/HD compared to the State. For HR level data, the comparison to State is based on evaluating the 95% confidence intervals of the HR with the State prevalence. For more information on interpreting the comparison to State variable, please refer to Section 4.2. Table 3 below outlines what measures for survey data are displayed in the PHA by geographical level.

Table 3. Survey data measures presented in the PHA by geographical level

Geographical Level	Measures (by area, year, and sex)
LGA	Modelled prevalence with 95% CI
	Modelled prevalence by age group with 95% credible interval (by area and year only)
	Modelled counts for selected conditions (PHP only)
	Comparison to State (based on posterior draw EPs)
HD	Modelled prevalence with 95% credible interval
	Modelled prevalence by age group with 95% credible interval (by area and year only)
	Modelled counts for selected conditions (PHP only)
	Comparison to State (based on posterior draw EPs)
HR	Raw prevalence with 95% confidence interval
	Raw prevalence by age group with 95% confidence interval (by area and year only)
	Comparison to State (based on prevalence 95% confidence interval comparison of HR and State)

Bayesian modelled measures for LGA and HD level data were derived using a hierarchical model incorporating sampling weights (Chen, Wakefield & Lumley, 2014). The process was similar to that for administrative counts in that the modelled prevalences were obtained using a mixture of forward and backward stepwise regression. Prevalences, and the associated credible intervals, were estimated via a Binomial modelling process.

Note that all survey measures are binary (i.e., yes, or no) and the model is a logistic model where the dependent variable is a binary outcome measure (e.g., current smoker or not). Combinations of fixed and random effects and interaction terms were fitted for each model. The process to determine best model fit is outlines in section (3. Model fit checks).

The advantage of this model is that it incorporates survey weights into the model. The input data required for this model includes individual level survey data, post strata data (all unique combinations of the covariates used in the model and census populations), and raw survey weights available from HWSS data. Note that the model itself does not output

counts, but probabilities, which can be used to derive estimated counts. The equation and output measures for the model are presented in Table 4. For information on the final regression equations for each condition please contact epi@health.wa.gov.au.

Table 4. Null and full survey data model equations.

Model	Equation
Null model	y ~ intercept + # fixed effects age group + year + sex + area y ~ Binomial(N_trials, proportion)
Full model	y ~ intercept + # random effects age group (iid) + sex (iid) + area (iid) + area (besag) + year (ar1) + # interaction terms area:sex + area:age group + area:year + age group:sex + sex:year + Age group:year + # area level covariates Total proportion of female population + Proportion of people with low income+ Remoteness[1] Socio-economic disadvantage category Indigenous population proportion Proportion of people with tertiary education Proportion of population with occupation as labourer or other manual workers Total proportion of males aged 35-39 years Total proportion of females aged 15-19 years y ~ Binomial (N_trials, proportion)

[1] Note: HD models do not have the remoteness covariate

2.4 Burden of disease data

BOD data is modelled using both administrative data sourced from the WA Mortality Dataset and HMDC, and survey data from the HWSS. A full list of BOD data indicators included in the PHA can be viewed in the PHA Data Dictionary in the BOD section.

Key measures reported for BOD data include YLL and YLD counts at the LGA, HD and HR geographical levels, YLL and YLD ASRs at the LGA, HD and HR geographical levels with associated 95% CIs presented at the LGA and HD levels only, and a 'Comparison to State' measure at the LGA and HD levels only. The comparison to State measure is based on the EPs of the posterior draws to identify whether the YLL or YLD ASRs are higher, lower, or similar between the specific LGA/HD compared to the State. For more information on interpreting the comparison to the State measures, please refer to Section 4.2.

Additionally at the HR level, Disability Adjusted Life Years (DALY) counts, and ASRs are also reported. DALY counts/ASRs are the sum of associated YLL and YLD counts/ASRs. At the HR level, YLL, YLD and DALY counts of minor category diseases/conditions are presented in a pie chart as a percentage contribution to the total YLL, YLD or DALY count for the intermediate category of interest for the total population. For each HR and year, the leading ten conditions of YLL, YLD and DALY burden are also displayed in the PHA by sex, as both a count, percentage, and ASR (per 100,000). Additionally, the contribution (%) of total YLL, YLD and DALY burden is presented by disease groups (intermediate category) in a tree-map figure.

Table 5 outlines what measures for BOD data are displayed in the PHA by geographical level. For LGA and HD level data, both the administrative and survey models were used. For YLL, the raw YLL estimates were modelled. For YLD the raw counts were modelled and then the modelled counts were transformed to YLD estimates (AIHW, 2023). Fitted counts and fitted probabilities obtained from the models were then used to calculate YLL or YLD as required.

Table 5. Burden of disease data measures presented in the PHA by geographical level

Geography Level	Measures (by area, year, and sex)
LGA	Modelled YLL count with 95% CI
	YLL ASR with 95% credible interval
	Modelled YLD count with 95% credible interval
	YLD ASR with 95% credible interval
	Comparison to State (based on posterior draw EPs)
HD	Modelled YLL count with 95% credible interval
	YLL ASR with 95% credible interval
	Modelled YLD count with 95% credible interval
	YLD ASR with 95% credible interval
	Comparison to State (based on posterior draw EPs)
HR	Raw YLL count
	YLL ASR
	Contribution (%) of minor category diseases/conditions to overall YLL count for intermediate category of interest for total population (by area and year only)
	Leading 10 conditions of YLL burden (count (%) and ASR (per 100,000)
	Contribution (%) of disease groups (intermediate categories) to overall YLL burden
	Raw YLD count
	YLD ASR
	Contribution (%) of minor category diseases/conditions to overall YLD count for intermediate category of interest for total population (by area and year only)
	Leading 10 conditions of YLD burden (count (%) and ASR (per 100,000)
	Contribution (%) of disease groups (intermediate categories) to overall YLD burden

	Raw DALY count
	DALY ASR
	Contribution (%) of minor category diseases/conditions to overall DALY count for intermediate category of interest for total population (by area and year only)
	Leading 10 conditions of DALY burden (count (%) and ASR (per 100,000)
	Contribution (%) of disease groups (intermediate categories) to overall DALY burden

3. Model fit checks

The following checks were performed on model outputs to assess the validity/goodness of fit of modelled results:

- Comparing observed counts and rates to the modelled counts and rates to see if the modelled results were plausible.
- Reducing unnecessary complexity of models by running all combinations of variables and removing terms where improved model fit was not observed. Models with the smallest WAIC and DIC scores were selected as final models.
- Simulation of counts from the best model formula was conducted to ensure the model provided a good representation of the data.
- Model performance assessment using AUC and ROC curves (survey data only).

If all attempts to improve model fit as outlined above failed, and a model for an individual disease/condition was not a good fit, the disease/condition was excluded from the PHA and therefore not shown in the drop-down disease/condition list.

4. Interpreting model output

4.1 Credible intervals

In Bayesian statistics, credible intervals (CIs) are a way to quantify the uncertainty or variability in the estimates of unknown parameters. Unlike frequentist confidence intervals, which are based on the long-run behaviour of repeated sampling, Bayesian CIs are derived

from the posterior distribution of the parameter, which represents our updated knowledge about the parameter after considering the observed data and any prior information.

To understand CIs, suppose we have a Bayesian model where we are interested in estimating the mean (μ) of a population. We have some prior belief about the possible values of μ, which is expressed as a prior distribution. After collecting data, we update our knowledge about μ and obtain the posterior distribution, which incorporates both the prior distribution and the observed data (Figure 1).

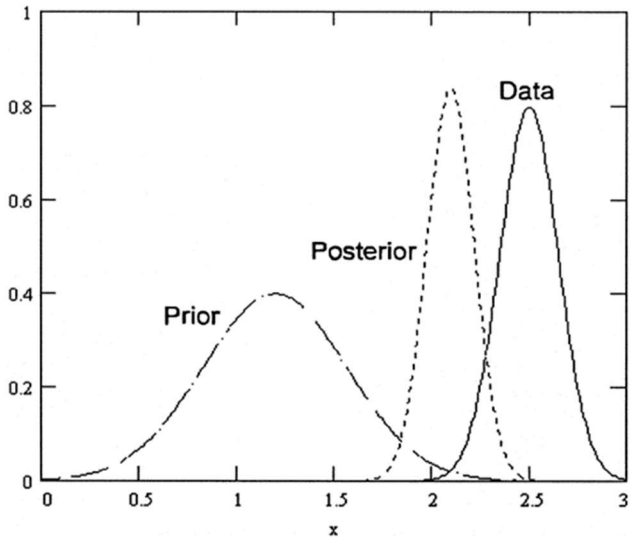

Figure 1. The relationship between Bayesian prior distribution, the data, and the posterior distribution (Matthews, 2001).

A CI is a range of values from the posterior distribution that is associated with a certain degree of credibility or probability. It represents a range of plausible values for the parameter, given the data and the prior information. For example, a 95% CI is constructed such that it contains the true parameter value with a probability of 0.95. (See a comparison of traditional 95% confidence interval vs. Bayesian 95% credible interval in https://www.statsdirect.com/help/basics/confidence_interval.htm).

The construction of CIs involves determining the limits of the interval based on the probability density of the posterior distribution ascertained from the INLA algorithms. The Bayesian models used by the Department of Health uses the Highest Density Interval (HDI), which constructs the CI to contain the highest density. This is a pragmatic method for skewed distributions (Figure 2).

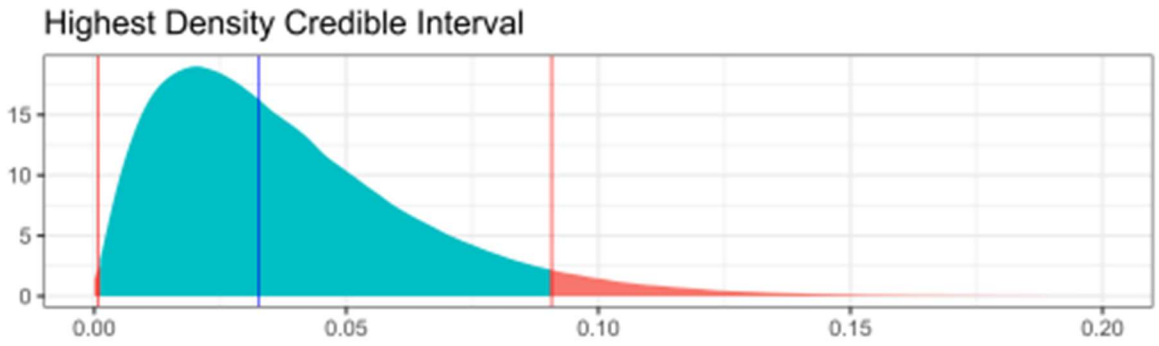

Figure 2. Credible Interval (red vertical lines) constructed using Highest Density Interval. The median is indicated by the blue line (Hogg & Cramb, 2022).

It's important to note that credible intervals (CIs) in Bayesian analysis, unlike frequentist statistics, measure uncertainty using available data and prior information, and are not based on repeated sampling. CIs offer a probabilistic assessment of a parameter's plausible values, influenced by the choice of prior distribution, and facilitate decision-making based on the entire posterior distribution rather than just point estimates.

Figure 3 shows the estimated ASRs and credible intervals for age standardised potentially preventable hospitalisation rates by health districts in WA. The length of the bar represents the ASR and the two whiskers, the lower and upper credible intervals. The wider the credible interval range, the more uncertain the ASR estimate, meaning the ASR for the area is less reliable compared to other areas. (Note that the darker the bar, the higher the ASR). Usually, areas with smaller populations have more uncertainty in their rates, often resulting in wider credible interval ranges. For example, as shown in Figure 3, the CI width from a small population area like Southern Wheatbelt is wider than that for an area with large population like Inner South, especially if their rate estimates are similar.

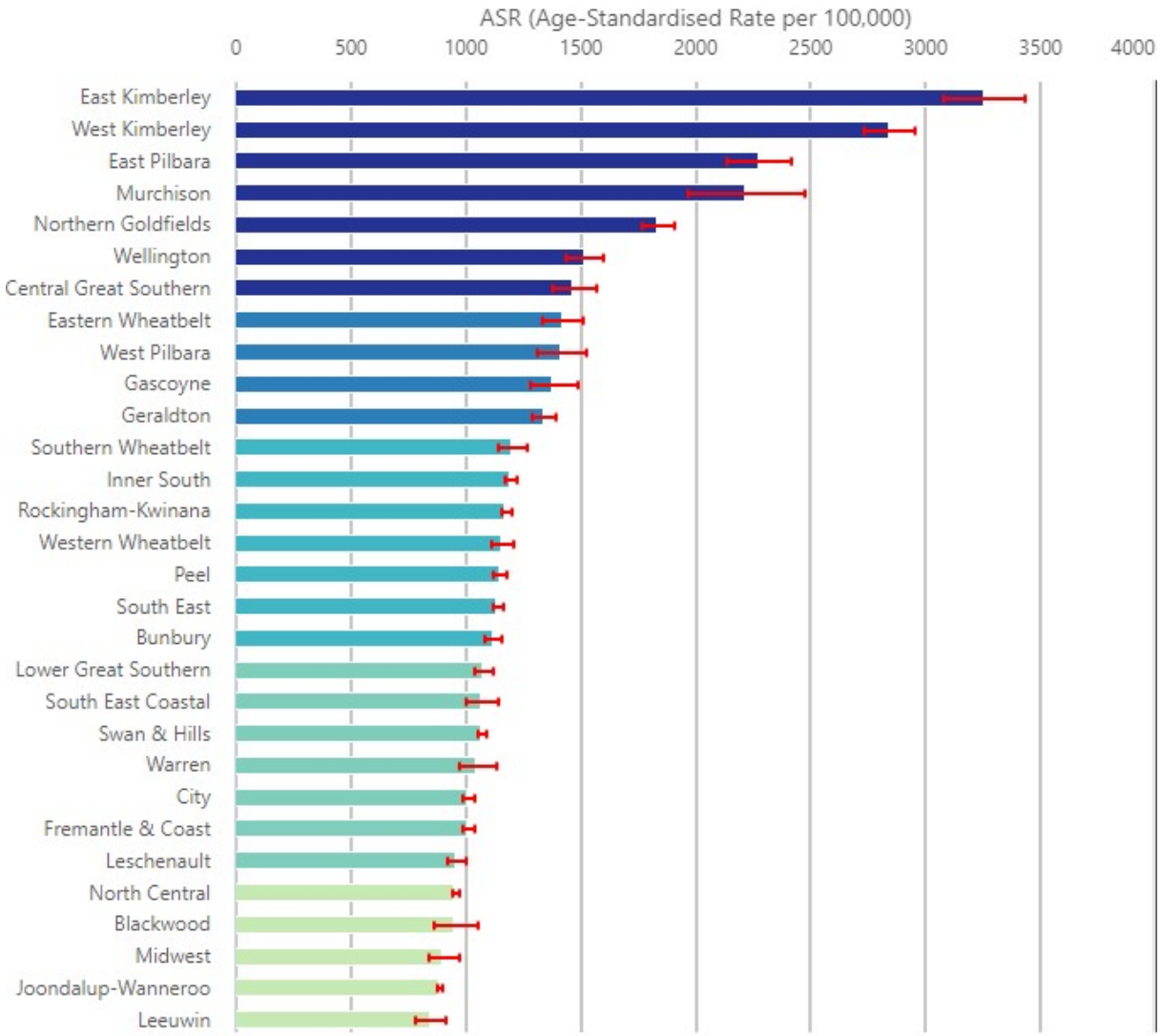

Figure 3. Comparison of credible Intervals for ASR by health district in WA in PHA.

4.2 Comparison to State

For all data types (administrative, survey and BOD data) the comparison to State value (higher/lower/similar) for LGAs and HDs is determined from the exceedance probabilities (EPs). This is the probability of the posterior draws being above a certain value. This was derived from the posterior draws using:

$$EP = \frac{1}{D}\sum_{d} I\left(\theta^{(d)} > c\right)$$

Where $I\left(\theta^{(d)} > c\right)$ is equal to 1 if $\theta^{(d)}$ is larger than the baseline value c and zero if $\theta^{(d)}$ is smaller than c (Hogg & Cramb, 2022). The EP was used to indicate whether the ASR, prevalence, age standardised YLL or age standardised YLD in a particular area is significantly higher than the state measure. EP values above 0.8 (i.e., 80% of the posterior)

were considered likely to be above the state value and therefore ‘higher’. Values below 0.2 (i.e., 20% of the posterior) on the other hand were considered likely to be below the state value and therefore ‘lower’. Values between 0.2 and 0.8 were then considered to be ‘similar’ to the state rate.

For HR level data however, where Bayesian modelling was not used, the comparison to State value was determined differently for each of the data types as detailed below.

Administrative Data

Comparison to State value was determined by analysing the 95% confidence interval of the SRR for the HR of interest with the State.

SRR is the ratio of observed disease/condition counts to expected disease/condition counts,

$$SRR_{HR} = \frac{\sum y}{\sum E}$$

where y is the observed disease/condition counts in a particular HR, and E (expected counts) is the state age specific rates multiplied by the HR age specific populations.

An SRR greater than 1 therefore indicates that the HR rate is higher; an SRR lower than 1 indicates the HR rate is lower.

To determine the appropriate value for the comparison to State, the 95% confidence intervals were used to ensure the difference was significant.

- If the SRR lower confidence interval value was greater than 1, then the HR rate was significantly higher than the State.
- If the SRR upper confidence interval value was less than 1, then the HR rate was significantly lower than the State.
- If the SRR confidence interval included 1 in its range, the HR rate was similar to the State.

Please note, the term SRR is used in this document for the sake of simplicity in describing this type of measure; and the estimation of similar measures such as standardised incidence ratio (e.g., for cancer incidence) or standardised mortality ratio (e.g., death data) will follow the identical process as for SRR.

Survey data

Comparison to State value was determined by analysing the 95% confidence interval of the prevalence for the HR of interest with the State. If the HR and the State prevalence confidence intervals overlapped, then the prevalence was similar to the State.

If the HR and the State prevalence confidence intervals did not overlap, then the prevalence estimates were significantly different. To determine whether the HR prevalence was higher or lower than the State, comparisons of the upper and lower limits of the confidence intervals were used as follows:

- If the HR prevalence was higher than the State prevalence and the lower limit of the confidence interval of the HR prevalence was higher than the upper limit of the State prevalence, then the HR prevalence was significantly higher than the State.
- If the HR prevalence was lower than the State prevalence and the upper limit of the confidence interval of the HR prevalence was less than the lower limit of the State prevalence, then the HR prevalence was significantly lower than the State.

Burden of disease data

A comparison to State measure is not presented in the PHA for BOD data at the HR level.

5. Advantages of modelled data over raw unmodelled data

Modelled estimates for LGAs and HDs include estimated count, age specific rate, age standardised rate, and prevalence. **Important note**: Modelled data are estimates only and are not actual counts, rates, or prevalence. Caution should therefore be exercised when using the modelled data.

There are several benefits of using modelled data. Firstly, modelled measures can be more stable compared to raw measures, particularly in areas with small counts and/or populations. Figure 4 shows a comparison of three prevalence measures (raw, modelled, and state prevalence) of adults who drink alcohol at levels that increase the risk of long-term harm for a particular LGA. The modelled prevalence (red) is a ‘smoothed’ version of the raw prevalence (green). Overall heterogeneity in modelled estimates were reduced, depicting a similar trend to the State prevalence (blue). Additionally, the modelled prevalence had increased stability and certainty compared to the raw data as indicated by

the narrower CIs for the modelled data compared to the wider confidence intervals observed in the raw data. This increased stability, mitigates the need for data suppression due to small counts or unreliable estimates. It also enables data users to observe clearer trends over time thereby contributing to a more comprehensive and insightful set of results.

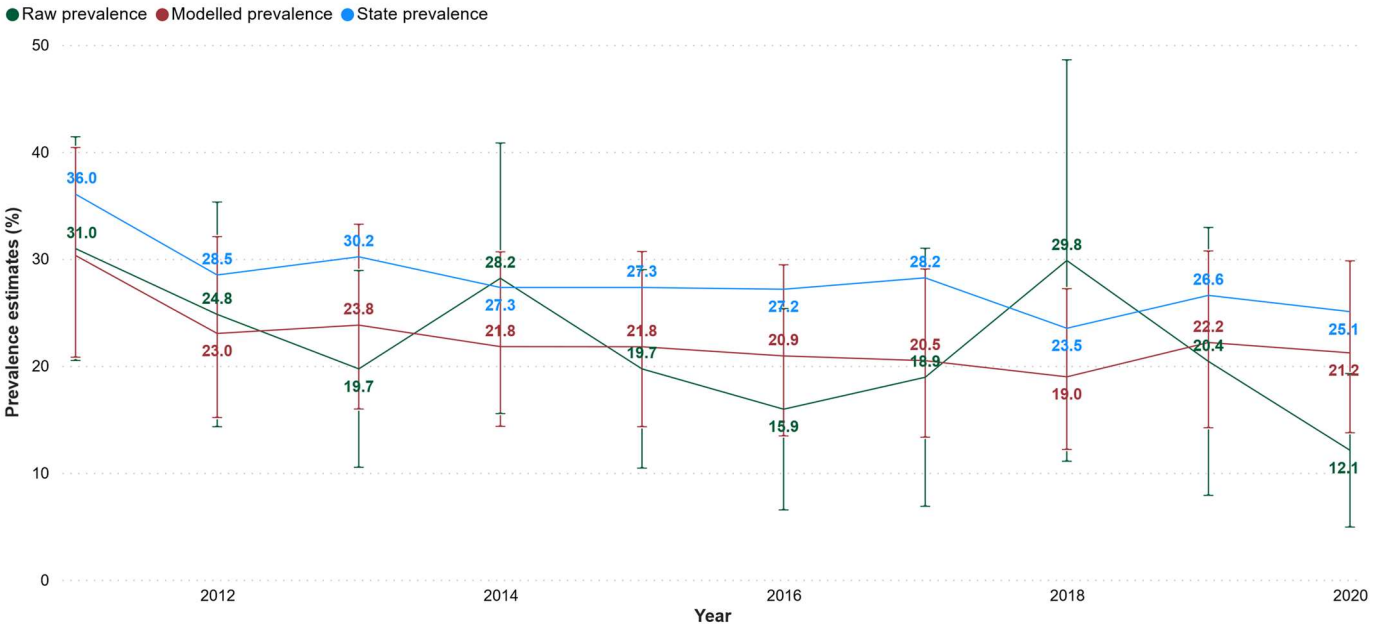

Figure 4. Raw, modelled, and state prevalence for long-term alcohol related harm in a WA LGA from 2011-2020

To further highlight the capabilities of Bayesian methods, Figure 5 illustrates a comparison in data coverage in ASRs for hospitalisations due to injury and poisoning across WA LGAs before and after applying Bayesian methods. When mapping the raw data, Figure 5(a) shows there are several gaps in data coverage (i.e., not reportable ASRs) primarily due to small event counts and/or small population sizes. Conversely, mapping the modelled data using Bayesian methods, Figure 5(b) shows complete coverage, leaving no data gaps across WA.

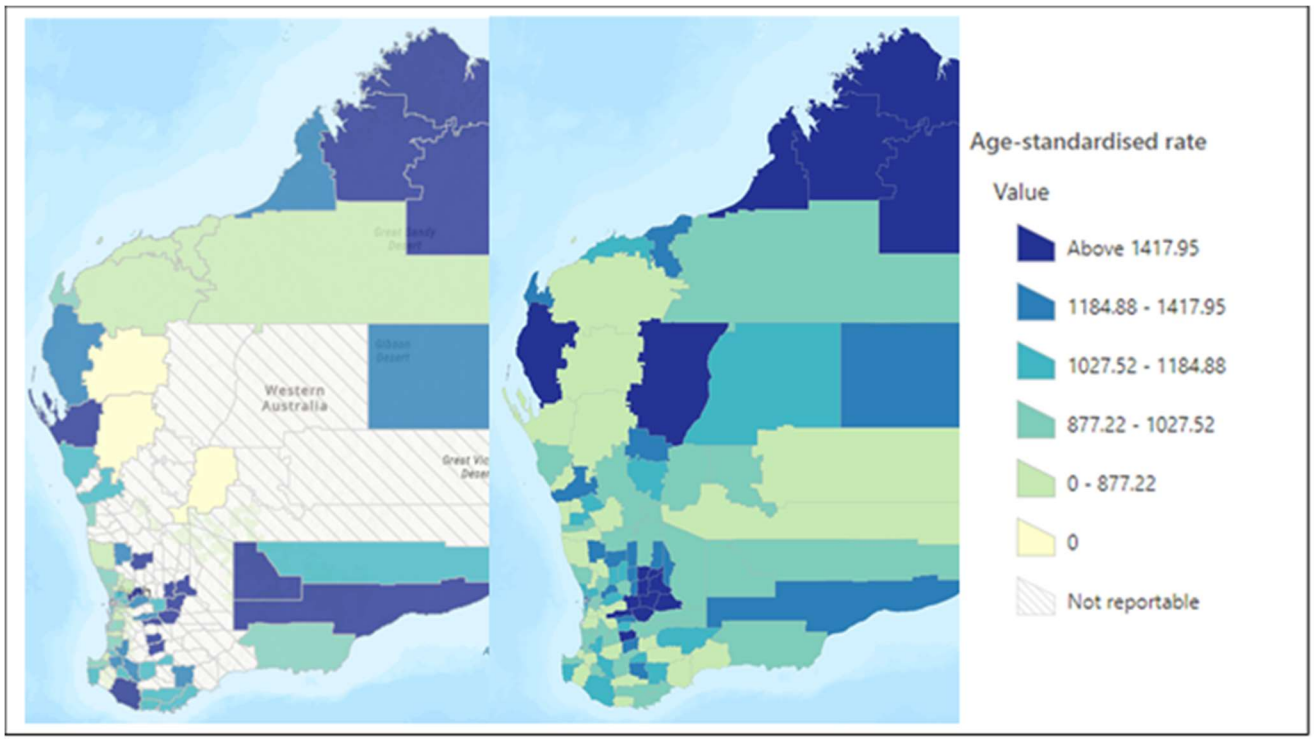

Figure 5. Data coverage across WA LGAs before (a) and after (b) using Bayesian methods to estimate ASRs for hospitalisations due to accidental falls in 2019

6. Disadvantages and limitations

The modelling itself comes with some limitations in that the priors used in the models for the spatio-temporal terms in this project, potentially smooth over all adjacent areas and time points, with the extent of smoothing determined by the data itself. These priors have previously been shown to work well in the Australian context (Cramb et al., 2020). However, if discontinuities in rates between adjacent areas are expected, alternative smoothing priors that allow for large differences in rates between neighbours are needed. While many priors that allow discontinuities are available, these have resulted in convergence difficulties for certain areas when used with sparse Australian health data (Cramb et al., 2020), so were not considered for this project.

While it is important to acknowledge these disadvantages and limitations, the benefits of producing Bayesian modelled data, outweigh its potential downfalls. The modelled estimates allow for a more complete picture of population health outcomes and trends across WA to be observed which in turn, provides essential epidemiological measures to inform public health planning, policy, and decision making that would otherwise not be available.

Further information

For feedback, queries or further analysis related to this document, contact epi@health.wa.gov.au.

Bayesian modelling project Deliverable 2: Modelling recommendations

Publication: 'Improving the spatial and temporal resolution of burden of disease measures with Bayesian models'

Hogg, J., Staples, K., Davis, A., Cramb, S., Patterson, C., Kirkland, L., Gourley, M., Xiao, J., & Sun, W. (2024). Improving the spatial and temporal resolution of burden of disease measures with Bayesian models. *Spatial and Spatio-temporal Epidemiology, 49, 100663*. https://doi.org/10.1016/j.sste.2024.100663.

Australian Cancer Atlas

The Australian Cancer Atlas (https://atlas.cancer.org.au/) has detailed information on how the Bayesian modelling is conducted and how output from the models is interpreted. The methodological document (Duncan et al, 2020) is a useful reference.

References

Australian Institute of Health and Welfare, 2023, Australian Burden of Disease Study; Technical notes, accessed 29 August 2024. https://www.aihw.gov.au/reports/burden-of-disease/australian-burden-of-disease-study-2023/contents/technical-notes.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. and Songini, M. (1995), Bayesian analysis of space-time variation in disease risk. *Statist. Med.*, 14, pp. 2433-2443. https://doi.org/10.1002/sim.4780142112

Chen C, Wakefield J, Lumely T. The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat Spatiotemporal Epidemiol*. 2014 11, pp. 33-43. https://doi.org/10.1016/j.sste.2014.07.002

Coles, E. & Sun, W. (2021). Methodology for developing tobacco smoking aetiological fractions for Western Australia. Perth: Department of Health WA.

Cramb, S., Duncan, E., Baade, P., & Mengersen, K. (2020). A comparison of Bayesian spatial models for cancer incidence at a small area level: Theory and performance. *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*. *Springer, Cham, pp. 245-274*. https://eprints.qut.edu.au/203868/

Duncan, E., Cramb, S., Baade, P., Mengersen, K., Saunders, T., & Aitken, J. (2020). Developing a Cancer Atlas using Bayesian Methods: A Practical Guide for Application and Interpretation. https://atlas.cancer.org.au/developing-a-cancer-atlas/

Hogg, J., & Cramb S. (2022). Bayesian modelling project, Deliverable 2: Modelling recommendations. (unpublished)

Hogg, J., Staples, K., Davis, A., Cramb, S., Patterson, C., Kirkland, L., Gourley, M., Xiao, J., & Sun, W. (2024). Improving the spatial and temporal resolution of burden of disease measures with Bayesian models. *Spatial and Spatio-temporal Epidemiology, 49, 100663*. https://doi.org/10.1016/j.sste.2024.100663.

Matthews, R. (2001). Methods for Assessing the Credibility of Clinical Trial Outcomes. *Drug Information Journal*, *35(4):1469-1478*. https://doi.org/10.1177/009286150103500442

Ridolfo, B. & Stevenson, C. (2001). The quantification of drug-caused mortality and morbidity in Australia,1998. AIHW cat. no. PHE 29. Canberra: AIHW (Drug Statistics Series no. 7). https://www.aihw.gov.au/reports/alcohol-other-drug-treatment-services/quantification-drug-caused-mortality-1998/summary

Van Diemen, A., Jian, L., Xiao, J. and Somerford, P. (2017). Methodology for Developing Western Australia Specific Alcohol-related Aetiological Fractions. Epidemiology Directorate, Department of Health Western Australia. Perth, Western Australia.

https://www.health.wa.gov.au/~/media/Files/Corporate/general-documents/Population-health/PDF/Specific-Alcohol-related-Aetiological-Fractions-Report.pdf

Appendix 1 - Mathematical Notation

Table A1. Legend for notations used in equations

Notation	Description
y	Count
μ	Mean
w	Weighted count (for survey weighted data)
n	Number of trials
p	Probability of success
E	Expected counts for area ($_i$) and year ($_t$)
∑	Sum
SRR_{HR}	Standardised Rate Ratio at the Health Region geographical level
D	Total number of posterior draws
I	Identity function (*I* is equal to 1 if $\left(\theta^{(d)} > c\right)$ is true, or equal to 0 if $\left(\theta^{(d)} > c\right)$ is false)
$\theta^{(d)}$	dth posterior draw of theta
c	Baseline value

Appendix 2 – R code for models

Data inputs

Note: The R INLA package requires separate ids for each variable in the model formula. The base variable is copied and renames to meet this requirement but is identical. For example, M_id = M_id2, M_id 3 ….

Administrative count models:

DATA – a data frame with the following variables:

- T_id – data year (numeric)
- A_id – age group (factor)
- M_id - area (LGA or HD) (numeric)
- G_id – sex (factor)
- Y - observed count (numeric)
- N – Population (numeric)

Administrative attributable fraction models:

DATA – a data frame with the following variables:

- T_id – data year (numeric)
- A_id – age group (factor)
- M_id - area (LGA or HD) (numeric)
- G_id – sex (factor)
- Y - observed count (numeric)
- N – Population (numeric)

DATA_2 – a data frame with the following variables (only required for alcohol-attributable fractions):

- T_id – data year (numeric)
- A_id – age group (factor)
- M_id - area (LGA or HD) (numeric)
- G_id – sex (factor)
- Y - observed count for the aboriginal population (numeric)
- N_trials – Total count (for alcohol-attributable models) (numeric)

Survey models:

DATA – a data frame with the following variables:

- T_id – data year (numeric)
- A_id – age group (factor)
- M_id - area (LGA or HD) (numeric)
- G_id – sex (factor)
- Y – Total of positive samples (numeric)
- N_trials – Total of weighted samples (numeric)

Prior specification

```
HC.prior  = "expression:
  sigma = exp(-theta/2);
  gamma = 10;
  log_dens = log(2) - log(pi) - log(gamma);
  log_dens = log_dens - log(1 + (sigma / gamma)^2);
  log_dens = log_dens - log(2) - theta / 2;
  return(log_dens);"
```

INLA formulas

```
formula_null<- Y ~ A_id + G_id + T_id + M_id


formula_best<- Y ~
  f(A_id, model = "iid",
    hyper = list(prec = list(prior = HC.prior)))+
  f(G_id, model = "iid",
    hyper = list(prec = list(prior = HC.prior)))+
  f(M_id, model ="besag",
    graph=map.adj,
    param=c(0.5,0.008),
    hyper = list(prec = list(prior = HC.prior))) +
  f(M_id2, model = "iid",
    hyper = list(prec = list(prior = HC.prior)))+
  f(T_id,
    model = "ar1",
    hyper = list(theta1 = list(prior = HC.prior),
                 theta2 = list(prior = "betacorrelation",
                               param = c(1,1)))) +
  f(M_id3, G_id2,
    model="iid")+
  f(M_id4, A_id2,
    model="iid")+
  f(M_id5, T_id2,
    model = "iid")+
```

```
  f(G_id3, A_id3,
    model="iid")+
  f(G_id4, T_id3,
    model="iid")+
  f(A_id4, T_id4,
    model="iid")
```

INLA model for administrative count models

```
model<-inla(formula=formula,
            data = DATA,
            family = c("poisson"),
            offset = log(DATA$N), #total of weighted samples
            control.compute = list(dic = TRUE,
                                   waic = TRUE,
                                   return.marginals.predictor=TRUE,
                                   config = TRUE),
            control.predictor=list(compute=TRUE,
                                   link=1))
```

INLA formula for alcohol-attributable fraction, binomial model

```
model<-inla(formula=formula,
            data = DATA, # DATA2 for alcohol-attributable binomial models)
            family = c("binomial"),
            Ntrials=N_trials, #total count
            control.compute = list(dic = TRUE,
                                   waic = TRUE,
                                   return.marginals.predictor =TRUE,
                                   config = TRUE),
            control.predictor = list(compute=TRUE,
                                   link = 1))
```

INLA formula for survey models

```
model<-inla(formula = formula,
          family = "xbinomial", #xbinomial if input is fractional
          data=DATA,
```

```
        N_trials=w, # total of weighted samples
        control.compute = list(dic = TRUE,
                               waic = TRUE,
                               return.marginals.predictor=TRUE,
                               config = TRUE),
        control.predictor=list(compute=TRUE,
                               link=1))
```

Sampling from the model (posterior draws)

```
#get samples
sample<-inla.posterior.sample(n=1000, model)

# define function to extract transformed estimate from sample
#Posson:
fun <- function(){
  exp(Predictor)
}
#binomial
fun <- function(){
    exp(Predictor)/(1+exp(Predictor))
  }
#get matrix of posterior draws
est.matrix<-inla.posterior.sample.eval(fun, sample)
```

This document can be made available in alternative formats on request for a person with disability.

health.wa.gov.au