Autism and ADHD prevalence in Chile through Bayesian prevalence analysis, machine learning and clinical record data linkage

Student 5526, Newnham College

2023-07-20

Declaration: 'This dissertation is submitted for the degree of Master of Philosophy.'

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The dissertation does not exceed the word limit for the respective Degree Committee.

Word count: 14,516 words

Acknowledgements

With thanks to my supervisor, Dr Andres Roman-Urrestarazu, Department of Psychiatry, University of Cambridge, for his time, expertise and encouragement throughout this project; to Jonathan Lewis at Cambridgeshire and Peterborough NHS Foundation Trust for assisting with access to their data; to Health Data Research UK for their generous scholarship; to Ant Bagshaw for suggesting a Cambridge adventure; and to Dad for first sparking my interest in health data science so many years ago.

Contents

1 Abstract **4**

2 Introduction **4**
2.1 Chilean health context . . . 4
2.2 Autism . . . 8
2.3 ADHD . . . 9
2.4 Prevalence analysis techniques . . . 10
2.5 Machine learning . . . 10
2.6 Probabilistic record linkage . . . 10

3 Aims **11**

4 Methods **11**
4.1 Deviation from research protocol . . . 11
4.2 Data management . . . 12
4.3 Data collection . . . 12
4.3.1 School data . . . 12
4.3.2 Clinical data . . . 12
4.3.3 Additional data . . . 14
4.4 Aim 1: Use school data and frequentist method to find a lower bound on autism and ADHD prevalence, and assess socio-demographic variation . . . 14
4.4.1 School data preparation . . . 14
4.4.2 Crude prevalence . . . 14
4.4.3 Frequentist age and sex adjustment . . . 15
4.5 Aim 2: Use clinical data and machine learning to identify autism diagnosis characteristics . . . 15
4.5.1 Clinical data preparation . . . 15
4.5.2 Multiple correspondence analysis . . . 16
4.5.3 Alternative machine learning approaches . . . 16
4.6 Aim 3a: Use machine learning to link school and clinical records . . . 16
4.6.1 School data preparation . . . 16
4.6.2 Clinical data preparation . . . 16
4.6.3 Selection of features for matching . . . 17
4.6.4 Manual record linkage . . . 17
4.6.5 Probabilistic record linkage . . . 17
4.6.6 Alternative record linkage methods . . . 17
4.6.7 Comparison of matched and unmatched records . . . 17
4.7 Aim 3b: Accurately estimate autism prevalence and unmet need, project estimates across health services using Bayesian prevalence prediction . . . 18
4.7.1 Updated autism prevalence estimation . . . 18
4.7.2 Prevalence projection . . . 18
4.7.3 Bayesian prevalence analysis . . . 18
4.7.4 Prior selection . . . 19
4.7.5 Markov chain Monte Carlo sampling . . . 20

5 Results **20**
5.1 School data . . . 20
5.2 Frequentist prevalence estimation . . . 21
5.3 Clinical data . . . 28
5.4 Machine learning with clinical data . . . 28
5.5 Linkage of school and patient records . . . 30
5.5.1 Manual record linkage . . . 48
5.5.2 Probabilistic record linkage . . . 48
5.6 Updated autism prevalence estimates and delta . . . 49

5.7 Bayesian prevalence projection 51

6 Discussion **54**
6.1 Findings 54
6.1.1 Aim 1: Use school data and frequentist method to find a lower bound on autism and ADHD prevalence, and assess socio-demographic variation 54
6.1.2 Aim 2: Use clinical data and machine learning to identify autism diagnosis characteristics 56
6.1.3 Aim 3a: Use machine learning to link school and clinical records 56
6.1.4 Aim 3b: Accurately estimate autism prevalence and unmet need, project estimates across health services using Bayesian prevalence prediction 56
6.2 Limitations 57
6.3 Extensions 58

7 Conclusions **59**

8 Supplementary materials **59**

9 References **81**

10 Appendix A | R code **85**

11 Appendix B | Research Protocol **155**

1 Abstract

The prevalence of autism and ADHD in Chile has not been well characterised and therefore the unmet need of affected children without school-based support is unknown. This investigation aims to quantify the prevalence of autism and ADHD in Chilean children and understand the characteristics of children with autism through analysis of 3 million school records and linkage to 1,300 supplementary clinical records. Found using frequentist age- and sex-adjustment, a lower bound on the prevalence of autism is 0.46% with male to female ratio of 6.00:1 and a lower bound on the prevalence of ADHD 1.50% with male to female ratio of 1.94:1. Autism prevalence peaks in 6-8 year-olds and ADHD in 9-11 year-olds. Prevalence of autism and ADHD is lower among Mapuche students than non-Indigenous students and is lower among students at urban schools than rural schools; no patterns are clear for SES. Multiple correspondence analysis of categories in clinical autism diagnosis data for Servicio de Salud Araucanía Sur showed ethnicity, age and commune of residence were important in distinguishing between patients with autism. Probabilistic linkage of school records and clinical autism diagnoses found autism prevalence of 1.22% with male to female ratio of 4.18 for Servicio de Salud Araucanía Sur and identified an unmet need of 1,132 students with autism that were not accessing school-based supports. Projecting findings across Chile, the national prevalence of autism is estimated to be 1.31%, with an unmet need of up to 25,903 students not accessing school support, among which females are disproportionately affected. Bayesian prevalence projection found credible intervals of autism prevalence for each Chilean health service. Collectively, these findings are a first step to creating a rich, national picture of autism and ADHD prevalence in Chile.

2 Introduction

2.1 Chilean health context

Chile is a high income country in Latin America with a population of 19.7 million, GDP of US$15,356 per capita, and US$2,699 per capita annual health care spend (1,2). Chile has a dual health care system with most people accessing care in the public health system that is partly funded by the Government through the Fondo Nacional de Salud (National Health Fund, FONASA), a social insurance fund (3). Approximately 15% of the population, mostly those of middle and high income, pay for private insurance through Instituciones de Salud Previsional (Social Security Institutions, ISAPRE) (3). All workers also pay a mandatory health contribution of 7% of their income which funds both FONASA and ISAPRE services (3). FONASA defines four coverage groups based on income with those in groups A and B having low income and receiving 100% subsidised care, and groups C and D having low to middle income and receiving 90% and 80% subsidised care respectively (3). Care is delivered through 29 Servicios de Salud (health services) that cover catchment areas within Chile's 16 regions, see Table 1 (4). The public and private systems have various additional pay-per-service options which creates high out-of-pocket expenses that effectively redistribute wealth with those in the top income quintile paying on average 7.0% of monthly household expenditure on out-of-pocket healthcare and those in the bottom quintile paying 3.9% (5,6). Many conditions are not covered by ISAPRE services meaning high needs and high risk patients are overrepresented in the public system (3). The public system is stretched thin with wait times of up to 1.6 years and 2 million people on specialist waiting lists in 2022 (3).

Chileans accessing public health care are on average poorer than those accessing private care with 17.2% below the poverty line in 2013, and they are less likely to access specialist medical services (7). Chile has considerable income inequality with a Gini Index of 44.9 in 2020 (8). In 2014, 21% of inequality was due to spatial inequalities, of which 50% was due to differences between communes meaning inequality in Chile has high granularity (9). Income is highest in the regions of Magallanes y Antártica Chilena, Santiago, Antofagasta and Aysén del Gral. Ibañez del Campo, see Figure 1 (10).

Chile has a large Indigenous population with nearly 10% of Chileans self-identifying as Mapuche and approximately 2.4% belonging to other Indigenous groups (11). Sandoval, Portaccio and Albala found life expectancy at birth of Indigenous Chileans as a whole group is seven years shorter than for non-Indigenous Chileans and 8.3 years shorter for Mapuche people as a group, and suggest lack of access to health services as a possible cause (12). Sandoval and Portaccio separately note that accurate analysis of Chile's health context

Table 1: Chilean health services by region

Region	Health services
Antofagasta	Antofagasta
Arica y Parinacota	Arica
Atacama	Atacama
Aysén del Gral. Ibañez del Campo	Aisén
Bío-Bío	Arauco
Bío-Bío	Biobío
Bío-Bío	Concepción
Bío-Bío	Talcahuano
Coquimbo	Coquimbo
La Araucanía	Araucanía Norte
La Araucanía	Araucanía Sur
Libertador Bernardo O'Higgins	Libertador B.O'Higgins
Los Lagos	Chiloé
Los Lagos	Osorno
Los Lagos	Reloncaví
Los Ríos	Valdivia
Magallanes y Antártica Chilena	Magallanes
Maule	Maule
Metropolitana de Santiago	Metropolitano Central
Metropolitana de Santiago	Metropolitano Norte
Metropolitana de Santiago	Metropolitano Occidente
Metropolitana de Santiago	Metropolitano Oriente
Metropolitana de Santiago	Metropolitano Sur
Metropolitana de Santiago	Metropolitano Sur Oriente
Tarapacá	Iquique
Valparaíso	Aconcagua
Valparaíso	Valparaíso San Antonio
Valparaíso	Viña del Mar Quillota
Ñuble	Ñuble

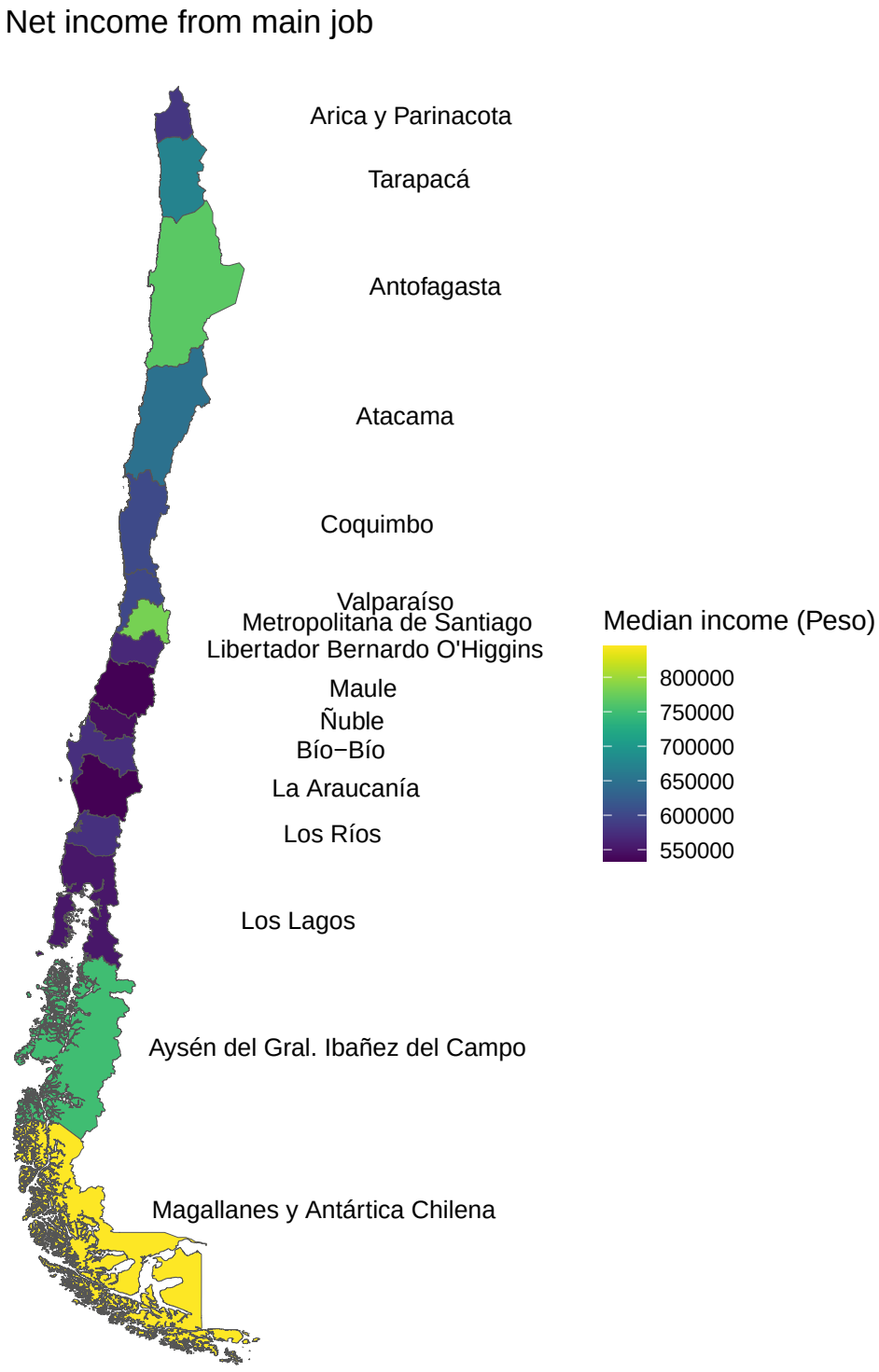


Figure 1: Net income from main job in Chile in 2021 by region, from the INE's Supplementary Income Survey (10).

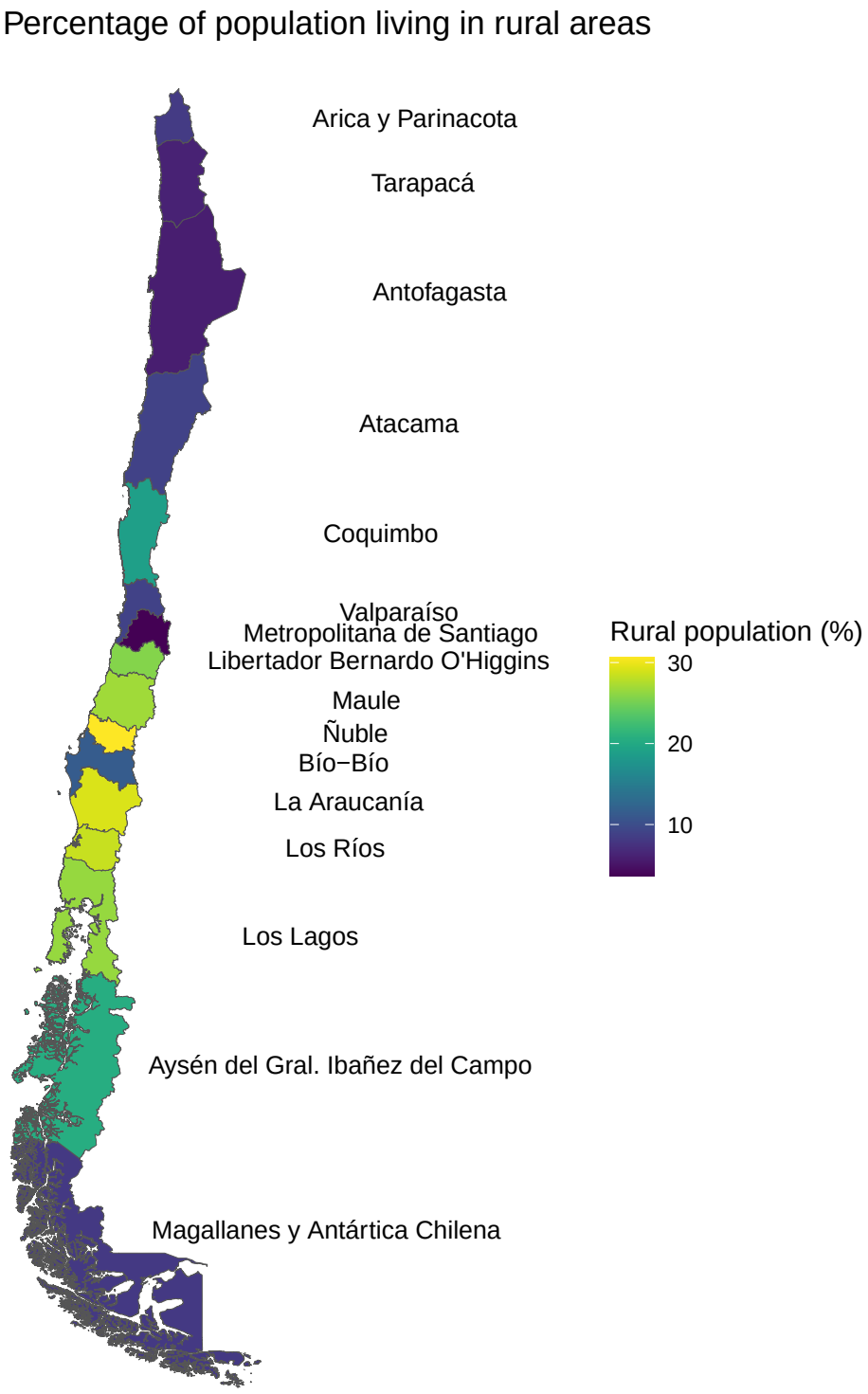


Figure 2: Percentage of population living in rural areas in Chile in 2017 from 2017 census (13).

for Indigenous groups is difficult because ethnicity is not required on death certificates (11). Large population health studies with self-reported ethnicity data are therefore particularly valuable in Chile.

Approximately 12% of Chile's population lives in rural areas; the regions of Ñuble, La Araucanía and Los Ríos have particularly high prevalence of rural dwelling, see Figure 2, and approximately 7 million people live in Santiago, Chile's largest city (13). Núñez and Manzano identified a lack of availability of rural health clinics as a barrier to accessing healthcare in Chile (6).

Chilean policy encourages children with special education needs and disabilities, including autism and ADHD, to be educated in mainstream schools and the Government provides funding to enable this (14). For each child with a registered disability that is integrated into a regular public school, the Ministerio de Educación (Ministry of Education) provides funding through the Subvención de Educación Especial Diferencial (Differential Special Education Grant, SEED) (15). SEED is provided for students with severe physical or mental disabilities, including autism and ADHD, that require education adjustments such as specialist classes or small class sizes (16). Specialists in the SEED network of providers conduct external assessments of students that apply for SEED funding to determine whether they have a special educational need. Chile's schooling system comprises 44% state-funded schools, 51% state-subsidised private schools and 5% non-subsidised private schools (14). Students in non-subsidised school are not eligible for SEED (15). Approximately 5% of Chilean students are registered as having special needs and there is evidence that they have not been completely successfully integrated into public schools as they experience higher rates of bullying and violence, and teachers report not feeling equipped to meet their needs (14). This may prompt some students with less severe needs to choose not to access support as it would single them out as different from their peers. The Programas de Integración Escolar (School Integration Programme, PIE) was established in 2013 to improve integration of students with special needs into public schooling (17).

2.2 Autism

Autism, or Autism Spectrum Disorder (ASD) describes a range of developmental conditions that cause significant challenges to functioning and include persistent social difficulties and restrictive or repetitive behaviours (18). The ICD-10 codes F84.0 and F84.1 describe autism and the code F84.2, F84.3, F84.4, F84.5, F84.8 and F84.9 describe other pervasive developmental disorders closely related to autism (19). Autism is diagnosed through specialist or caregiver assessment of behaviour, cognitive function and developmental milestone attainment and generally uses standardised screening tools (20). The Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS), the gold standard autism diagnostic tools, have been translated into Spanish, and the Modified Checklist for Autism in Toddlers (M-CHAT), Autism Screening Questionnaire (ASQ/SCQ), Autism Behaviour Checklist (ABC), Social Responsiveness Scale (SRS), Autism Detection in Early Childhood (ADEC), the Childhood Autism Rating Scale (CARS), the Structured Observation for Autism Screening (OERA), and the Autism Mental Status Examination (AMSE) have been validated in Chile (21). Roman-Urrestarazu showed that a 10 question version of the Quantitative Checklist for Autism in Toddlers (Q-CHAT-25) is valid for use in routine health checks in primary health clinics in Chile (22).

The worldwide population prevalence of autism is estimated to be 1-2% (20), and if this prevalence holds true for Chile it represents a burden of disease of approximately 200,000 – 400,000 people. Given autism is typically diagnosed in early childhood (23), much of this burden falls on schools. Very little research has been conducted on the prevalence of autism in Chile. Yáñez et al found autism prevalence of 1.95% (95% CI 0.81-4.63%) in their small study of 272 children in Estación Central commune of Servicio de Salud Metropolitano Central and Santiago commune of Servicio de Salud Metropolitano Occidente, both in Santiago (24). They used consecutive sampling of children attending optional checkups at clinicals that primarily treat low to middle income families which is not fully representative of Chile's broader population, therefore their prevalence estimate is likely to contain selection bias. Their figure may be an underestimate as participants were aged 18-30 months which is well below the UK's median diagnosis age of 55 months (23) and the CDC's recommended peak prevalence assessment age of 8 (25). Thus Yáñez et al. may have missed children with less severe autism symptoms and hence underestimated prevalence.

Some data on autism prevalence is available for other Latin American countries. The estimated prevalence

of autism was 0.17% (95% CI 0.015 - 0.019%) among 254,905 children aged 3-9 in Venezuela in 2008 (26). Prevalence of autism in Buenos Aires between 2010 and 2017 was estimated as 1.03% from 22,750 clinical records (27). In Ecuador in 2015, prevalence of autism diagnosis among 51,000 students aged 5-15 was 0.11% and prevalence of suspected autism was 0.21%, with these low estimates thought to be due to children with disabilities having low school attendance rates (28). The older prevalence calculations are likely to be underestimates as the prevalence of known autism diagnoses has been increasing for the last 20 years, though this is thought to be due to more appropriate diagnostic approaches and the true prevalence has remained similar (29). To the best of our knowledge, this investigation is the largest study of autism prevalence in Chile and Latin America to date. It is important to find the prevalence of autism in Chile and quantify the population that does not have full school support to ensure people can access the interventions they need to live healthy and fulfilling lives.

It is also valuable to understand the distribution of autism cases in Chile and the characteristics of people that do have diagnoses to identify population groups that are under-diagnosed or under supported. The autism prevalence is known to differ by age, sex, socio-economic status, ethnicity and geography (20,30). Individuals with more severe symptoms are more likely to be diagnosed earlier (23) and autism prevalence increases as children age, reflecting the condition's long diagnostic window (30).

Loomes, Hull and Mandy's meta-analysis of male-to-female odds ratios of having autism showed three males are diagnosed for every female in childhood and adolescence (31). This difference is thought to be due to a combination of sex-linked causes, differences in symptom presentation, and bias of diagnostic tools toward traits that are more commonly observed in males and may be masked in females (20,31). Giarelli et al. found that females are more likely than males to meet autism diagnostic criteria but not have a documented diagnosis and that there is no significant difference in mean diagnosis age between female and male children that already had a documented autism diagnosis (32).

Research on the relationship between autism prevalence and socio-economic status is conflicting. Two US studies found high socio-economic status was associated with up to 2.2 times higher prevalence, thought to be due to easier access to diagnostics (33,34), while a French study found the highest prevalence was in the most deprived areas (35) and Roman-Urrestarazu et al's study of UK school students found higher prevalence among those who had ever been eligible for free school meals, an indicator of low SES (30).

Roman-Urrestarazu et al. also found differences in autism prevalence across ethnicities in UK school students; after adjusting for sex, age band and location and in comparison to students from White backgrounds, students from Roma/Irish Traveller, Asian and Other backgrounds were significantly less likely to have autism, and students from Chinese, Black and un-classified backgrounds were more likely to have autism (30). These differences likely reflect a combination of true prevalence differences, differing access to diagnosis and cultural factors that impact awareness of autism (30). Bailey and Arcaiuli's literature review found no significant difference in the actual prevalence of autism between Australian Indigenous and non-Indigenous people but did find Indigenous Australians were less likely to have a diagnosis of autism, and suggest lack of access to diagnosis services, mis-diagnosis and different cultural understandings of disability as probable causes (36).

A systematic review by Zeidan et al. found no difference in autism prevalence between rural and urban areas when using pooled estimates but note that one component study in India found higher prevalence in rural populations and another in Taiwan found higher prevalence in urban populations (37).

2.3 ADHD

ADHD is a neurodevelopmental condition and is characterised by severe inattention and hyperactivity (38). It is described by the ICD-10 codes F90.0, F90.1, F90.8 and F90.9 (19). Like autism, ADHD is typically diagnosed in childhood, using similar diagnostic tools (38). Fioravante, Lozano-Lozano Martella noted in their 2022 article that Chile does not have a standardised, objective ADHD diagnosis process and recommend combining diagnostic instruments such as the Conners Comprehensive Behaviour Scale, Conners Parent Scale, the Strengths and Difficulties Questionnaire (SDQ-Cas) and the Diagnostic and Statistical Manual Version 5 (DVM-V) diagnostic criteria (39). ADHD and autism often occur together and Salazar et al. estimate that 59.1% of people with autism have a co-diagnosis of ADHD (40).

The prevalence of ADHD in Chile was estimated to be 10.0% by de la Barra et al. using data on 1,570 school students in Cautin (La Araucanía region), Santiago, Iquique (Tarapacá region) and Concepción (Bío-Bío region) between 2007 and 2009 (41). They found no significant differences by sex or age and found prevalence was higher in children aged 4-11 than in adolescents (41). They also note that their findings differed from the established literature that shows higher prevalence in males than females and among people with low socio-economic status (41). Assuming applicability across all ages and regions, de la Barra et al.'s estimate represents a burden of disease of nearly 2 million people in Chile. In an earlier publication on what is presumably a subset of the same data, Vicente et al. reported ADHD prevalence of 12.6% in 792 students in Santiago (42).

ADHD prevalence varies by other socio-demographic features. It is now considered to be more common in people with low socio-economic status but is less well diagnosed in this group (43). The prevalence of ADHD among Indigenous people is typically found to be lower than in non-Indigenous people; Borges e Azevêdo et al. found it to be 4.3% among Karajá children in Brazil (44). A study in Vietnam found ADHD prevalence of 7.7% overall, comprising 14.9% in urban areas and just 6.7% in rural areas (45). Finding up-to-date estimates of ADHD prevalence stratified by socio-demographic factors is important to understand differences in access to school supports and thus address inequities.

2.4 Prevalence analysis techniques

Analysing the population prevalence of a condition can be considered either a population-level task in which the population is assumed to have a particular distribution and inference is made on its mean, or an individual-level task in which the condition is observed in individuals and this informs calculation of the population prevalence, as proposed by Ince et al. (46). They illustrate a hierarchical binomial model of population prevalence, from which frequentist or Bayesian estimates can be made and their Bayesian estimate of population prevalence uses a conjugate beta prior to find the exact formula for the posterior predictive distribution (46). Downing et al. demonstrate a powerful application of Bayesian prevalence analysis to incomplete, multi-source data in their opioid-dependence study (47). Their study combined opioid treatment data with opioid-related mortality, hospitalisation and arrest datasets and fitted a multi-source Bayesian regression model to extrapolate the prevalence observed in the treatment data to the broader population (47). Bayesian methods are particularly useful when prior information about the condition's prevalence exists, when data is incomplete or has patchy coverage, when some indication of the likelihood of output values is needed, when sample sizes are small, or when frequentist approaches are insufficient to model the complexity of the system under study (48).

2.5 Machine learning

Machine learning provides an increasingly valuable suite of tools in population health research and clustering techniques are well suited to understanding the complexity inherent in epidemiological datasets (49). Multiple correspondence analysis (MCA) is a descriptive, unsupervised machine learning technique that is well suited to categorical data (50). It uses Euclidean distances to cluster data points based on similarity of categorical feature values, without making assumptions about features' distributions (50). MCA is most useful in exploratory analysis of large datasets to find feature categories that group together and associations are often assessed visually (49,50). Costa et al. used MCA to explore the relationship between healthy aging and categorical variables describing 1,051 patients' cognition, health, lifestyle and demographic features (50).

2.6 Probabilistic record linkage

Probabilistic record linkage is a technique for joining two datasets that do not share a common feature that uniquely identifies individuals in each dataset (51). It was proposed by Fellegi and Sunter in 1969 (52) and refined by many since then, such as Enamorado, Fifield and Imai who parallelised the algorithm to improve efficiency, added category frequency information to calculations of pair weights and accounted for uncertainty in the merging processes (51).

Probabilistic record linkage generates all possible record pairs across both datasets, compares the similarities

of each pair's values for each common feature and calculates their weight, typically using an expectation maximisation (EM) algorithm that includes consideration of the relative contributions of each feature and their categories' relative frequencies (52–54). Pairs are then allocated as a link, a non-link or a possible link based on user-defined thresholds for the pair weights (52). The similarity comparison is also user defined with Jaro-Winkler distance being common for character features such as names and exact matches being common for numeric features (51,53). R's RecordLinkage package uses Contiero et al.'s EM algorithm, EpiLink, to calculate pair weights (53,55). Blocking can be used to improve efficiency by requiring a perfect match on some pre-specified features and therefore reducing the total number of pairs to be considered (53).

Enamorado, Fifield and Imai recommend using probabilistic rather than deterministic models to merge large datasets because probabilistic models allow the certainty of matches to be quantified and deterministic models do not, meaning the latter are prone to bias due to measurement error (51). They also recommend blocking the data before running the matching algorithm to reduce the number of false matches identified, especially when the overlap between datasets is low (51).

A simulation study by Hejblum et al. validated the use of probabilistic linkage for matching anonymised clinical records using data on patients' diagnoses (56). They found the Fellegi-Sunter method underperformed compared to their proposed Bayesian method because it is not well suited to one-to-one matching and does not distinguish between the feature values of matches – it weights 0-0 matches the same as 1-1 matches - which in a clinical data context where missingness is common would cause overweighting of low information matches (56).

3 Aims

This project aims to better understand the prevalence of autism and ADHD in children in Chile. To that end, its objectives are to:

1. Find a lower bound on Chile's prevalence of autism and ADHD and assess variation across sex, health service, SES, ethnicity and rurality from school data using a frequentist method.
2. Identify common characteristics of Chilean children with autism from clinical data using machine learning clustering.
3. Link Chilean school data and clinical data to obtain accurate autism prevalence and unmet need estimates in one health service, and project estimates across all health services using Bayesian prevalence prediction.

4 Methods

4.1 Deviation from research protocol

This investigation differed substantially from the research protocol provided at Appendix B. The protocol intended to investigate autism prevalence in the Cambridgeshire region using clinical data from the Cambridgeshire and Peterborough NHS Foundation Trust and school data from the UK Department for Education's National Pupil Database. Unfortunately this clinical data was found to be of insufficient size and quality to conduct the proposed investigation and while this school data was of high quality, it had previously been well analysed by Roman-Urrestarazu (30,57). Therefore school and clinical data from Chile were used instead as they were of high quality and autism and ADHD prevalence is under-researched in Latin America. The available data from Chile did not have fields on age at autism diagnosis and it was therefore not possible to progress the research protocol's first aim of using machine learning to analyse autism diagnosis age. Instead, machine learning has been used to analyse characteristics of patients with autism. The protocol's second aim of linking school autism diagnoses with clinical diagnoses was adapted and appears here as supplementation of the Chilean school data diagnoses with clinical diagnoses. Additional aims and analysis that were appropriate to the Chile data were developed and are detailed in Section 3 above and the rest of this section.

4.2 Data management

The statistical software package R was used to clean, link, analyse and visualise data. R analysis scripts and version control were managed in GitHub and are available at https://github.com/delatee/Autism-diagnosis-age-ML. Raw data was not uploaded to GitHub. All data was stored on local devices and will be deleted at conclusion of the project.

4.3 Data collection

See Figure 4 for an illustration of the data sources and how they were combined through the analysis detailed below.

4.3.1 School data

This research used data from the Chilean Government's Ministerio de Educación that was provided under a freedom of information request and Chile's Law 21545 (58). This dataset contained anonymised records of 3.6 million students in all Chilean schools in 2021. It was collected and curated by the Chilean Unidad de Estadísticas (Statistics Unit), Centro de Estudios (Study Centre) and Ministerio de Educación and is housed in the Sistema de Información General de Estudiantes (General Student Information System, SIGE). It included data on school type and location, student characteristics, academic performance, special needs and monthly school fee contributions. This dataset will be referred to as the school data.

The school data showed whether students access SEED funding, and if they did it provided the disability under which they qualified for funding. The school data included four groups of students who had autism but were recorded in this dataset as not having autism: students with autism who attended private schools which are not eligible for SEED; students with autism who chose not to apply for SEED; students with autism who applied for SEED but were found to not be sufficiently severely affected to receive SEED; and students with autism who received SEED but for a different disability as only one can be recorded, perhaps one that more severely affects them. The number of students in each of these groups was unknown and these groups were analogous for students with ADHD. Thus all estimates of autism and ADHD prevalence from this dataset alone were likely to be underestimates of the true population prevalence.

4.3.2 Clinical data

This research also used data from Chile's Servicio de Salud Araucanía Sur (South Araucanía health service, SSAS), obtained under a freedom of information request. These data were collated from secondary care clinical records, particularly from mental health community care services. They comprised clinical records for public sector specialist health visits of patients aged 6-18 with a primary diagnosis of autism between February 2014 and December 2021 for all communes in the SSAS catchment. Figure 3 shows the communes in the SSAS and Servicio de Salud Araucanía Norte (North Araucanía health service, SSAN) catchments. The clinical records included wage deducted health insurance contributions from Chile's Fondo Nacional de Salud (National Health Fund, FONASA), from which a colleague previously inferred socio-economic status of patients' families. The majority of records were for patients resident in the SSAS catchment area and they did not include any records for privately provided healthcare. These data will be referred to as the clinical data.

For a subset of the clinical data, those for appointments provided by Villarrica Hospital in the Villarrica commune of Chile's La Araucanía region, patients' autism diagnoses were manually validated prior to this investigation. Validation involved a neurologist and a psychiatrist in the SEED network of providers checking the clinical records against central health records to confirm individuals did have autism, adding secondary and tertiary diagnoses where relevant, and adding demographic information including ethnicity, rurality of residence, existence of another disability and experience of foster care status where available. Most patients in this subset were resident in Loncoche, Pucón and Villarrica communes in the SSAS catchment area. These data will be referred to as the validated clinical data, see Figure 4 and the rest of the clinical data that was not validated will be referred to as the remaining clinical data.

Figure 3: Communes in the La Araucanía region, with Servicio de Salud Araucanía Sur (SSAS) communes in green and Servicio de Salud Araucanía Norte (SSAN) communes in blue (60).

4.3.3 Additional data

Chile's 2017 census data, held by the Instituto Nacional de Estadisticas (National Statistics Institute, INE), was used to create a standard population of Chile's age and sex distribution from projections for 2021 of population size by age and sex (59). It was also used to obtain projections of the population of youth aged 0-14 in 2021 by region and the percentages of the population living in rural areas in 2017 by region (13).

Data from the INE's Encuesta Suplementaria de Ingresos (Supplementary Income Survey, ESI) of 2021 (10) was used to obtain the nominal median income by region in 2021.

For mapping, shapefiles of administrative areas were obtained from The Humanitarian Data Exchange of the United Nations Office for the Coordination of Humanitarian Affairs (60). Additional region and commune naming information was taken from the R package 'chilemapas'.

4.4 Aim 1: Use school data and frequentist method to find a lower bound on autism and ADHD prevalence, and assess socio-demographic variation

Using the school data, the crude prevalence of autism and ADHD were found. Frequentists methods were used to find the age- and sex-adjusted autism and ADHD prevalence among Chilean school students. Variation in prevalence across age, sex, commune, SES, ethnicity and rurality was assessed by comparing confidence intervals.

4.4.1 School data preparation

The school dataset was restricted to students aged 6-18 as of 30th June 2021 to capture children of school age in Chile. No restriction on sex was necessary as it contained only females and males. Students' commune of residence was mapped to their respective health service catchment areas. Two issues due to boundary changes were corrected. Firstly, the commune now called Tocopilla which falls across the Antofagasta and Tarapacá regions was formerly two communes, Tocopilla in Antofagasta and Pozo Almonte in Tarapacá. The old names have been retained to ensure appropriate region mapping and they were mapped to Antofagasta and Iquique health services respectively. Secondly, the communes recorded as belonging to the former Ñuble sub-region of the Bío-Bío region were mapped to their corresponding communes in the recently formed Ñuble region.

Commune of residence was missing for 4078 students (0.13%). The commune of their school was imputed as most students were likely to go to school near their place of residence.

Students' age as of 30th June 2021 was mapped to age-bands of 6-8, 9-11, 12-14 and 15-18, preserving the divide between primary and secondary school as secondary school starts at age 12 in Chile, and with a larger final band as fewer students were expected to be diagnosed older and the 18 years old group may be small due to students leaving schooling. Students' ethnicity was mapped to being a member of the Mapuche Indigenous group, being a member of another Chilean Indigenous group, or not being a member of an Indigenous group based on their recorded ethnicity which can take at most one value. The 6859 students (0.22%) with ethnicity recorded as 'no registry' were mapped to not being a member of an Indigenous group.

Students' school fee status was mapped to a proxy socio-economic status feature. Students with free schooling were given SES of 1 indicating low status as families with low SES are entitled to schooling rebates. Students paying Chilean Peso $1,000-$100,000 monthly were given SES of 2, indicating medium status, and students paying more than $100,000 monthly were given 3 to indicate high status. For school fee status, 49973 students (1.64%) were coded as 'No information' and 29 students (0.00095%) were missing school fee values and these were re-categorised as 'No information' to reflect their unknown school fee status. No other features of interest for this analysis had missingness.

4.4.2 Crude prevalence

The crude prevalence of autism and of ADHD were calculated as the count of cases divided by the relevant population size. Grouping features of sex, health service of residential commune, monthly school fee as a

proxy for SES, ethnicity and school's rurality were used, and calculations were made by sex and age band as calculating by single year ages resulted in too many counts of less than 20.

4.4.3 Frequentist age and sex adjustment

Nationally and for each grouping feature, prevalence was standardised by age and sex using direct standardisation following the method presented by Fay and Feuer (61). The 2017 Chile census projections of the distribution of people by age and sex in 2021 was used as the standard population. Gamma confidence limits were calculated at the 95% level using chi-squared distributions. The adjusted prevalence rates were multiplied by their respective population sizes to give adjusted prevalence counts which were rounded to integers, therefore losing a small amount of precision.

Ethnicity analysis compared individuals from the Mapuche Indigenous group to those from other Indigenous groups or with no Indigenous group and used data from the La Araucanía, Aysén, Biobío, Los Lagos, Los Rios, Magallanes and Región Metropolitana de Santiago regions only as these regions were collectively the five regions with the largest Mapuche populations and the five regions with the highest proportion of Mapuche people in their population.

4.5 Aim 2: Use clinical data and machine learning to identify autism diagnosis characteristics

The machine learning technique of multiple correspondence analysis for categorical data was used to explore the feature categories in the validated clinical data that explain its variance. The features examined were sex, age band, ethnicity, commune of residence, rurality of residence, health insurance contribution level, other disability status and foster care status.

4.5.1 Clinical data preparation

The full clinical dataset was cleaned to ensure internal consistency in feature values and translated into English where appropriate. Patients were categorised as having autism if any of their diagnosis codes were in ICD F84-F89 for any appointment. No restriction on sex was necessary as these data contained only females and males. The clinical data included economic data in the form of patients' family's health insurance contribution level. Values taken were 'FONASA-A', 'FONASA-B', 'FONSAS-C' and 'FONASA-D' which are respectively larger contributions to this public health fund and were assumed to map to increasing socio-economic status, and 'Private health insurance' which indicates contributions to a private health insurance provider rather than the public fund and was therefore assumed to map to the highest socio-economic level.

For multiple correspondence analysis, only the validated clinical dataset was used as it contained additional demographic fields not present in the remaining clinical data. In the validated clinical data, patients' age as of 30th June 2021 was mapped to age-bands of 0-2, 3-5, 6-8, 9-11, 12-14 and 15-18. Students below school age were retained for MCA to more fully explore the variation in the autism patient data. The validated clinical dataset had many values for ethnicity, other disability status and foster care status that were recorded as 'no information'. For other disability status and foster care status, additional features were created with the 'no information' values imputed as 'no disability' and 'no foster care' respectively, as it is likely that patients who did have another disability or had experienced foster care would have this recorded.

The validated clinical data was aggregated to patient level. First, only appointments for patients resident in a commune in the SSAS catchment area were included. Then appointments were aggregated to patient level by selecting each patient's most recent commune of residence and rurality status, and most common health insurance contribution level, hospital and medical specialty of appointment. For ethnicity, Mapuche was selected if present in any of that patient's appointment records, then Chilean if present, then Foreign if present and else 'No information' was selected. Similarly, other disability and foster care status were selected as 'Yes' if recorded for any appointment, then 'No' was selected if present, then 'No information' if no status was recorded for any appointment.

4.5.2 Multiple correspondence analysis

The machine learning technique of multiple correspondence analysis (MCA) was conducted using R's FactoMineR package with the number of output dimensions equal to the number of input features. MCA was an appropriate technique to use here because the small clinical dataset contains primarily categorical features which MCA is designed to handle.

Data-led analysis was used to identify features in the small clinical dataset that explained the variance in this patient-level data. Initially, all available features with suspected association with autism were included in MCA, specifically: sex, age band, commune of residence, health insurance contribution level as a proxy for SES, ethnicity, rurality of residence, other disability status and foster care status. Hospital and medical specialty of appointment were not included as they were not thought likely to be associated with autism diagnosis. The data grouped well into having information about other disabilities and about foster care and not having information about each, therefore the MCA was rerun with the disability and foster care features exchanged for their respective imputed versions. This identified age, commune and ethnicity as important for explaining the variance, so the MCA was rerun using only these three features.

4.5.3 Alternative machine learning approaches

Other machine learning techniques including component factor analysis (CFA) and principle component analysis (PCA) were also considered to cluster patients with autism. Both are powerful methods for uncovering the latent structure of data, however CFA typically requires ordered categorical variables and PCA requires continuous variables. As the commune feature, an unordered categorical variable, was thought to be important to explain variation between patients with autism and therefore needed to be included in analysis, using CFA or PCA would require one-hot-encoding of commune which would reduce the appropriateness of these tests.

4.6 Aim 3a: Use machine learning to link school and clinical records

The school data and clinical data were linked using manual and probabilistic record linkage.

4.6.1 School data preparation

For data linkage, the school data was restricted to students with autism that were living in communes in the SSAS catchment in 2021 to maximise comparability with the clinical data. A false empty record was added to the school dataset before linkage that allowed the algorithm to correctly match on SES. This false record was only used during linkage, did not match to any patient records and was removed before comparing matched and unmatched records. This will be referred to as the SSAS school data, see Figure 4.

4.6.2 Clinical data preparation

For data linkage, the full clinical dataset was used. It was restricted to appointments for individuals resident in communes in the SSAS catchment as the data is believed to be complete for this catchment area only. It was also restricted to patients aged 6-18 as of 30th June 2021 to maximise compatibility to the school data. Appointment year was not restricted in order to retain more data and thus maximise linkage opportunities, and only patients of female and male sex were present. Patients with familial socio-economic level of FONASA-A were interpreted to have low SES and given proxy status of 1 (equivalent to students with free school fees); patients with FONASA-B, FONASA-C and FONASA-D were interpreted to have moderate SES and given 2 (equivalent to students paying $1,000-$100,000 monthly for state schooling); and patients with private health insurance were interpreted to have high SES and given 3 (equivalent to students paying more than $100,000 monthly).

The clinical dataset had no missingness in the features of interest for data linkage.

The clinical data was aggregated to one row per patient per commune in the SSAS catchment, to maximise the opportunity for matching patients that had moved within this health service. Most patients lived in only one commune in SSAS during the study period. Aggregation used the most common SES value for each patient. The resulting dataset will be referred to as the patient data, see Figure 4.

4.6.3 Selection of features for matching

The features available for matching that occurred in both the SSAS school data and the patient data were sex, date of birth, commune of residence, and the proxies for socio-economic status – monthly school fees in the SSAS school data and mode health insurance contributions in the patient data.

4.6.4 Manual record linkage

The SSAS school and patient data were blocked on sex and date of birth to improve match quality and reduce runtime. As both datasets were from trusted, large-scale data collections, it was reasonable to assume sex and dates of birth were highly accurate in both datasets, and it would not have been reasonable to accept any proposed matches that did not agree on either sex or date of birth.

Two versions of manual record linkage were tried. First, the SSAS school and patient data were merged with perfect matches required on sex, date of birth, commune of residence and proxies for SES. Second, the SSAS school and patient data were merged with perfect matches required only on sex, date of birth and commune of residence, as the proxies for SES were known to be approximate and therefore requiring perfect matches on SES was not reasonable.

4.6.5 Probabilistic record linkage

The machine learning technique of probabilistic matching was used to link the SSAS school and patient data. All possible pairs of blocked matches across these two datasets were generated and agreement weights were calculated for each feature using expectation maximisation, then aggregated into a weight for the pair. Records with missing values were retained as this linkage method is robust to missingness. The similarity comparison method was exact matching for commune of residence, diagnosis with autism and socio-economic status; there was no value in using a string comparison method for commune of residence as all commune names were already standardised and two communes having similarly spelled names did not increase the likelihood of a match between those communes. Linkage was implemented using R's RecordLinkage package. This included consideration of the average frequencies of categories in each feature and estimated errors rates were supplied: the default estimated error rate of 0.01 was supplied for the commune of residence and diagnosis with autism features as they were expected to be fairly accurate features, and an estimated error rate of 0.1 was supplied for the socio-economic status feature to reflect that it was a loosely defined proxy.

No specific weight value was set as a cut-off to define matches. Instead, pairs were selected based on high weight to create a 1-1, bipartite, matching between SSAS school records and patients. These matches were examined to ensure each patient that lived in multiple communes had matched to only one SSAS school record.

4.6.6 Alternative record linkage methods

Record linkage using Bayesian methods in which matched status is modelled, as developed by Sadinle (62) and refined by Stringham (63), was considered for record linkage here. This technique is typically more complex and has longer runtimes than probabilistic matching but can more easily enforce one-to-one matching and is particularly well suited to matching names. As the datasets in this investigation did not include names, there was limited benefit from using Bayesian linkage methods.

Record linkage using machine learning to classify matched and unmatched pairs, as explored by Pita et al, was also considered as it is generally thought to be more accurate than probabilistic matching alone (64). However this was not pursued because the datasets under investigation had very few common features to match on, meaning machine learning algorithms would have few options to trial, and there were no known true matches with which to assess the accuracy of machine learning models.

4.6.7 Comparison of matched and unmatched records

For the SSAS school and patient datasets, each record was classified as either matched or unmatched based on whether it appeared in the bipartite matching. The discrete Kolmogorov-Smirnov test was used to compare

matched and unmatched records within each dataset for each of the features used for matching, excluding date of birth which had too many categories to have meaningful results. Missing values in the socio-economic status feature were omitted before testing as the Kolmogorov-Smirnov test is not robust to missingness.

Permutation tests were then performed for each of the features tested in each dataset by permuting the matched status 2000 times and recomputing the discrete Kolmogorov-Smirnov test for each permutation. The p-values for the Kolmogorov-Smirnov tests on the observed data were then compared to the distributions of p-values for the permuted data to determine the significance of the observed results.

4.7 Aim 3b: Accurately estimate autism prevalence and unmet need, project estimates across health services using Bayesian prevalence prediction

An updated estimate of autism prevalence in SSAS was made and projected across health services. Unmet need was calculated and projected nationally. Bayesian random effect models were used to predict autism prevalence across health services for several priors.

4.7.1 Updated autism prevalence estimation

The patient data was de-duplicated to a single row per patient with the commune of their matched record chosen if they had multiple communes of residence. The unmatched patients were aggregated to counts per age and sex group. The full school dataset was restricted to students resident in SSAS, then was aggregated to counts per age, sex and autism status group. It was assumed that the patients that did not match to the SSAS school data, which only includes students with SEED for autism, do exist in the larger school data of students resident in SSAS but they did not have a diagnosis of autism in the school data because they did not receive SEED or did not receive it for autism. Thus the count of students with autism for each age and sex group in the restricted school data was increased by the number of unmatched patients in that age and sex group, and the number of students without autism was decreased by the same amount. This effectively reallocated the appropriate number of SSAS students recorded as not having autism to having autism and retained the school data sample sizes. This will be referred to as the linked data, see Figure 4.

The crude and age- and sex-adjusted updated prevalence of autism for SSAS was calculated from the updated counts in the linked data, following the same method as in sections 4.4.2 and 4.4.3. The adjusted prevalence delta was calculated as the difference between the adjusted updated prevalence for SSAS (from the linked data) and the adjusted prevalence for SSAS (from the school data only). It represented the prevalence of clinical autism diagnosis without SEED for autism in SSAS. Unmet need for SSAS was calculated as the adjusted prevalence delta multiplied by the number of students in SSAS.

4.7.2 Prevalence projection

The adjusted updated national prevalence of autism was calculated as the adjusted national prevalence (from the school data only) plus the adjusted prevalence delta (found for SSAS). This assumed the delta was nationally applicable.

Prevalence was projected for each of the health services other than SSAS by adding the adjusted prevalence delta to the adjusted prevalence (from the school data only) for each health service. This projected the prevalence of individuals with a clinical autism diagnosis that did not access SEED across the health services. It assumed that all SSAS patients existed in the school data and were recorded in this dataset as resident in SSAS communes. Estimated confidence intervals were calculated for the projections by finding a band around the projection that was the maximal of either the width of the 95% gamma confidence interval for each health service's adjusted prevalence (from the school data only) or the width of the confidence interval for the adjusted prevalence delta.

4.7.3 Bayesian prevalence analysis

Bayesian prevalence analysis of autism was conducted for the school data by health service. Bayesian methods were appropriate here because they allowed calculation of prevalence with incomplete data. Here two different

types of data, school records and patient appointments, were combined and both were known to be incomplete. Bayesian analysis is robust to this messiness and allowed plausible prevalence predictions to be made. It also provided information about the likelihood of these predictions given the observed school data.

To conduct the Bayesian analysis, a random-effects model was constructed with the random effect on health service as follows.

Set

$$y_i = \text{adjusted count of autism cases in health service } i$$

$$n_i = \text{number of students in health service } i$$

$$\theta_i = \text{prevalence of autism in health service } i$$

For each health service i, the model formula was

$$y_i|(n_i, \theta_i) \sim \text{Binomial}(n_i, \theta_i)$$

With

$$\theta_i \sim \text{Beta}(a, b)$$

And posterior distribution

$$\theta_i|(y_i, n_i) \sim \text{Beta}(y_i + a, n_i - y_i + b)$$

Fitting a binomial model required integer valued counts of autism cases. As adjusted case counts were used throughout, the adjusted counts had to be rounded to integer values which caused a small amount of precision to be lost. This likely caused the posterior credible intervals to be slightly wider but was not expected to have a large effect on findings.

4.7.4 Prior selection

Four priors for θ_i were used when fitting the Bayesian prevalence model. First, a conjugate beta prior common to all health services was constructed with the national age- and sex-adjusted prevalence of autism in the school dataset and its standard deviation used respectively as the mean and standard deviation of the prior. This prior was suitable because the national adjusted prevalence in the school data provided a lower bound on the plausible prevalence of autism in Chile.

Second, a conjugate beta prior specific to each health service using the health service specific age- and sex-adjusted autism prevalence in the school data and their standard deviations as the prior means and their standard deviations respectively. This prior was suitable because it extended the previous prior to each of the random effect categories and it reflected the students know to be receiving SEED for autism. On its own this prior was expected to give uninformative posteriors because it was effectively duplicating the information in the sample data. However it was used here as a health service specific lower bound on the plausible prevalence of autism that can then be compared to the projected prevalence values.

Third, a conjugate beta prior specific to each health service with the health service specific projections of adjusted updated prevalence from data linkage and their standard deviations from their maximal 95% confidence intervals as the prior means and prior standard deviations respectively. This prior was suitable as it captured the extra information provided by the linkage and thus represented all individuals with clinical autism diagnoses, not only those that received SEED for autism. Additionally, this prior had narrow standard deviations which modelled a plausible upper bound on the prevalence of autism in each health service.

Fourth, a uniform prior specific to each health service with the adjusted autism prevalence from the school data for each health service as its lower bounds and the adjusted projected prevalence from data linkage for each health service as its upper bounds. This prior was suitable because it captured the information from both the school and clinical datasets, without specifying where within these bounds the true prevalence values were likely to be.

4.7.5 Markov chain Monte Carlo sampling

Bayesian prevalence modelling used the JAGS (Just Another Gibbs Sampler) R package which uses Markov chain Monte Carlo (MCMC) sampling to produce posterior density distributions when given the above priors and adjusted prevalence observations. A burn-in period of 2000 samples was used to ensure models converged, then 2000 iterations were used to model the posterior densities. Visual inspection of trace plots showed no evidence of a lack of convergence, $\hat{r}$ values were less than 1.1, effective sample sizes were reasonably large and median estimates had at least two significant figures unaffected by Monte Carlo standard error.

5 Results

Figure 4 shows the school and clinical data and how they contribute to the three aims of this investigation. The school data is used to address aim 1, finding a lower bound on autism and ADHD prevalence and understanding their variation, see sections 5.1 and 5.2. The validated clinical data for patients resident in SSAS is used to address aim 2, identifying common characteristics of children with autism, see sections 5.3 and 5.4. Linkage of school and clinical data is used to address aim 3, obtaining and projecting accurate autism prevalence estimates, see sections 5.5, 5.6 and 5.7.

5.1 School data

Table 2: Count and percentage of features' values in the school dataset. Metro. is short for Metropolitano and these health services are all in Santiago.

Feature	Available values	Count (%)
Sex	Female	1,487,224 (48.66%)
	Male	1,569,082 (51.34%)
Age band	6-8	748,406 (24.49%)
	9-11	767,350 (25.11%)
	12-14	749,693 (24.53%)
	15-18	790,857 (25.88%)
Health service	Aconcagua	46,840 (1.53%)
	Aisén	19,890 (0.65%)
	Antofagasta	119,378 (3.91%)
	Araucanía Norte	36,651 (1.20%)
	Araucanía Sur	132,242 (4.33%)
	Arauco	31,318 (1.02%)
	Arica	44,609 (1.46%)
	Atacama	58,743 (1.92%)
	Biobío	71,411 (2.34%)
	Chiloé	30,908 (1.01%)
	Concepción	109,502 (3.58%)
	Coquimbo	141,152 (4.62%)
	Iquique	69,935 (2.29%)
	Magallanes	28,031 (0.92%)
	Maule	182,352 (5.97%)
	Metro. Central	122,576 (4.01%)
	Metro. Norte	180,230 (5.90%)
	Metro. Occidente	277,282 (9.07%)
	Metro. Oriente	182,798 (5.98%)
	Metro. Sur	200,984 (6.58%)
	Metro. Sur Oriente	236,817 (7.75%)

Table 2: Count and percentage of features' values in the school dataset. Metro. is short for Metropolitano and these health services are all in Santiago. *(continued)*

Feature	Available values	Count (%)
	O'Higgins	161,335 (5.28%)
	Osorno	40,266 (1.32%)
	Reloncaví	79,767 (2.61%)
	Talcahuano	54,678 (1.79%)
	Valdivia	66,206 (2.17%)
	Valparaíso	78,598 (2.57%)
	Viña del Mar	172,456 (5.64%)
	Ñuble	79,351 (2.60%)
School fee	Free	2,190,359 (71.67%)
	$1,000-$10,000	1,120 (0.04%)
	$10,001-$25,000	36,477 (1.19%)
	$25,001-$50,000	206,952 (6.77%)
	$50,001-$100,000	270,875 (8.86%)
	$100,001+	300,521 (9.83%)
	No information	50,002 (1.64%)
Ethnicity	Mapuche	176,302 (5.77%)
	Other Indigenous group	35,779 (1.17%)
	No Indigenous group	2,844,225 (93.06%)
Rurality	Rural	238,948 (7.82%)
	Urban	2,817,358 (92.18%)
Accesses SEED	Yes SEED	339,968 (11.12%)
	No SEED	2,716,338 (88.88%)
Autism	Yes autism	14,549 (0.48%)
	No autism	3,041,757 (99.52%)
ADHD	Yes ADHD	46,224 (1.51%)
	No ADHD	3,010,082 (98.49%)

The school dataset contained records for 3,056,306 Chilean students aged 6-18 in 2021, see Table 2. Of these, 1,487,224 (48.66%) were female and the rest were male. Students were evenly distributed across age bands and fairly evenly distributed across communes with generally more students in the metropolitan Santiago communes. Most students received free schooling and after this value, the proportion that paid for schooling increased with monthly school fees. 176,302 (5.77%) students were Mapuche and 2,844,225 (93.06%) did not belong to an Indigenous group. 238,948 (7.82%)a ttended a school in a rural area. A special needs code was recorded for 339,968 (11.12%) students, indicating they accessed SEED funding in 2021. Of these students, 14,549 (4.28%) received SEED for autism and 46,224 (13.6%) received SEED for ADHD.

5.2 Frequentist prevalence estimation

The national crude prevalence of autism in the school data was 0.48% (0.47-0.48%). After using frequentist methods to adjust for the age and sex distribution of the 6-18 year-old population of Chile in 2021, the national adjusted prevalence of autism was 0.46% (0.46-0.47%). The national crude prevalence of ADHD was 1.51% (1.50-1.53%) and the national age- and sex-adjusted prevalence was 1.50% (1.48-1.51%). The crude prevalence of autism and ADHD varied with age, as shown in Figures 5 and 6. Autism prevalence was highest in 6-8 year-olds and decreased with age while ADHD prevalence peaked in 9-11 year-olds then decreased. Both conditions showed a small increase in prevalence for age 18. See also Supplementary Tables 17 and 18.

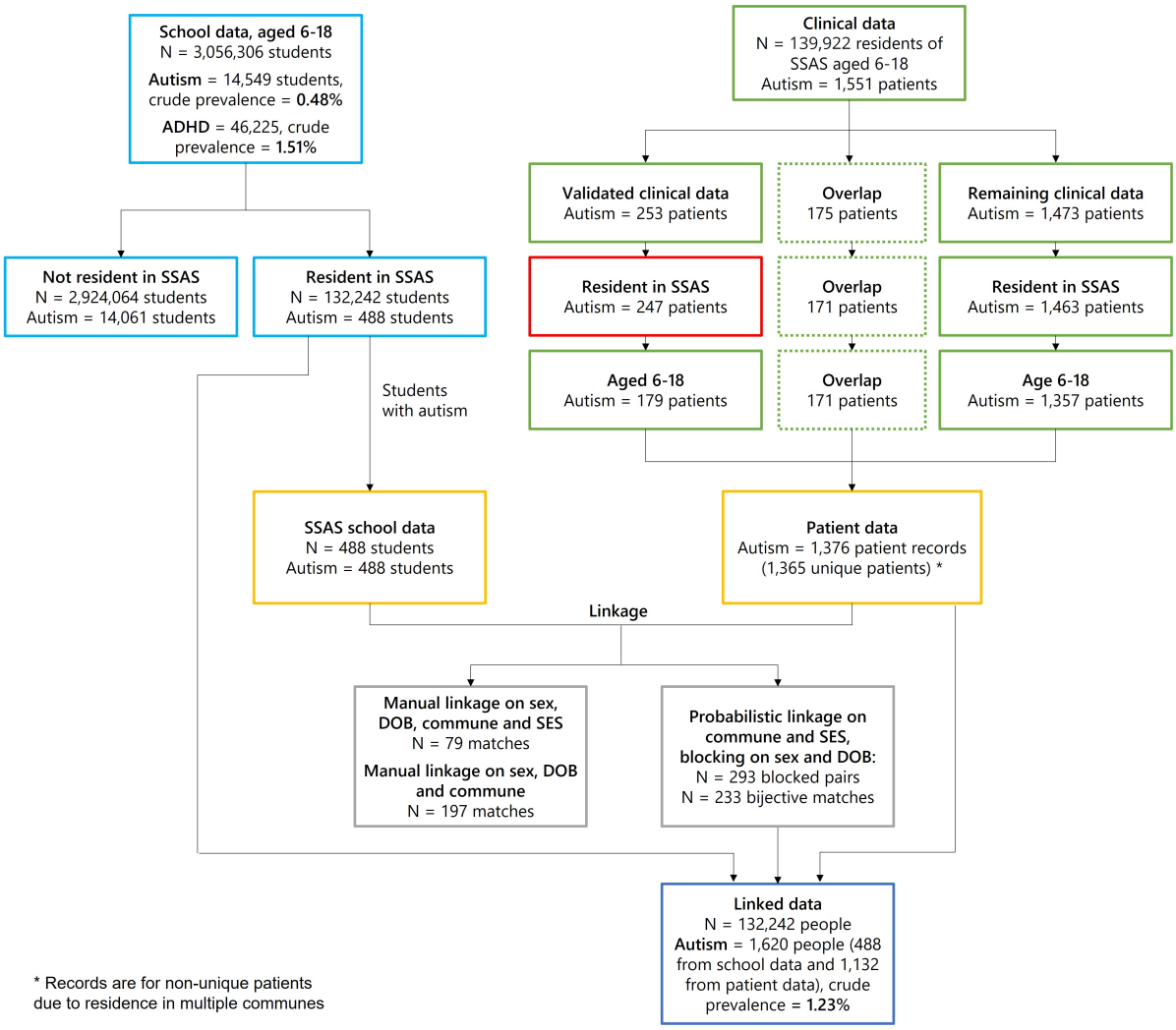


Figure 4: The datasets used in this investigation, their size (N) and the number of people with autism and ADHD and their crude prevalence where applicable. Light blue shows the school data which was used to calculate lower bounds on autism and ADHD prevalence, and subsets of the school data for students resident in Servicio de Salud Araucanía Sur (SSAS) and those that were not. In the school data, autism and ADHD indicate students that accessed SEED for autism and for ADHD respectively. Green shows the clinical data and its subsets of validated and unvalidated clinical data and autism here indicates clinical diagnoses. Validation involved confirming autism diagnosis against central clinical records, adding 2nd and 3rd diagnoses, and adding demographic data. Red shows the validated clinical data for patients resident in SSAS that was used for multiple correspondence analysis. Yellow shows the subsets of school and clinical data used for data linkage; the patient data was the combination of validated and unvalidated clinical datasets and some patients were represented in both these datasets, shown by the dotted overlap boxes. Grey shows the matches resulting from three linkage approaches. Dark blue shows the result of data linkage and was used to calculated updated estimates of autism prevalence.

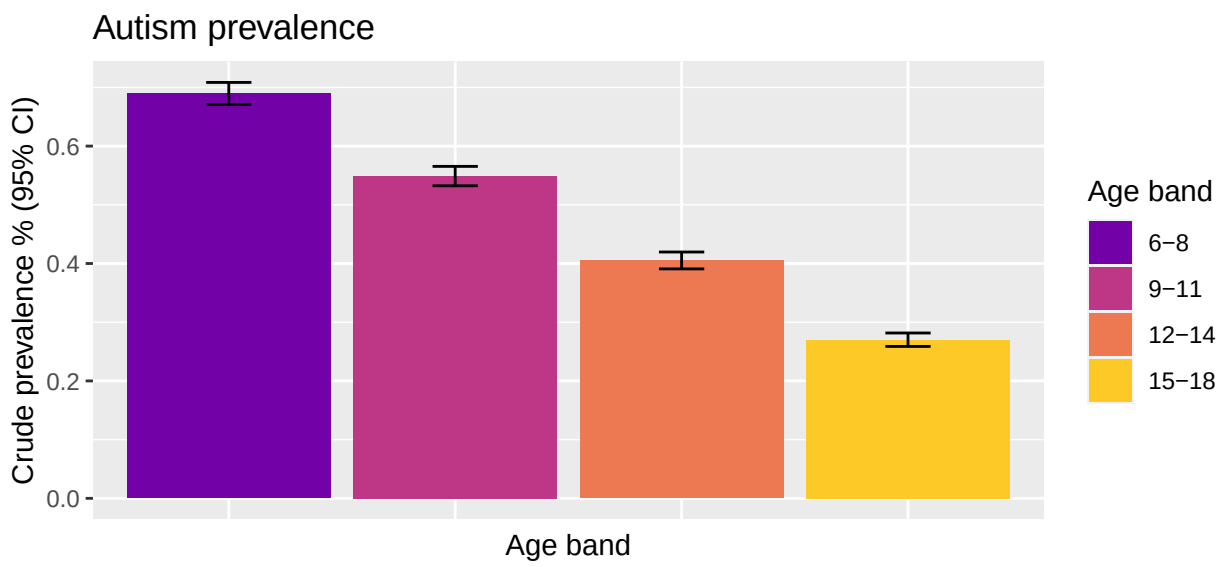


Figure 5: Crude prevalence of autism in school data by age band. Bars show 95% normal confidence intervals.

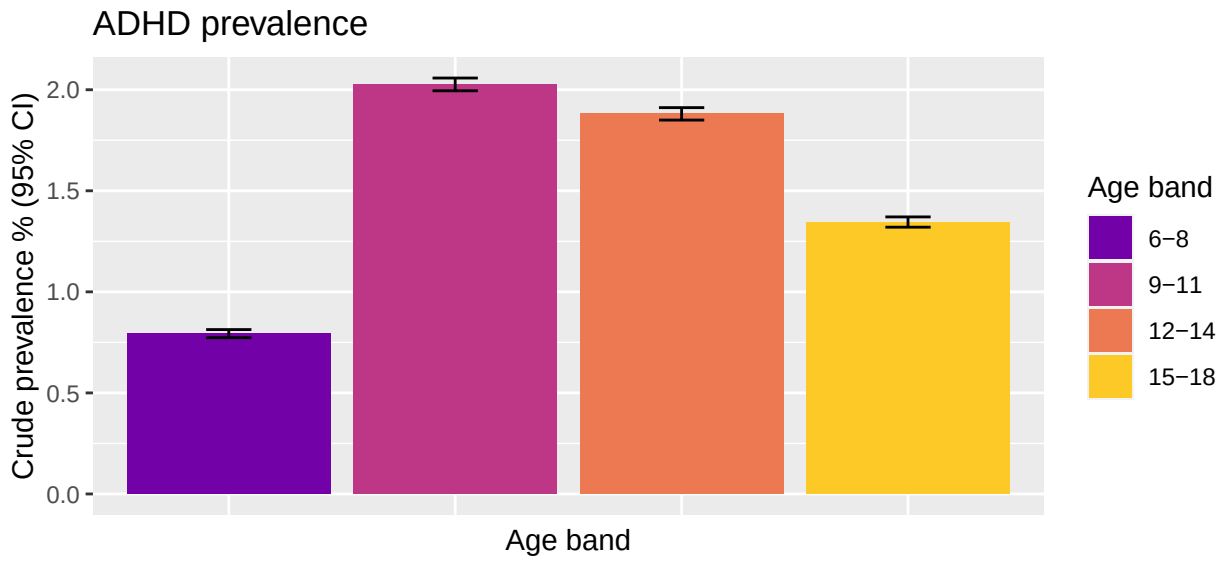


Figure 6: Crude prevalence of ADHD in school data by age band. Bars show 95% normal confidence intervals.

Disggregated by sex, autism prevalence in the school data was 0.13% (0.13-0.14%) for females and 0.80% (0.79-0.82%) for males, with male to female ratio of 6.02:1, see Figure 7 and Supplementary Table 19. Among females, the age- and sex-adjusted autism prevalence was 0.13% (0.13-0.14%) and among males it was 0.79% (0.77-0.80%), thus giving an adjusted male to female ratio of 6.00:1. ADHD prevalence was 1.01% (1.00-1.03%) for females and 1.98% (1.96-2.01%) for males, with male to female ratio of 1.96:1, see Figure 8 and Supplementary Table 20. After adjustment, the prevalence of ADHD was 1.01% (1.00-1.03%) for females and 1.97% (1.94-1.99%) for males with adjusted male to female ratio of 1.94:1.

Given the prevalence of autism and ADHD in these data were likely to be underestimates, the age- and sex-adjusted prevalence values can be considered a lower bound on the true prevalence of autism and of ADHD in Chile.

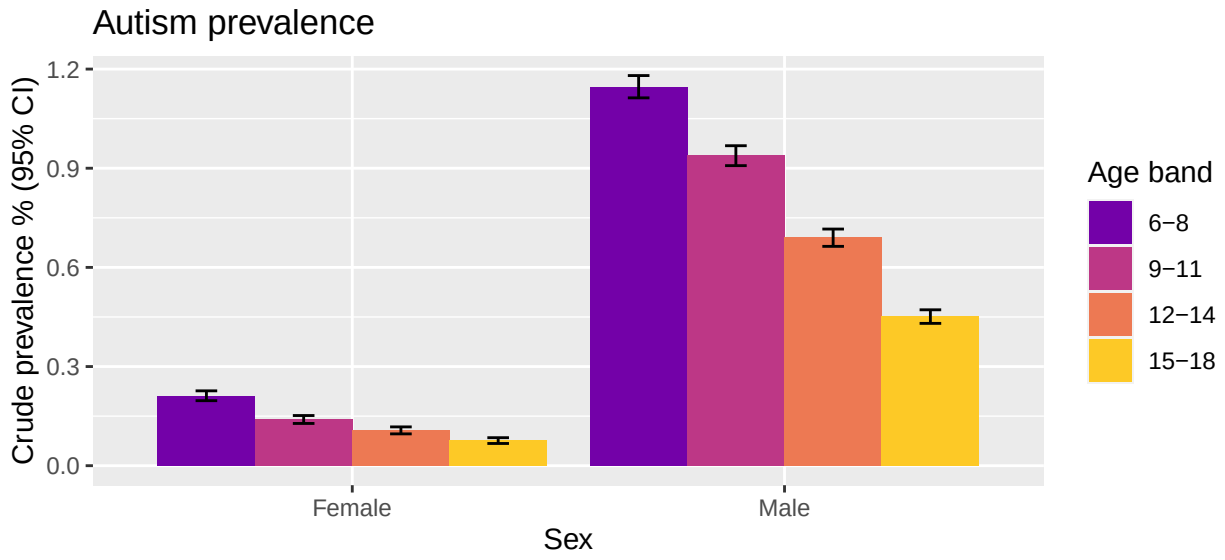


Figure 7: Crude prevalence of autism in school data by age band and sex. Bars show 95% normal confidence intervals.

Table 3: Crude and age- and sex-adjusted autism prevalence by health service in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

Health service	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Aconcagua	0.44 (0.38, 0.50)	0.43 (0.37, 0.50)
Aisén	0.75 (0.63, 0.87)	0.75 (0.63, 0.90)
Antofagasta	0.84 (0.79, 0.89)	0.83 (0.77, 0.88)
Araucanía Norte	0.30 (0.24, 0.36)	0.30 (0.24, 0.38)
Araucanía Sur	0.37 (0.34, 0.40)	0.37 (0.34, 0.41)
Arauco	0.73 (0.64, 0.83)	0.72 (0.62, 0.82)
Arica	0.61 (0.54, 0.68)	0.61 (0.54, 0.70)
Atacama	0.31 (0.26, 0.35)	0.31 (0.27, 0.37)
Biobío	0.43 (0.38, 0.48)	0.42 (0.37, 0.47)
Chiloé	0.45 (0.38, 0.52)	0.43 (0.36, 0.52)
Concepción	0.78 (0.73, 0.84)	0.77 (0.72, 0.83)
Coquimbo	0.41 (0.38, 0.45)	0.40 (0.36, 0.43)
Iquique	0.45 (0.40, 0.50)	0.43 (0.38, 0.49)
Magallanes	0.83 (0.72, 0.94)	0.83 (0.72, 0.96)

Health service	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Maule	0.31 (0.28, 0.33)	0.30 (0.28, 0.33)
Metro. Central	0.42 (0.38, 0.46)	0.42 (0.38, 0.46)
Metro. Norte	0.29 (0.27, 0.32)	0.29 (0.26, 0.31)
Metro. Occidente	0.36 (0.34, 0.38)	0.34 (0.32, 0.36)
Metro. Oriente	0.30 (0.28, 0.33)	0.30 (0.27, 0.33)
Metro. Sur	0.41 (0.39, 0.44)	0.40 (0.37, 0.43)
Metro. Sur Oriente	0.37 (0.34, 0.39)	0.36 (0.34, 0.39)
O'Higgins	0.43 (0.40, 0.47)	0.42 (0.39, 0.46)
Osorno	0.44 (0.38, 0.51)	0.43 (0.37, 0.51)
Reloncaví	0.42 (0.38, 0.47)	0.42 (0.37, 0.47)
Talcahuano	0.84 (0.76, 0.92)	0.81 (0.74, 0.90)
Valdivia	0.31 (0.27, 0.35)	0.30 (0.26, 0.35)
Valparaíso	0.69 (0.63, 0.74)	0.68 (0.62, 0.74)
Viña del Mar	0.67 (0.63, 0.71)	0.66 (0.62, 0.70)
Ñuble	1.32 (1.24, 1.40)	1.29 (1.21, 1.37)

Table 4: Crude and age- and sex-adjusted ADHD prevalence by health service in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

Health service	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Aconcagua	2.08 (1.95, 2.21)	2.04 (1.91, 2.19)
Aisén	2.23 (2.03, 2.44)	2.17 (1.97, 2.40)
Antofagasta	1.00 (0.94, 1.06)	0.98 (0.93, 1.04)
Araucanía Norte	1.33 (1.22, 1.45)	1.29 (1.18, 1.43)
Araucanía Sur	1.42 (1.36, 1.49)	1.38 (1.32, 1.45)
Arauco	1.64 (1.50, 1.78)	1.64 (1.50, 1.81)
Arica	1.14 (1.04, 1.24)	1.12 (1.02, 1.24)
Atacama	0.49 (0.44, 0.55)	0.49 (0.43, 0.56)
Biobío	2.27 (2.16, 2.38)	2.26 (2.15, 2.38)
Chiloé	2.96 (2.77, 3.15)	2.87 (2.68, 3.07)
Concepción	2.94 (2.84, 3.04)	3.00 (2.89, 3.11)
Coquimbo	1.98 (1.91, 2.05)	2.00 (1.92, 2.08)
Iquique	1.49 (1.40, 1.58)	1.50 (1.40, 1.60)
Magallanes	3.07 (2.87, 3.27)	3.06 (2.85, 3.29)
Maule	1.19 (1.14, 1.24)	1.15 (1.11, 1.21)
Metro. Central	1.53 (1.46, 1.59)	1.49 (1.43, 1.57)
Metro. Norte	1.42 (1.36, 1.47)	1.42 (1.36, 1.48)
Metro. Occidente	1.09 (1.05, 1.13)	1.12 (1.08, 1.17)
Metro. Oriente	1.22 (1.17, 1.27)	1.20 (1.15, 1.26)
Metro. Sur	1.42 (1.37, 1.48)	1.41 (1.35, 1.46)
Metro. Sur Oriente	1.56 (1.51, 1.61)	1.53 (1.48, 1.58)
O'Higgins	1.73 (1.66, 1.79)	1.69 (1.63, 1.76)
Osorno	1.05 (0.95, 1.14)	1.02 (0.92, 1.13)
Reloncaví	1.02 (0.95, 1.09)	0.99 (0.92, 1.07)
Talcahuano	3.07 (2.93, 3.22)	3.02 (2.87, 3.18)
Valdivia	1.08 (1.00, 1.16)	1.06 (0.98, 1.15)
Valparaíso	1.19 (1.12, 1.27)	1.19 (1.11, 1.27)
Viña del Mar	1.17 (1.12, 1.22)	1.15 (1.10, 1.20)

Health service	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Ñuble	2.12 (2.02, 2.22)	2.11 (2.00, 2.22)

Considering school data by health service, shown in Table 3, the adjusted prevalence of autism was significantly higher in Ñuble at 1.29% (1.21 - 1.37%) than other health services with non-overlapping confidence intervals. Adjusted prevalence was lowest in Metropolitano Norte at 0.29% (0.26 - 0.31%) and Araucanía Norte at 0.30% (0.24- 0.38%). Autism prevalence peaked in the 6-8 age band across all health services except Chiloé and Magallanes where it peaked in the 9-11 band, see Supplementary Table 21 and Supplementary Figure 41.

ADHD prevalence also varied across health services, as shown in Table 4. Prevalence was significantly lower in Atacama than other regions at 0.49% (0.43 - 0.56%). Magallanes, Talcahuano, Concepción, Chiloé, Biobío, Aisén, Ñuble, Aconcagua and Coquimbo all had adjusted prevalence of 2.00% or higher and there was a significant gap between these health services and those with lower prevalence. There was also a significant gap separating Magallanes, Talcahuano, Concepción and Chiloé as the highest prevalence health services. See also Supplementary Table 22 and Supplementary Figure 42.

Table 5: Crude and age- and sex-adjusted autism prevalence by monthly school fee (Peso) in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

School fee	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Free	0.57 (0.56, 0.58)	0.55 (0.54, 0.56)
$1,000-$10,000	0.71 (0.22, 1.21)	0.69 (0.29, 3.25)
$10,001-$25,000	0.20 (0.15, 0.25)	0.20 (0.16, 0.27)
$25,001-$50,000	0.30 (0.28, 0.33)	0.32 (0.29, 0.34)
$50,001-$100,000	0.39 (0.36, 0.41)	0.40 (0.37, 0.43)
$100,001+	0.05 (0.04, 0.05)	0.05 (0.04, 0.06)
No information	0.50 (0.44, 0.57)	0.46 (0.40, 0.52)

Table 6: Crude and age- and sex-adjusted ADHD prevalence by monthly school fee (Peso) in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

School fee	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Free	1.65 (1.63, 1.67)	1.62 (1.60, 1.64)
$1,000-$10,000	0.18 (0.00, 0.43)	0.11 (0.01, 2.74)
$10,001-$25,000	1.06 (0.95, 1.16)	1.05 (0.94, 1.17)
$25,001-$50,000	1.59 (1.54, 1.65)	1.65 (1.59, 1.72)
$50,001-$100,000	1.90 (1.84, 1.95)	1.90 (1.84, 1.96)
$100,001+	0.22 (0.21, 0.24)	0.23 (0.21, 0.25)
No information	1.21 (1.11, 1.31)	1.22 (1.13, 1.33)

For school fees, which were used here as a proxy for SES, autism prevalence was highest among students that received free or low fee education, though the sample size for students that paid $1,000-$10,000 monthly was very small, see Table 5. ADHD prevalence was variable across school fee levels, with the $1,000-$10,000 band having low prevalence and very few cases, see Table 6. For both autism and ADHD, prevalence was very low among students paying more than $100,000 monthly, suggesting students from wealthier families may not

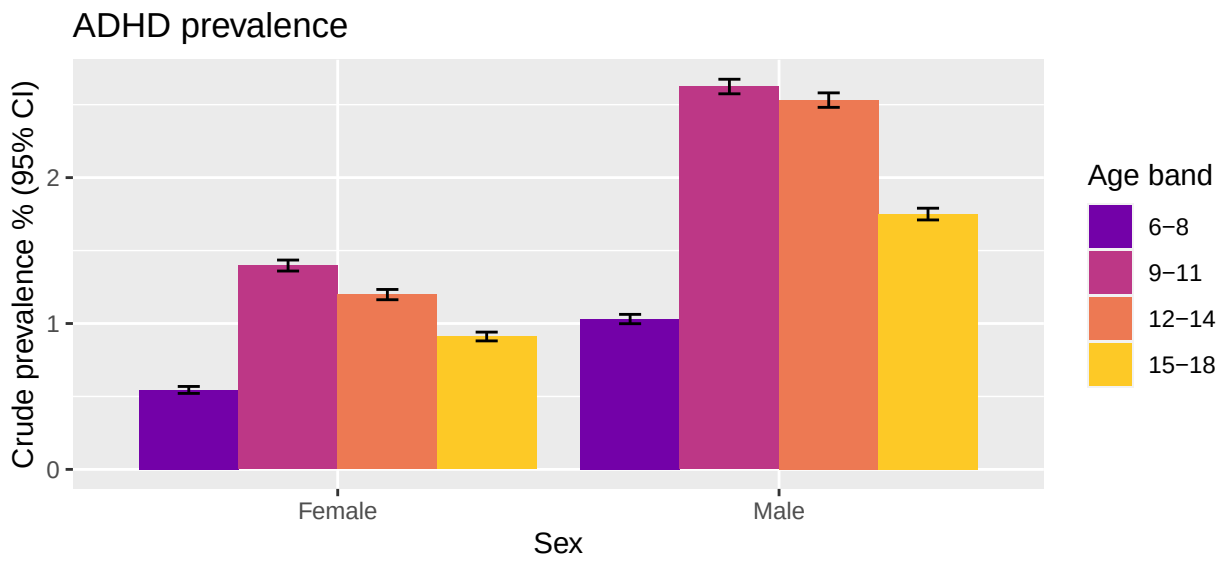


Figure 8: Crude prevalence of ADHD in school data by age band and sex. Bars show 95% normal confidence intervals.

have accessed SEED or may not have been eligible for it due to attending non-subsidised private schools. See also Supplementary Tables 23 and 24 and Supplementary Figures 43, 44.

Table 7: Crude and age- and sex-adjusted autism prevalence by ethnicity in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

Ethnicity	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Mapuche	0.36 (0.33, 0.39)	0.34 (0.31, 0.37)
Other Indigenous group	0.37 (0.20, 0.54)	0.38 (0.21, 0.75)
No Indigenous group	0.43 (0.42, 0.44)	0.42 (0.41, 0.43)

Table 8: Crude and age- and sex-adjusted ADHD prevalence by ethnicity in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

Ethnicity	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Mapuche	1.38 (1.32, 1.44)	1.33 (1.28, 1.39)
Other Indigenous group	1.12 (0.82, 1.41)	1.10 (0.83, 1.55)
No Indigenous group	1.59 (1.57, 1.61)	1.58 (1.56, 1.60)

For both autism and ADHD, adjusted prevalence was significantly lower for Mapuche students than non-Indigenous students in the Aysén, Biobío, La Araucanía, Los Lagos, Los Rios, Magallanes and Región Metropolitana de Santiago regions, as shown in Tables 7 and 8. Autism prevalence for students belonging to other Indigenous groups fall in between and was not significantly different, and ADHD prevalence for other Indigenous groups was lower but not significantly. See also Supplementary Tables 25 and 26 and Supplementary Figures 45 and 46.

Table 9: Crude and age- and sex-adjusted autism prevalence by school's rurality in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

School rurality	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Rural	0.66 (0.63, 0.69)	0.57 (0.54, 0.61)
Urban	0.46 (0.45, 0.47)	0.46 (0.45, 0.46)

Table 10: Crude and age- and sex-adjusted ADHD prevalence by school's rurality in Chile school data. Crude prevalence has 95% normal confidence intervals and adjusted prevalence has 95% gamma confidence intervals.

School rurality	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
Rural	1.77 (1.72, 1.82)	1.67 (1.61, 1.74)
Urban	1.49 (1.48, 1.50)	1.48 (1.47, 1.50)

Autism and ADHD were both significantly more prevalent for students at rural schools than urban schools at 0.57% (0.54 - 0.61%) and 0.46% (0.45 - 0.46%) respectively for autism and 1.67% (1.61 - 1.74%) and 1.48% (1.47 - 1.50%) respectively for ADHD, as shown in Tables 9 and 10. See also Supplementary Tables 27 and 28 and Supplementary Figures 47 and 48.

These results together indicate that for autism and ADHD there were differences in prevalence across location and demographic features.

5.3 Clinical data

The validated clinical dataset has data on 1,570 appointments for 253 unique patients with autism aged 1-18 in 2021, of which 247 patients have lived in a commune in the SSAS catchment area during the period the data covers. Among these, most were male, live in urban areas and have Chilean but not Mapuche ethnicity. The mean age is 9.4 years. Many patients live in Villarica commune as the validated clinical data is from Villarica Hospital. FONASA-A, indicative of low SES, is the most common level of health insurance, and few patients have another recorded disability or have experienced foster care.

5.4 Machine learning with clinical data

Multiple correspondence analysis was first conducted with all features thought to be associated with autism diagnosis with no imputation. Figure 9 shows approximately 14.6% of the variance in these data were captured by the first two dimensions of MCA. Disability and foster care status were well separated by the first dimension but Figures 18 and 19 show that this separation was primarily driven by whether information was available for these features. Ethnicity, age band and commune of residence were well separately by the second dimension of MCA, as shown in Figure 9, and were somewhat separated by the first dimension. Figures 10 and 11 further show the importance of categories within the foster care, disabilities, ethnicity, age band and commune features for explaining the variance in this data. In particular, categories that explained more of the variance include not having disability or foster care status, or not having information on these, having age of 0-2 or 3-5, being foreign or Chilean, living in Toltén and having private health insurance. Examining the clustering of individual patients by the first two dimensions of MCA, Figure 12 demonstrates that patients in age bands 0-2 and 3-5 clustered well and those in older age bands did not. There was some clustering by ethnicity in Figure 13 but it was obscured by the separation of points into the two larger clustered defined by having or not having information on disability and foster care status. Figure 14 shows clear separation of Toltén and Nueva Imperial communes, however it is important to note here that these and several other

Table 11: Count and percentage of features' values in the validated clinical dataset.

Feature	Available values	Count (%)
Sex	Female	55 (22.27%)
	Male	192 (77.73%)
Age band	Age 0-2	<20
	Age 3-5	55 (22.27%)
	Age 6-8	37 (14.98%)
	Age 9-11	56 (22.67%)
	Age 12-14	45 (18.22%)
	Age 15-18	41 (16.60%)
Commune	Curarrehue	<20
	Freire	<20
	Gorbea	<20
	Loncoche	39 (15.79%)
	Nueva Imperial	<20
	Pitrufquén	<20
	Pucón	50 (20.24%)
	Temuco	<20
	Teodoro Schmidt	<20
	Toltén	<20
	Villarrica	137 (55.47%)
Private health level	FONASA - A	99 (40.08%)
	FONASA - B	67 (27.13%)
	FONASA - C	35 (14.17%)
	FONASA - D	38 (15.38%)
	Private Health Insurance	<20
Ethnicity	Mapuche	<20
	Chilean	131 (53.04%)
	Foreign	32 (12.96%)
	No ethnicity information	80 (32.39%)
Rurality	Rural	26 (10.53%)
	Urban	221 (89.47%)
Disability	Yes disability	<20
	No disability	78 (31.58%)
	No disability information	157 (63.56%)
Foster care	Yes foster care	<20
	No foster care	88 (35.63%)
	No foster care information	157 (63.56%)

communes were represented by only one patient, see Table 11. There was decent clustering of patients in Temuco and Pitrufquén communes. Figure 15 shows possible separation for patients with private health insurance and Figures 16 and 17 show little clustering by sex or rurality of residence.

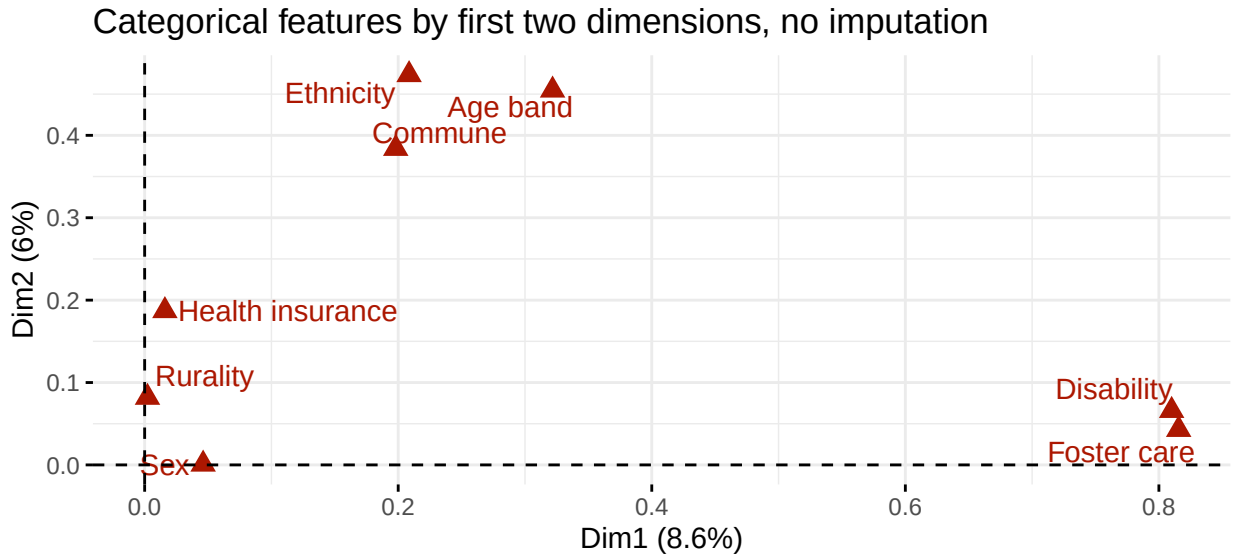


Figure 9: Categorical features by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation.

Exchanging disability and foster care status for their imputed versions led to the MCA capturing approximately 14.1% of the variance in that data with its first two dimensions, as shown in Figure 20. Disability and foster care status were no longer well separated by the first dimension and Figures 29 and 30 show that the patients who had experienced foster care did cluster well but the patients with a disability did not. In Figure 20, age band and ethnicity were now well separated by both dimensions and commune mostly by the second. With the reduced importance of disability and foster care, age bands 0-2 and 3-5, foreign, no information and Chilean ethnicity and Toltén and Nueva Imperial communes contributed most to the first two dimensions, see Figures 21 and 22. Again patients with age bands 0-2 and 3-5 clustered well and older age bands did not, as shown in Figure 23. Figure 24 shows much clearer clustering of ethnicity than before. For communes with more than one patient resident, clustering was less clear than before with only Pitrufquén well separated by these dimensions, see Figure 25. Clustering by SES, sex and rurality was weak, see Figures 26, 27 and 28.

MCA on the validated clinical data using only age band, ethnicity and commune resulted in the first two dimensions capturing 17.8% of the variance in that data, see Figure 31, and these features were fairly well represented by the first two dimensions. Again age bands 0-2 and 3-5, foreign, no information and Chilean ethnicity and Toltén and Nueva Imperial communes contributed most to the first two dimensions, see Figures 32 and 33. By patient, age bands 0-2 and 3-5 still clustered well (Figure 34), ethnicity clusters were very distinct (Figure 35), and communes showed more structure than previously (Figure 36).

5.5 Linkage of school and patient records

Table 12 shows that students living in SSAS were significantly different to students living in the other health services with regard to age band, SES, ethnicity, rurality and whether they accessed SEED at all, but did not differ by sex.

In the school records, 132,242 students lived in the SSAS health service catchment, of which 488 (0.37%) had autism. Among the students with autism, shown in Table 13, there were many more males than females, Temuco was the most common commune of residence and most students had a proxy SES of 1.

Aggregating the combined clinical data to patient-level data for linkage resulted in the patient dataset with

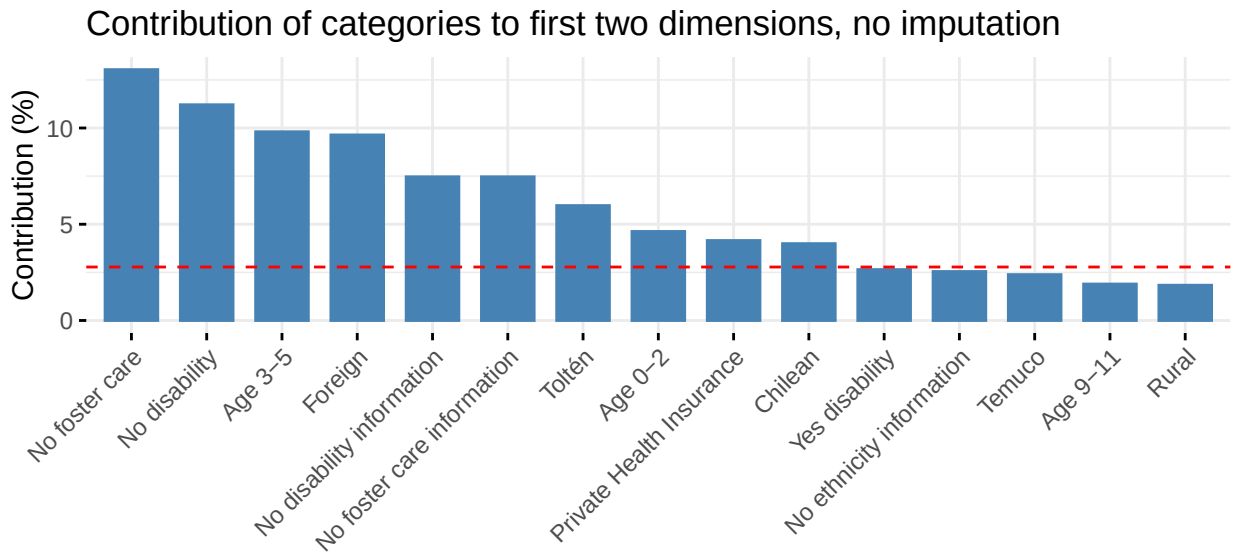


Figure 10: Contribution of the top 15 categories to the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation. The red line shows the expected average if contributions were uniform.

Table 12: Percentage of features' values for each category for the students in the school data resident in SSAS and those not resident in SSAS with the p-value resulting from Chi squared tests of the difference between SSAS and non-SSAS students for each feature.

Feature	Available values	% for SSAS	% for non-SSAS	p-value
Sex	Female	48.95%	48.65%	0.0328
	Male	51.05%	51.35%	
Age band	6-8	24.08%	24.51%	0.00243
	9-11	25.11%	25.11%	
	12-14	24.69%	24.52%	
	15-18	26.13%	25.86%	
School fee	Free	84.15%	71.10%	«0.001
	$1,000-$10,000	0.00%	0.04%	
	$10,001-$25,000	0.56%	1.22%	
	$25,001-$50,000	2.74%	6.95%	
	$50,001-$100,000	6.42%	8.97%	
	$100,001+	4.67%	10.07%	
	No information	1.46%	1.64%	
Ethnicity	Mapuche	33.58%	4.51%	«0.001
	Other Indigenous group	0.13%	1.22%	
	No Indigenous group	66.29%	94.27%	
Rurality	Rural	16.68%	7.42%	«0.001
	Urban	83.32%	92.58%	
Accesses SEED	No	83.25%	89.13%	«0.001
	Yes	16.75%	10.87%	

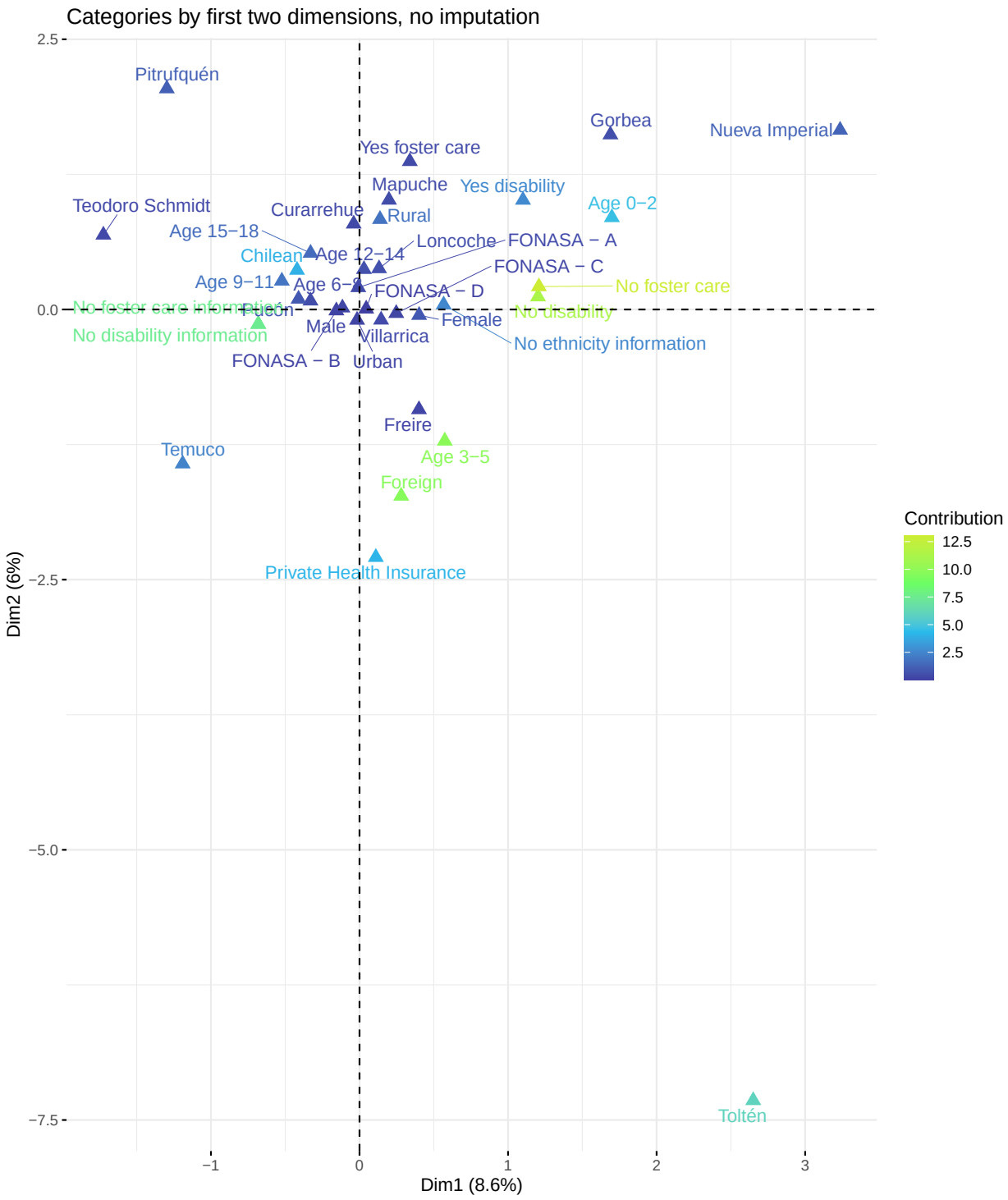


Figure 11: Available categories by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation. Brigther, more yellow colours indicate larger contribution to the first two dimensions.

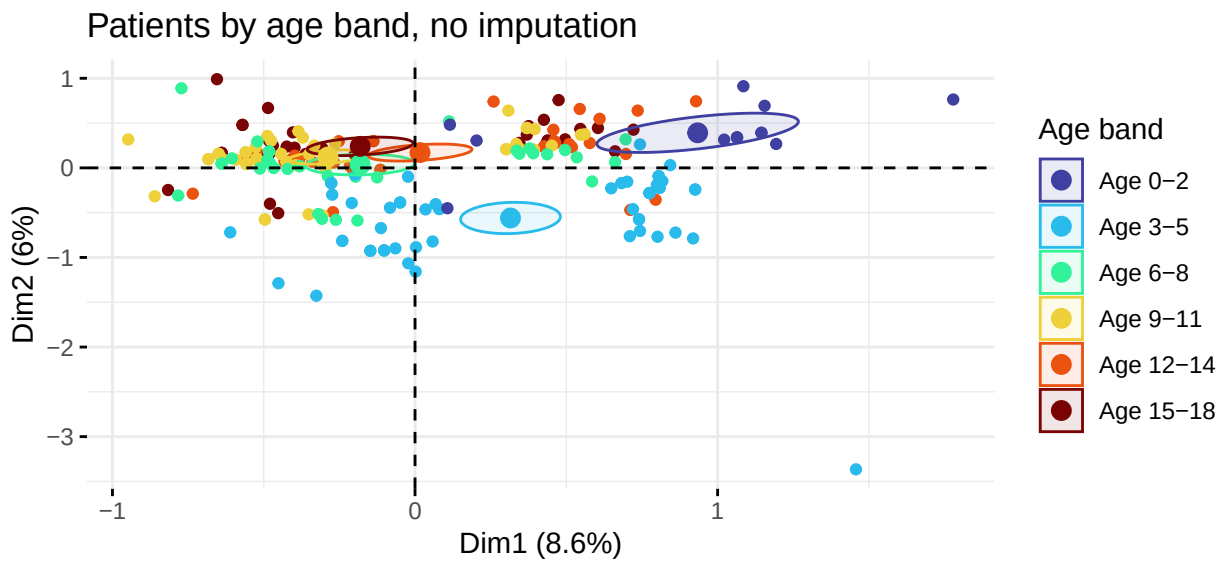


Figure 12: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by age band.

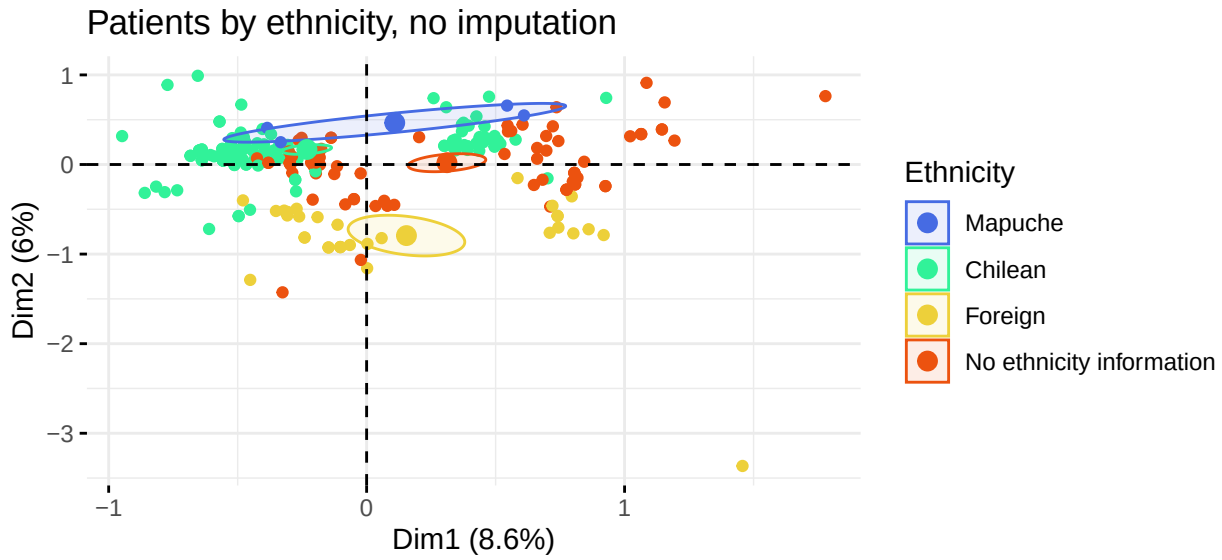


Figure 13: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by ethnicity.

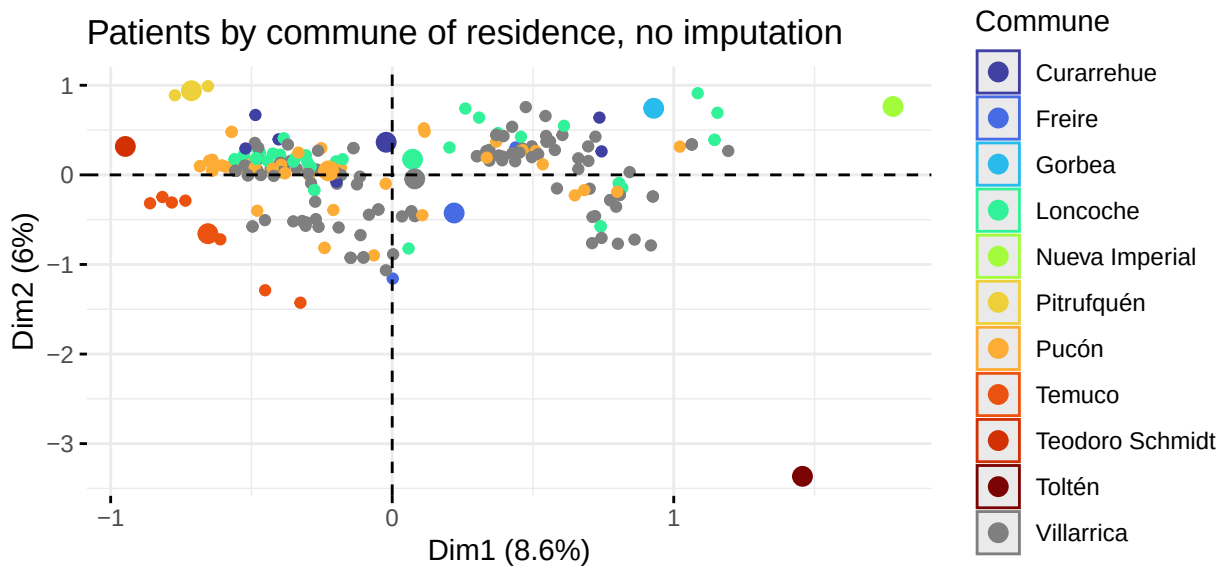


Figure 14: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by commune of residence.

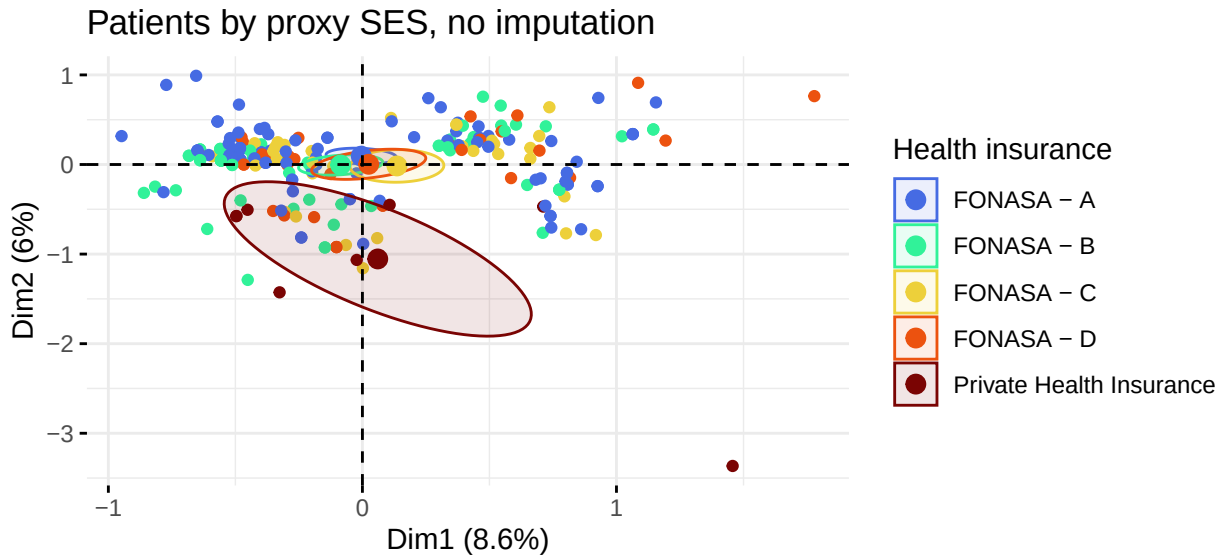


Figure 15: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by health insurance level.

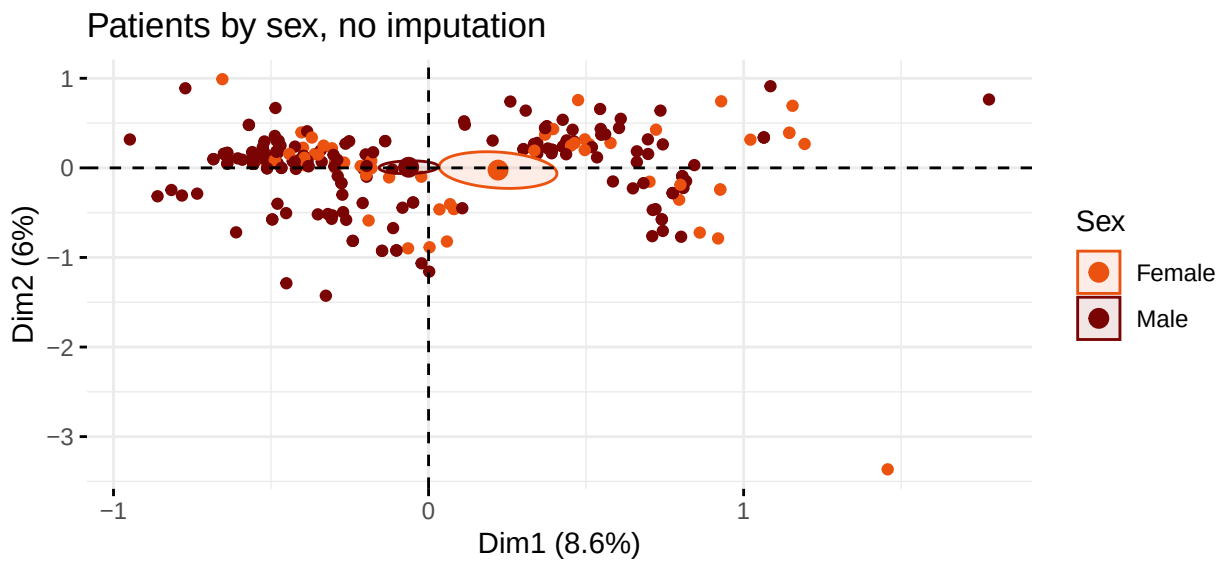


Figure 16: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by sex.

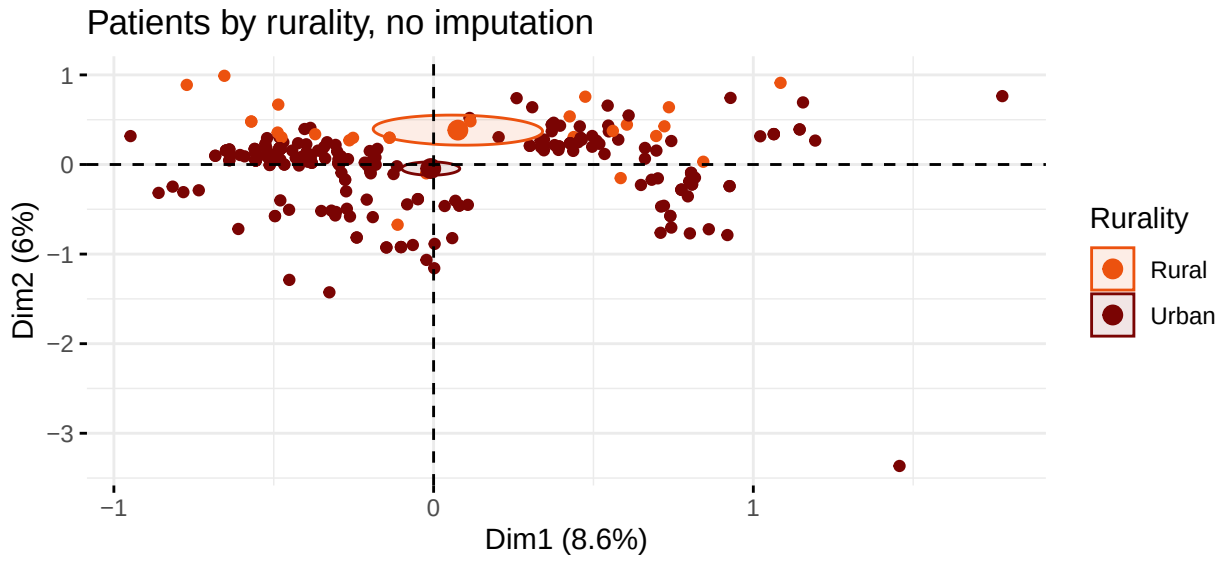


Figure 17: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by rurality of residence.

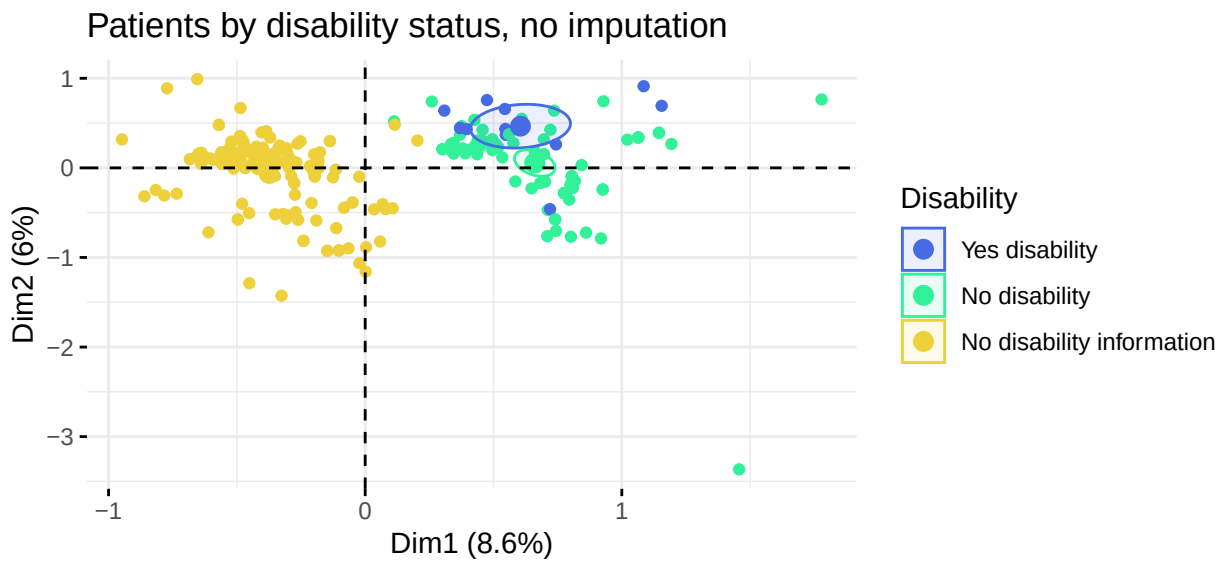


Figure 18: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by disability status.

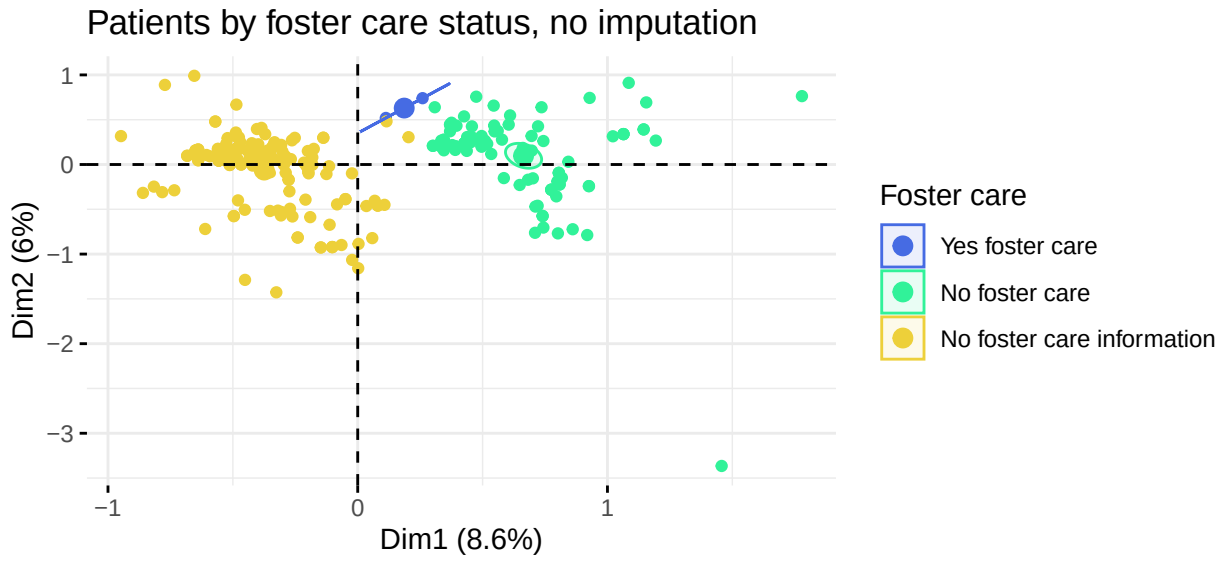


Figure 19: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features without imputation, coloured by foster care status.

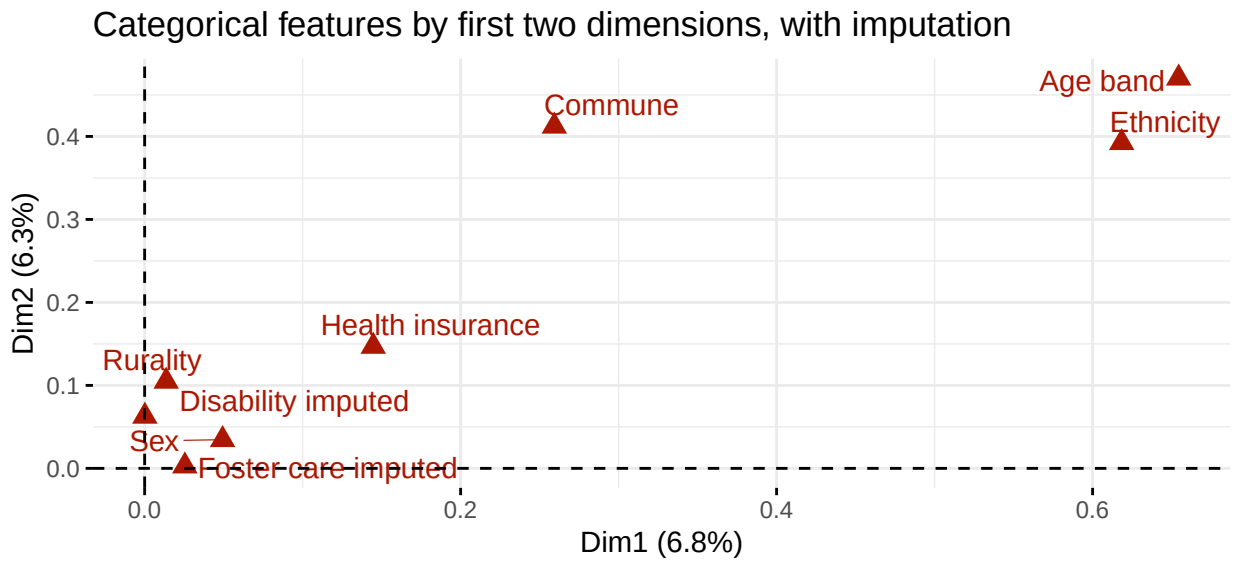


Figure 20: Categorical features by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation.

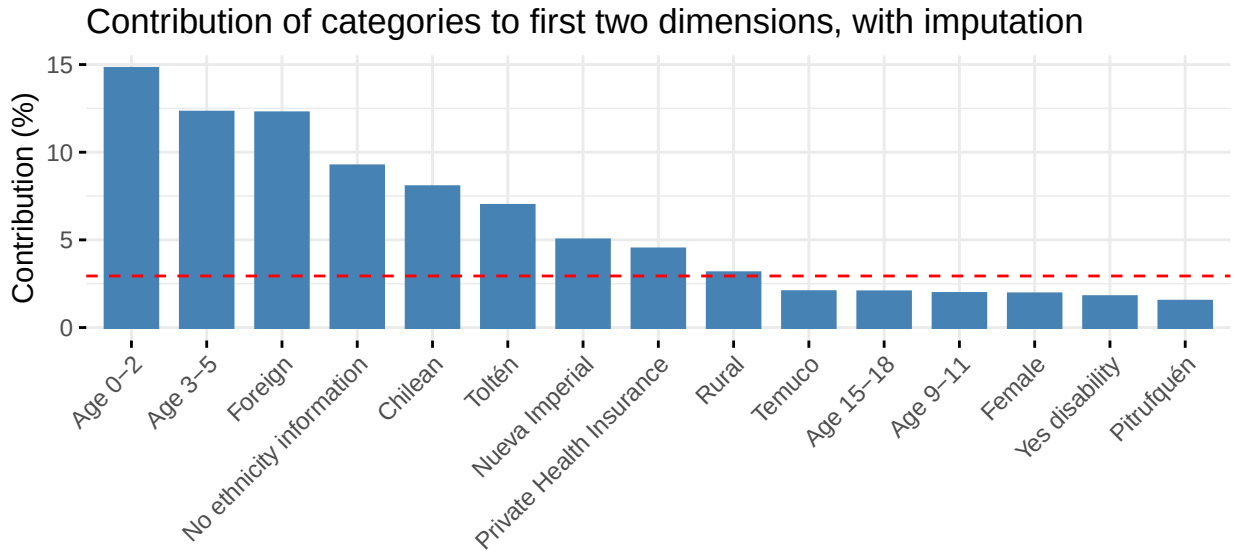


Figure 21: Contribution of the top 15 categories to the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation. The red line shows the expected average if contributions were uniform.

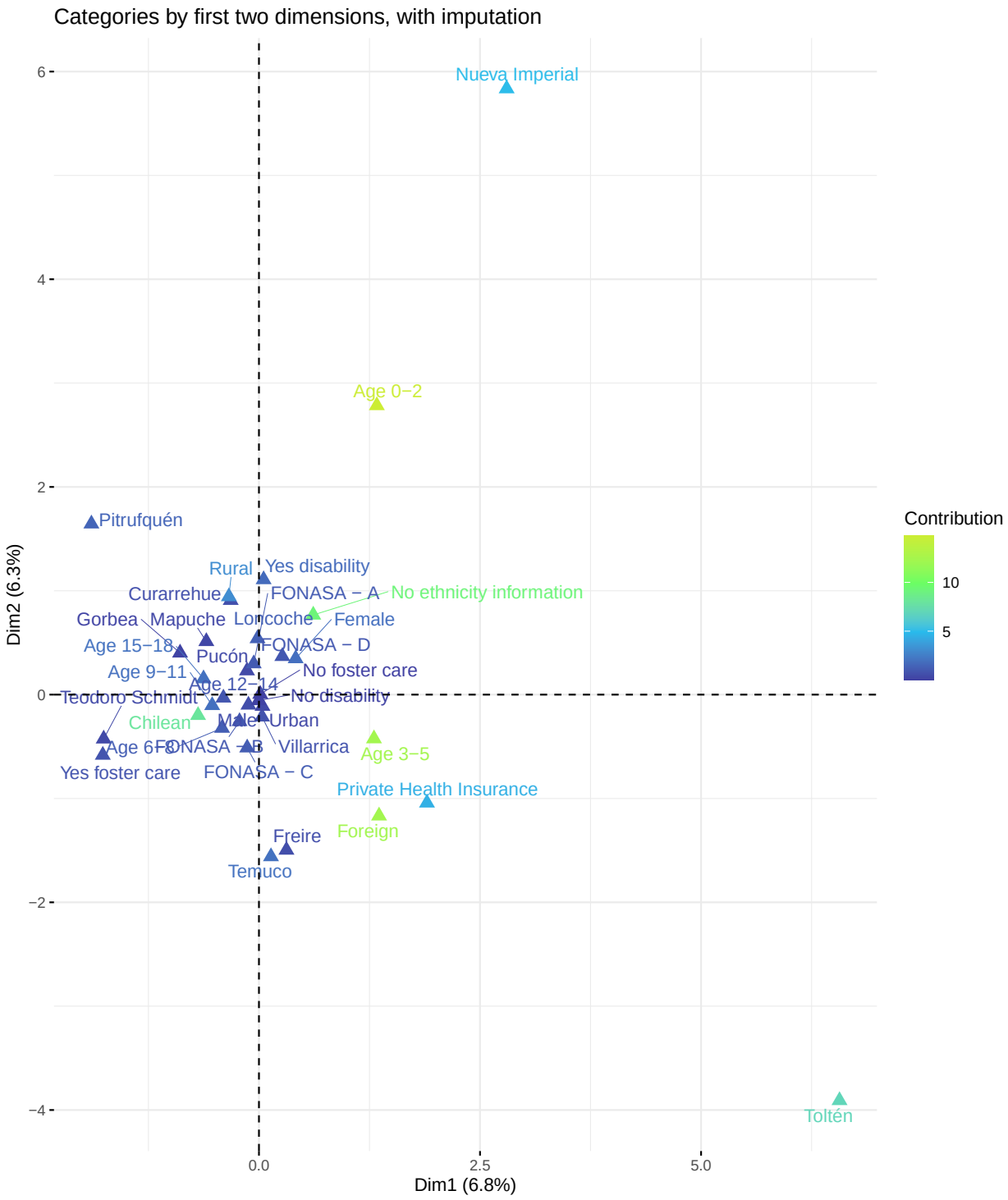


Figure 22: Available categories by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation. Brigther, more yellow colours indicate larger contribution to the first two dimensions.

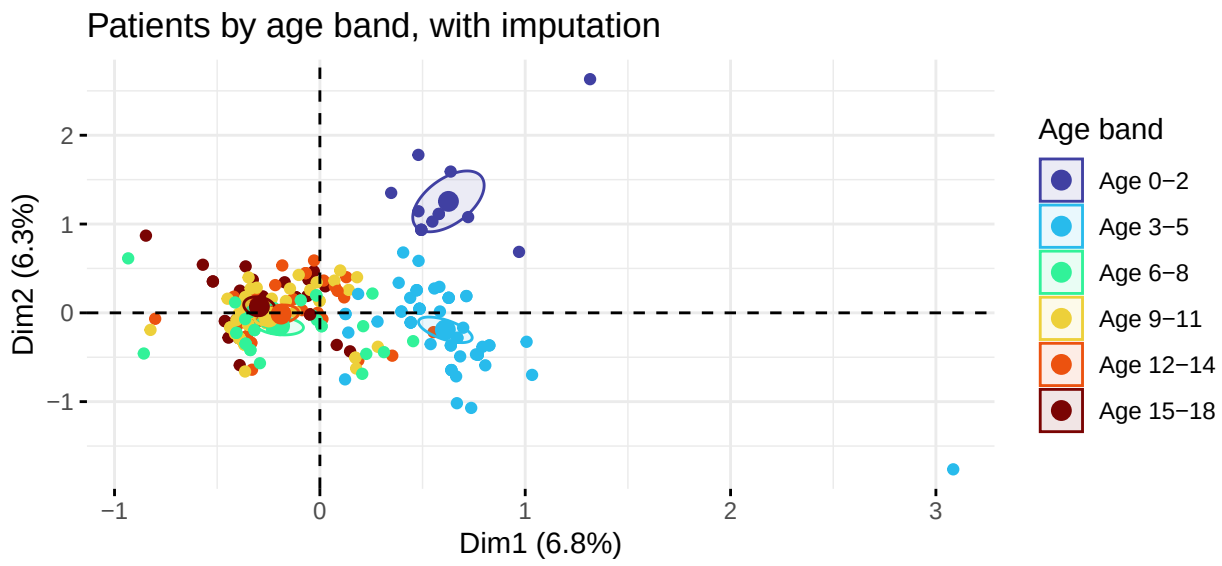


Figure 23: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by age band.

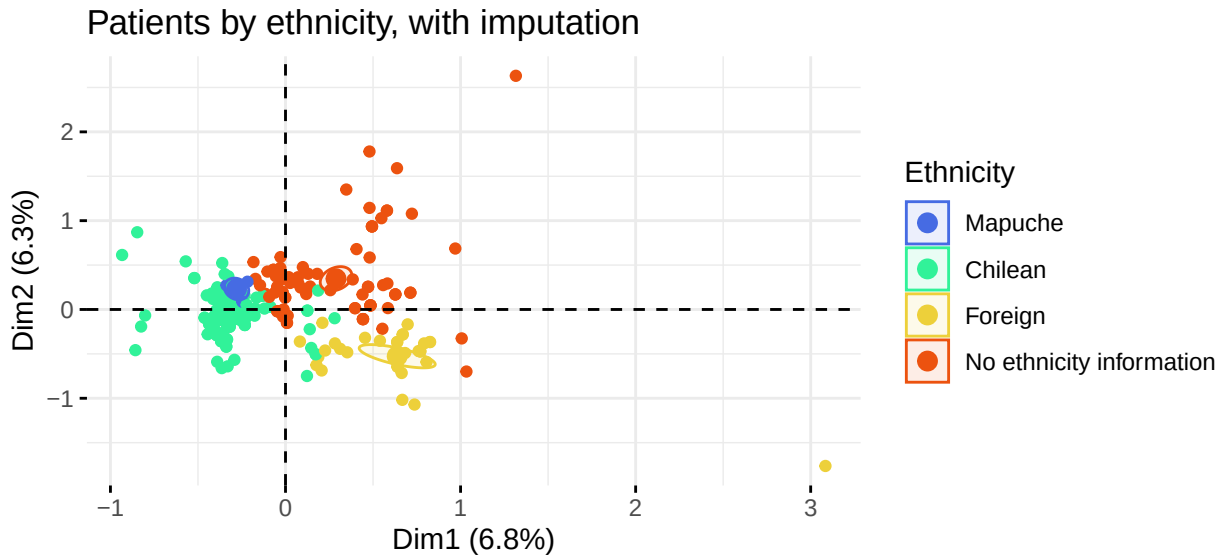


Figure 24: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by ethnicity.

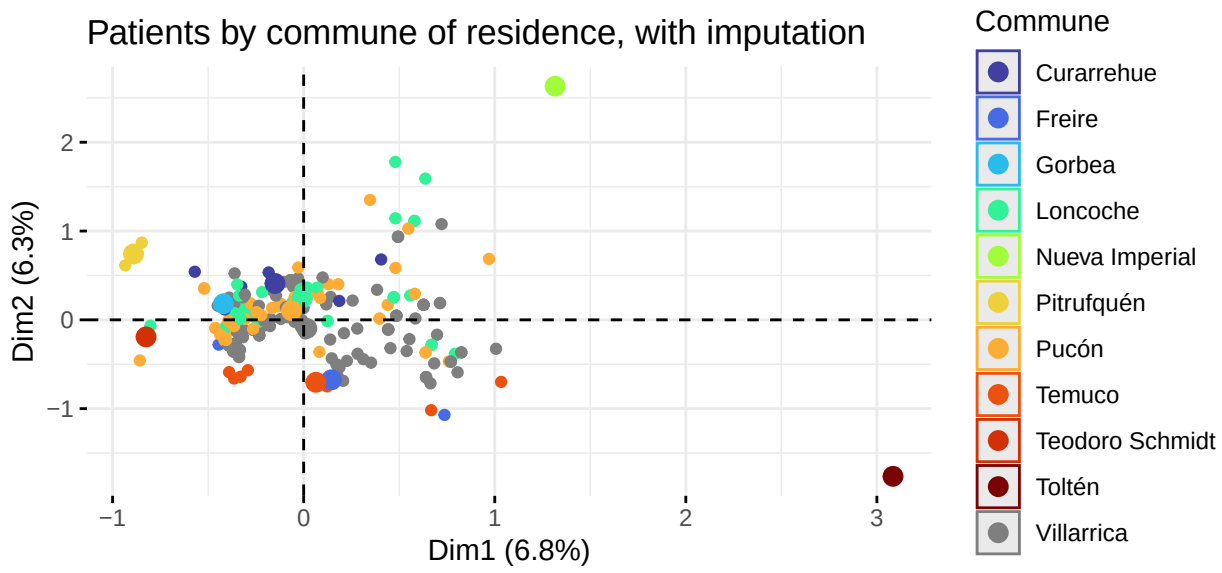


Figure 25: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by commune of residence.

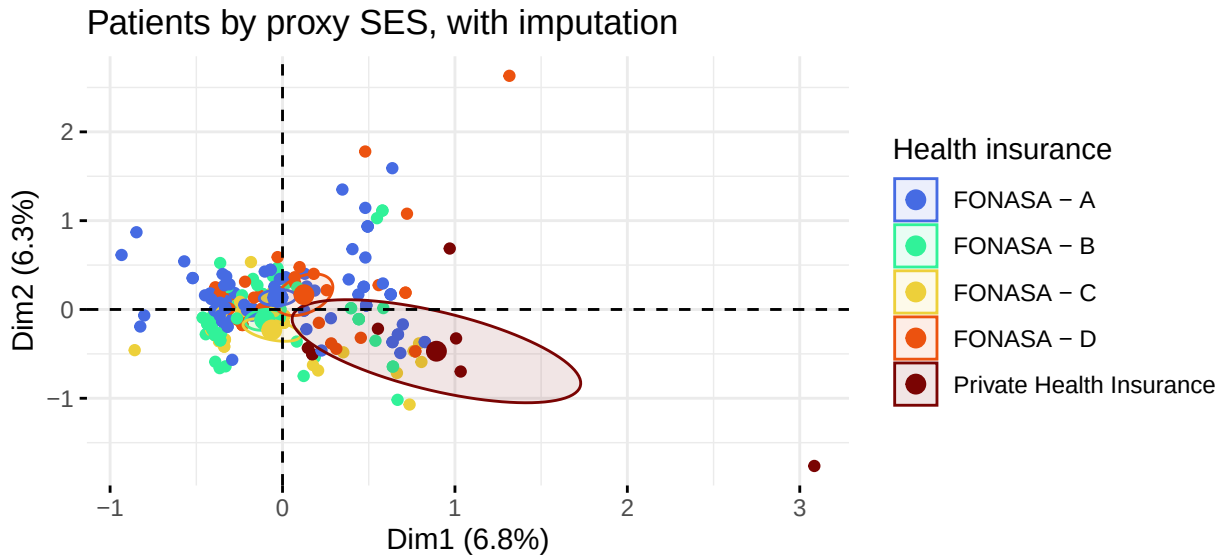


Figure 26: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by health insurance level.

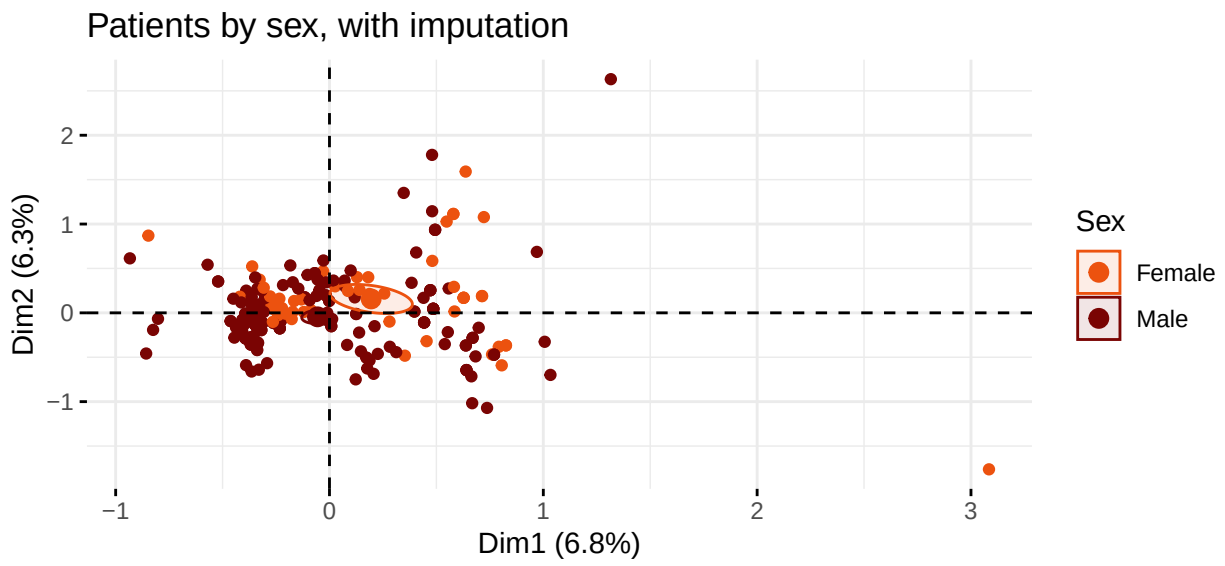


Figure 27: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by sex.

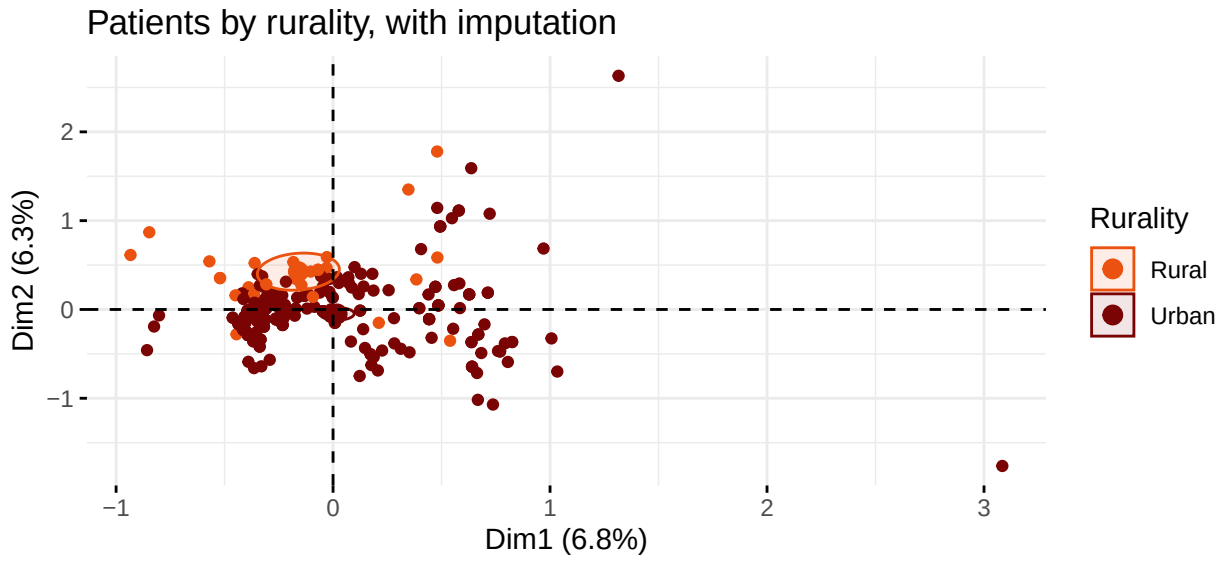


Figure 28: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by rurality of residence.

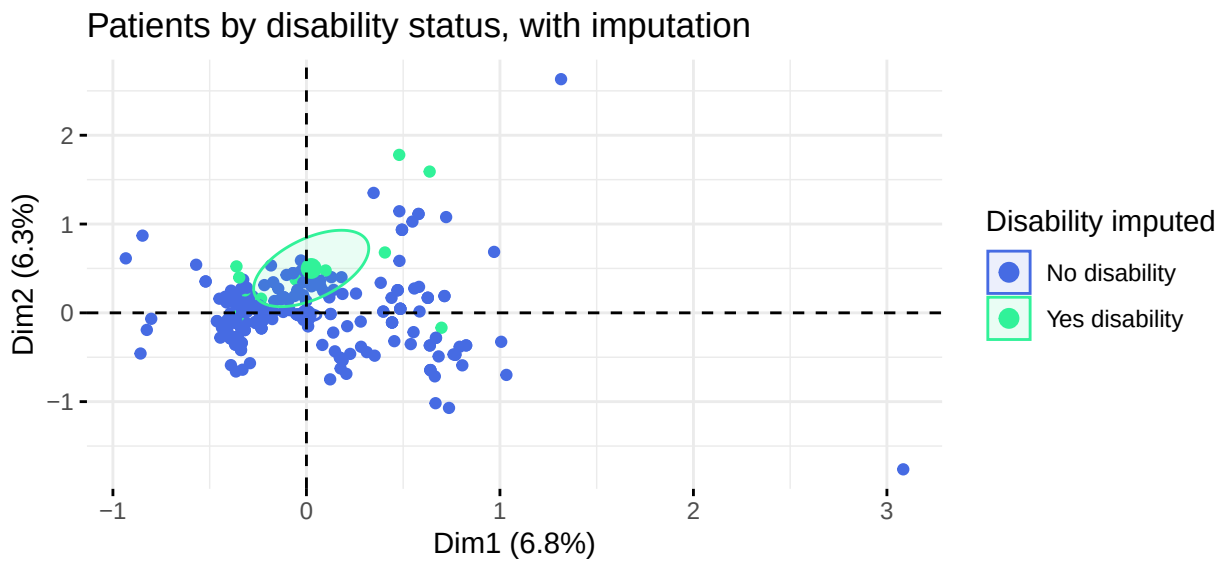


Figure 29: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by disability status.

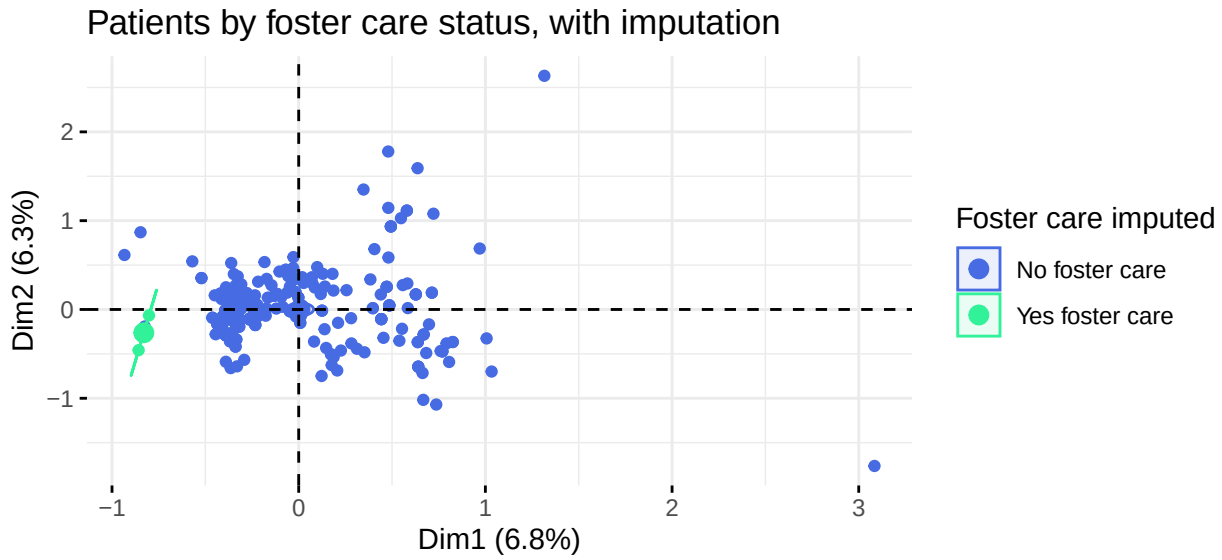


Figure 30: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using all features with imputation, coloured by foster care status.

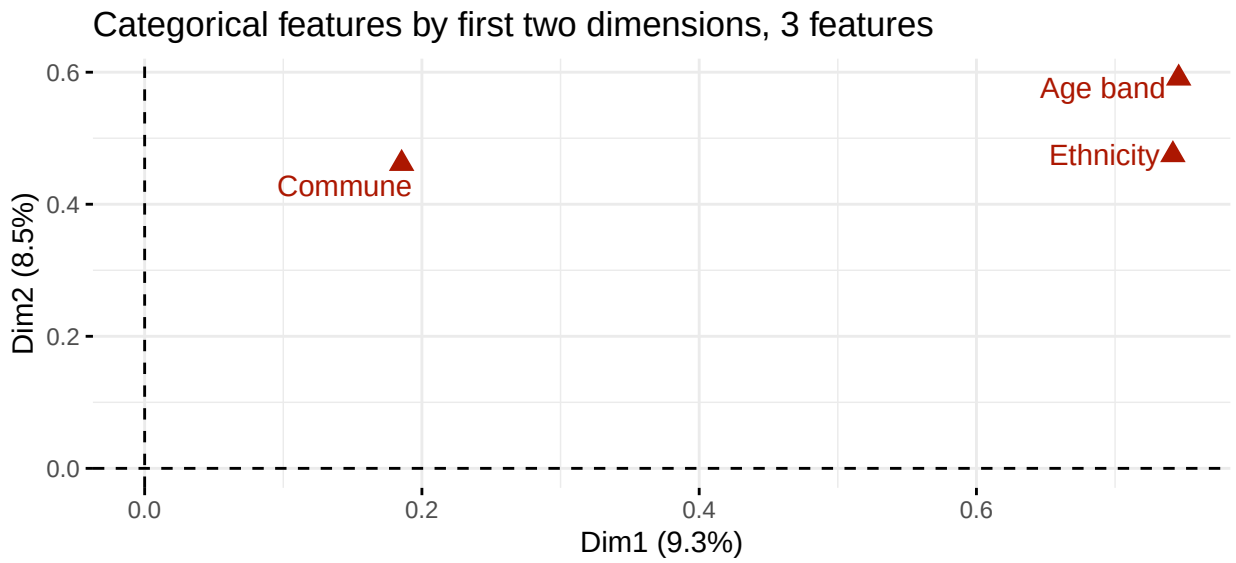


Figure 31: Categorical features by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using patients' age band, commune of residence and ethnicity.

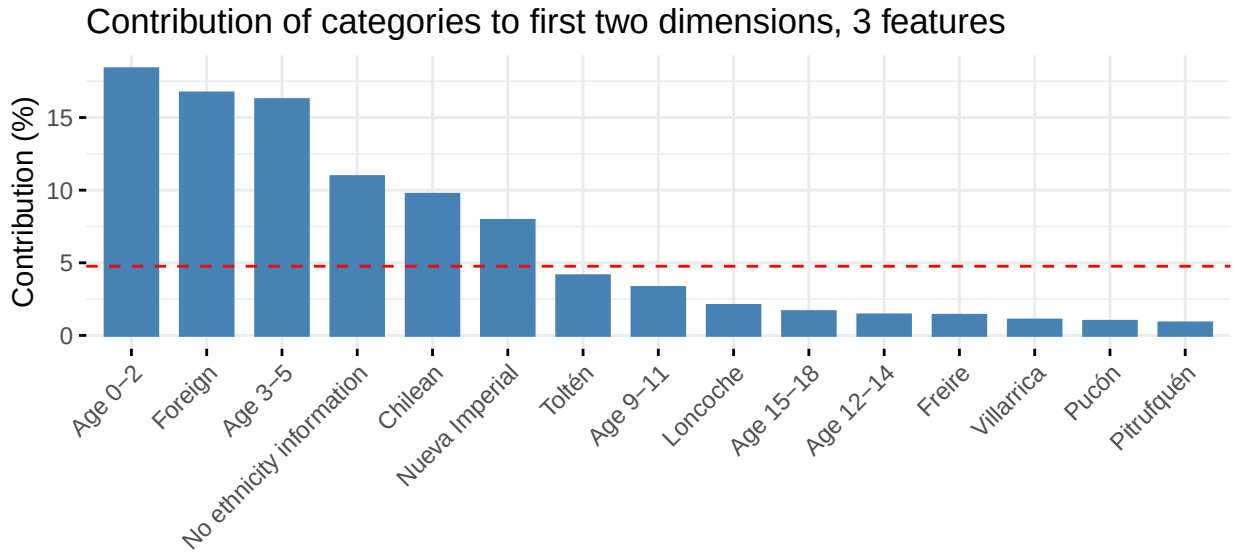


Figure 32: Contribution of the top 15 categories to the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using patients' age band, commune and ethnicity. The red line shows the expected average if contributions were uniform.

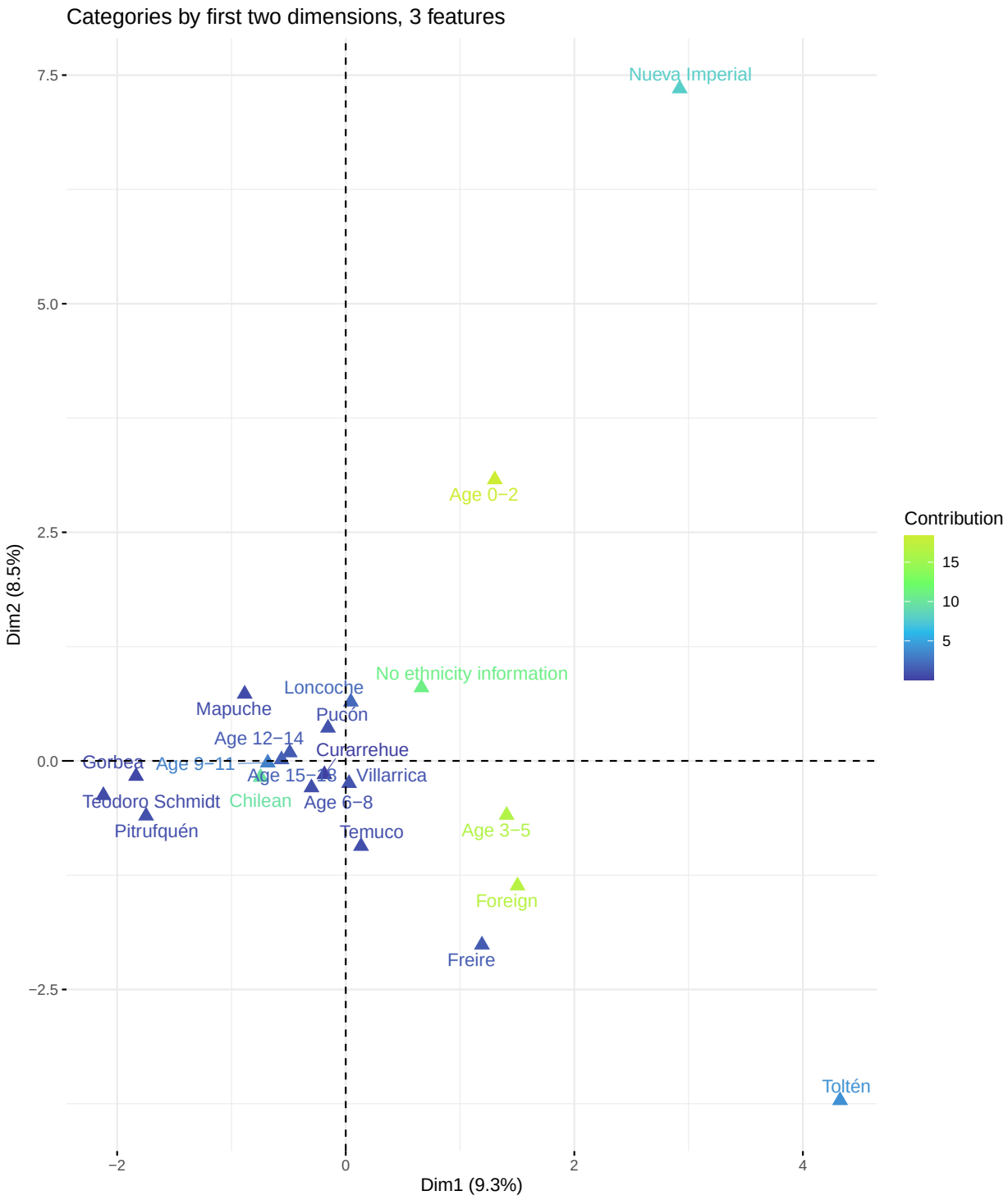


Figure 33: Available categories by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using patients' age band, commune and ethnicity. Brigther, more yellow colours indicate larger contribution to the first two dimensions.

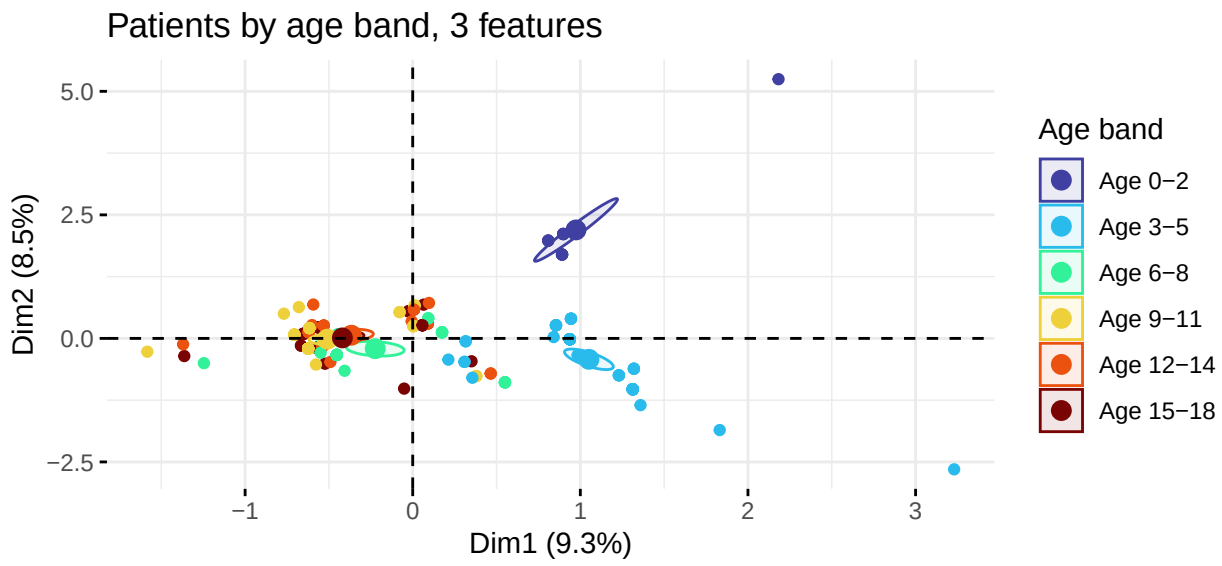


Figure 34: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using patients' age band, commune and ethnicity, coloured by age band.

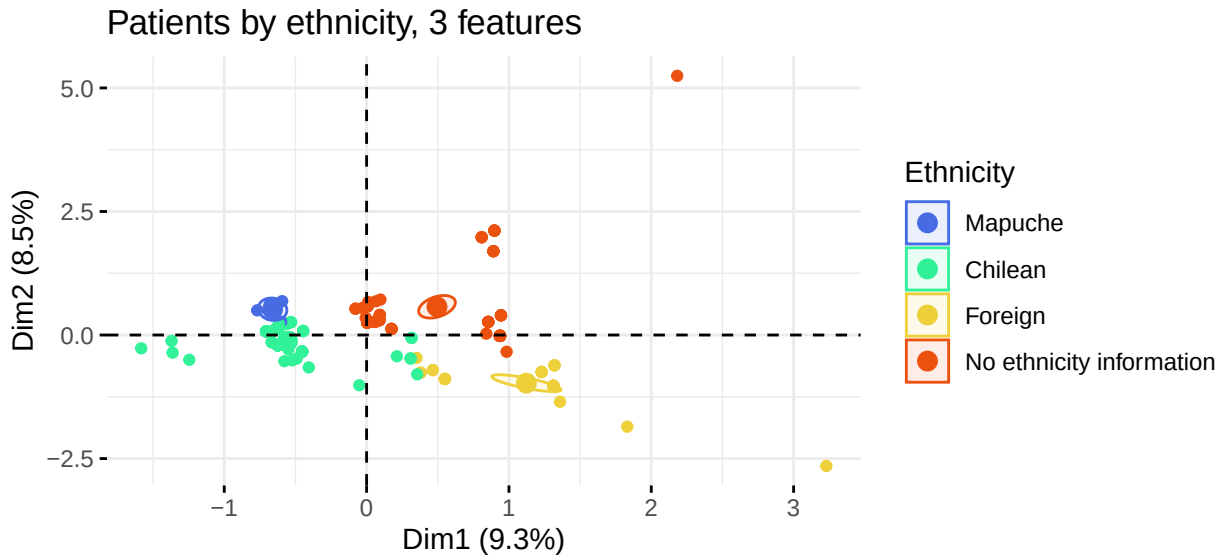


Figure 35: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using age band, commune and ethnicity, coloured by ethnicity.

Table 13: Count and percentage of features' values in the SSAS school dataset which comprises students with autism resident in the SSAS catchment. Proxy SES is school fee group.

Feature	Available values	Count (%)
Sex	Female	61 (12.50%)
	Male	427 (87.50%)
Age band	6-8	165 (33.81%)
	9-11	136 (27.87%)
	12-14	106 (21.72%)
	15-18	81 (16.60%)
Commune	Carahue	<20
	Cholchol	<20
	Cunco	<20
	Curarrehue	<20
	Freire	<20
	Galvarino	<20
	Gorbea	<20
	Lautaro	29 (5.94%)
	Loncoche	<20
	Melipeuco	<20
	Nueva Imperial	<20
	Padre Las Casas	35 (7.17%)
	Perquenco	<20
	Pitrufquén	<20
	Pucón	<20
	Saavedra	<20
	Temuco	236 (48.36%)
	Teodoro Schmidt	<20
	Toltén	<20
	Vilcún	<20
	Villarrica	30 (6.15%)
Proxy SES	1	417 (85.45%)
	2	48 (9.84%)
	3	<20
	NA	22 (4.51%)

Table 14: Count and percentage of features' values in the patient dataset. Proxy SES is health insurance contribution group.

Feature	Available values	Count (%)
Sex	Female	276 (20.06%)
	Male	1100 (79.94%)
Age band	6-8	413 (30.01%)
	9-11	373 (27.11%)
	12-14	301 (21.88%)
	15-18	289 (21.00%)
Commune	Carahue	44 (3.20%)
	Cholchol	<20
	Cunco	30 (2.18%)
	Curarrehue	<20
	Freire	26 (1.89%)
	Galvarino	21 (1.53%)
	Gorbea	20 (1.45%)
	Lautaro	101 (7.34%)
	Loncoche	33 (2.40%)
	Melipeuco	<20
	Nueva Imperial	78 (5.67%)
	Padre Las Casas	136 (9.88%)
	Perquenco	<20
	Pitrufquén	42 (3.05%)
	Pucón	42 (3.05%)
	Saavedra	<20
	Temuco	553 (40.19%)
	Teodoro Schmidt	<20
	Toltén	<20
	Vilcún	53 (3.85%)
	Villarrica	124 (9.01%)
Proxy SES	1	577 (41.93%)
	2	772 (56.10%)
	3	27 (1.96%)

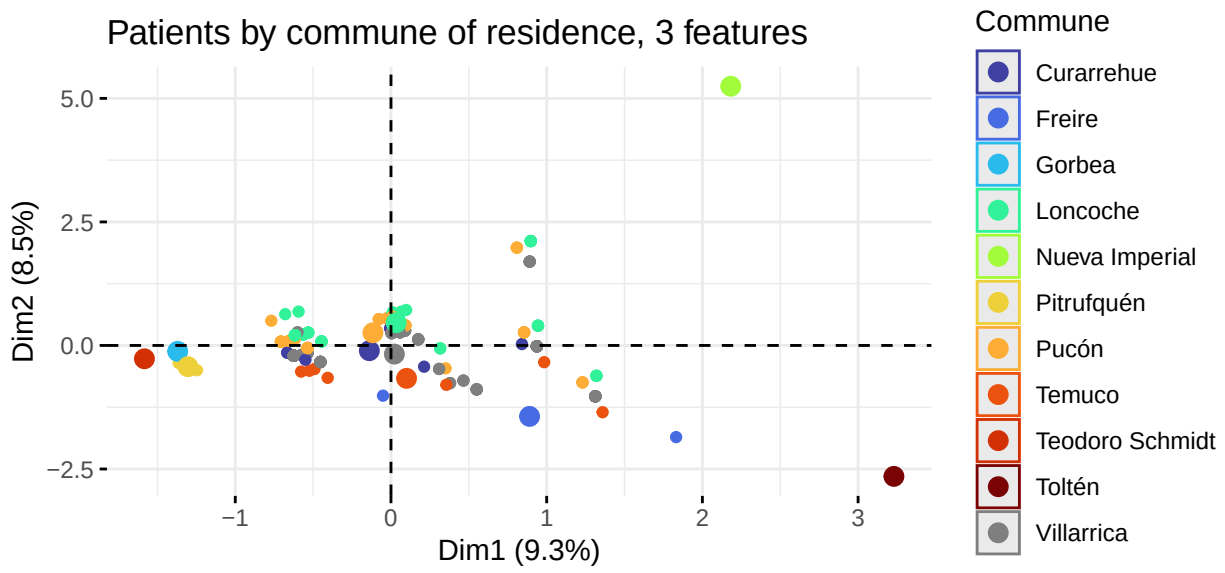


Figure 36: Patients by the first two dimensions of multiple correspondence analysis on autism patients in the validated clinical data using age band, commune and ethnicity, coloured by commune of residence.

1,376 records for 1,365 unique patients as 9 patients lived in 2 communes and 1 lived in 3 communes during the period covered by the data see Figure 4. Table 14 shows a similar pattern of category frequencies in the patient data as in the SSAS school data: males were more prevalent and Temuco was the most common commune of residence, however most patients had a proxy SES of 2 rather than 1.

5.5.1 Manual record linkage

Using perfect match on sex, date of birth, commune of residence and the proxies for SES, 79 perfect matches were manually found between the SSAS school and patient records. Of these, 77 unique SSAS school records were perfectly matched to SSAS patient records, and all perfect matches were for unique patients. When mismatch on SES proxy was allowed, 197 perfect matches were manually found between the SSAS school and patient records. Of these, 188 unique school records were perfectly matched to SSAS patient records and 193 SSAS patients were perfectly matched to SSAS school records.

5.5.2 Probabilistic record linkage

Blocking on sex and date of birth, resulted in 293 blocked pairs. Probabilistic matching on sex, date of birth, commune of residence and the proxies for SES with selection of possible matches to create a bijective set of matches resulted in 233 matches of unique SSAS school and patient records. This corresponds to 47.65% of the school records for students with autism in SSAS having a match in the SSAS patient records, 16.93% of the patient records having a match in the SSAS school records and 17.07% of the unique patients having a match in the SSAS school records. For each patient that had lived in more than one commune and therefore appeared more than once in the patient data, only one match to an SSAS school record was made, meaning the matching was bijective for SSAS school records and unique patients.

Analysis of differences between matched and unmatched records in the SSAS school data and in the patient data by sex, commune and proxy for SES are provided in the Supplementary Figures. Kolmogorov-Smirnov permutation tests found no significant difference in frequency of sexes between matched (12.88% female) and unmatched SSAS school records (12.16% female), see Supplementary Figure 49. They found a strongly significant difference in the frequency of sexes between matched (12.88% female) and unmatched patient records (21.52% female), see Supplementary Figure 50. This difference was likely due to the male to female ratios differing across the datasets: the SSAS school data of students with autsim was 12.47% female, the

patient data was 20.06% female and the matches were 12.50% females. Permutation testing found that for the SSAS school data, matched (39.91% resident in Temuco) and unmatched records (56.08% resident in Temuco) differed significantly by commune, see Figure 51, and that there was no significant difference in the patient data by commune between matched (40.34% resident in Temuco) and unmatched records (40.16% resident in Temuco), see Figure 51. This appeared to be driven by the matchability of students and patients living in Temuco, the most prevalent commune. For the SSAS school data by proxy SES there was a strongly significant difference between matched (90.56% status of 1, 5.58% status of 2, 0.00% status of 3 and 3.86% with unknown status) and unmatched (80.78% status of 1, 13.73% status of 2, 0.39% status of 3 and 5.10% with unknown status) records, see Figure 53. For the patient data there was a somewhat significant difference between matched (38.2% status of 1, 58.37% status of 2, and 3.43% status of 3) and unmatched (42.69% status of 1, 55.64% status of 2 and 1.66% status of 3) records, see Figure 54. Again this likely reflected different frequencies of the SES values across datasets. Kolmogorov-Smirnov permutation testing was not conducted for date of birth as this feature contained too many categories for results to be meaningful.

5.6 Updated autism prevalence estimates and delta

Table 15: Age- and sex-adjusted updated autism prevalence from data linkage in SSAS by age band with 95% gamma confidence intervals.

Age band	Crude prevalence (95% CI)	Adjusted prevalence (95% CI)
6-8	1.54 (1.40, 1.67)	1.54 (1.41, 1.68)
9-11	1.34 (1.21, 1.46)	1.33 (1.21, 1.46)
12-14	1.08 (0.97, 1.19)	1.08 (0.97, 1.20)
15-18	0.96 (0.86, 1.07)	0.98 (0.87, 1.11)

After linking the school and patient data for SSAS, 1,132 patients with autism could not be matched to students. This represents the unmet need of SSAS students with autism that did not access school-based support for it. Combining these additional cases with the 488 students that did access SEED for autism gave 1,620 people with autism in SSAS. Thus the crude updated prevalence of autism in SSAS was 1.23% (1.17-1.28%) and the age- and sex-adjusted updated prevalence of autism was 1.22% (1.16-1.28%). The sample size here was 132,242, the same as the number of students resident in SSAS, because no individuals were added, only non-cases in the school data reallocated as cases based on the patient data linkage findings. For females, the adjusted updated prevalence was 0.47% (0.41-0.53%) and for males it was 1.95% (1.84-2.06%). This gave an updated male to female ratio of 4.18 after data linkage, smaller than the ratio of 7.00:1 in the SSAS school data before linkage. Updated autism prevalence was highest among individuals aged 6-8 at 1.54% (1.41-1.68%) and decreased with age, see Table 15.

Table 16: Adjusted prevalence and adjusted updated prevalence of autism by health service in Chile. Adjusted prevalence is from school data only. Adjusted updated prevalence is from linkage of school data and patient data. Prevalence for Servicio de Salud Araucanía Sur (SSAS) was calculated directly from linkage results. Prevalence for other health services was calcuated by adding the adjusted prevalence delta to each health service's adjusted prevalence from the school data only. Adjusted prevalence has 95% gamma confidence intervals. The width of the adjusted updated prevalence confidence intervals is the maximum of the school data adjusted prevalence confidence intervals for each health serce, and the adjusted prevalence delta confidence interval, except for SSAS which has the 95% gamma confidence intervals found earlier.

Health service	Adjusted prevalence (95% CI)	Adjusted updated prevalence (Maximal 95% CI)
Aconcagua	0.43 (0.37, 0.50)	1.28 (1.21, 1.34)
Aisén	0.75 (0.63, 0.90)	1.60 (1.47, 1.73)
Antofagasta	0.83 (0.77, 0.88)	1.67 (1.61, 1.74)
Araucanía Norte	0.30 (0.24, 0.38)	1.15 (1.08, 1.21)
Araucanía Sur	0.37 (0.34, 0.41)	1.22 (1.16, 1.28)
Arauco	0.72 (0.62, 0.82)	1.56 (1.46, 1.66)
Arica	0.61 (0.54, 0.70)	1.46 (1.38, 1.54)
Atacama	0.31 (0.27, 0.37)	1.16 (1.10, 1.22)
Biobío	0.42 (0.37, 0.47)	1.27 (1.20, 1.33)
Chiloé	0.43 (0.36, 0.52)	1.28 (1.20, 1.36)
Concepción	0.77 (0.72, 0.83)	1.62 (1.56, 1.68)
Coquimbo	0.40 (0.36, 0.43)	1.24 (1.18, 1.31)
Iquique	0.43 (0.38, 0.49)	1.28 (1.22, 1.34)
Magallanes	0.83 (0.72, 0.96)	1.68 (1.56, 1.80)
Maule	0.30 (0.28, 0.33)	1.15 (1.09, 1.21)
Metro. Central	0.42 (0.38, 0.46)	1.26 (1.20, 1.33)
Metro. Norte	0.29 (0.26, 0.31)	1.13 (1.07, 1.20)
Metro. Occidente	0.34 (0.32, 0.36)	1.19 (1.12, 1.25)
Metro. Oriente	0.30 (0.27, 0.33)	1.15 (1.08, 1.21)
Metro. Sur	0.40 (0.37, 0.43)	1.25 (1.18, 1.31)
Metro. Sur Oriente	0.36 (0.34, 0.39)	1.21 (1.15, 1.27)
O'Higgins	0.42 (0.39, 0.46)	1.27 (1.21, 1.34)
Osorno	0.43 (0.37, 0.51)	1.28 (1.21, 1.35)
Reloncaví	0.42 (0.37, 0.47)	1.26 (1.20, 1.33)
Talcahuano	0.81 (0.74, 0.90)	1.66 (1.58, 1.74)
Valdivia	0.30 (0.26, 0.35)	1.15 (1.08, 1.21)
Valparaíso	0.68 (0.62, 0.74)	1.52 (1.46, 1.59)
Viña del Mar	0.66 (0.62, 0.70)	1.51 (1.44, 1.57)
Ñuble	1.29 (1.21, 1.37)	2.13 (2.05, 2.21)

The adjusted prevalence delta for SSAS was 0.85% (0.78, 0.91%) and represents the unmet need of children with autism in SSAS that did not access SEED for it. Table 16 shows the projection of adjusted updated prevalence from the data linkage for SSAS onto the other health services. The patterns of prevalence across health services were retained and Ñuble has the highest adjusted updated prevalence at 2.13% (2.05, 2.21%). Adding the adjusted prevalence delta to the national adjusted autism prevalence found from the school data alone gave national adjusted updated autism prevalence of 1.31%. This in turn gave a posited 40,113 children aged 6-18 with autism in Chile and corresponds to an estimated unmet need of 25,903 children that did not access SEED for autism (posited count and unmet need are age- and sex-adjusted).

5.7 Bayesian prevalence projection

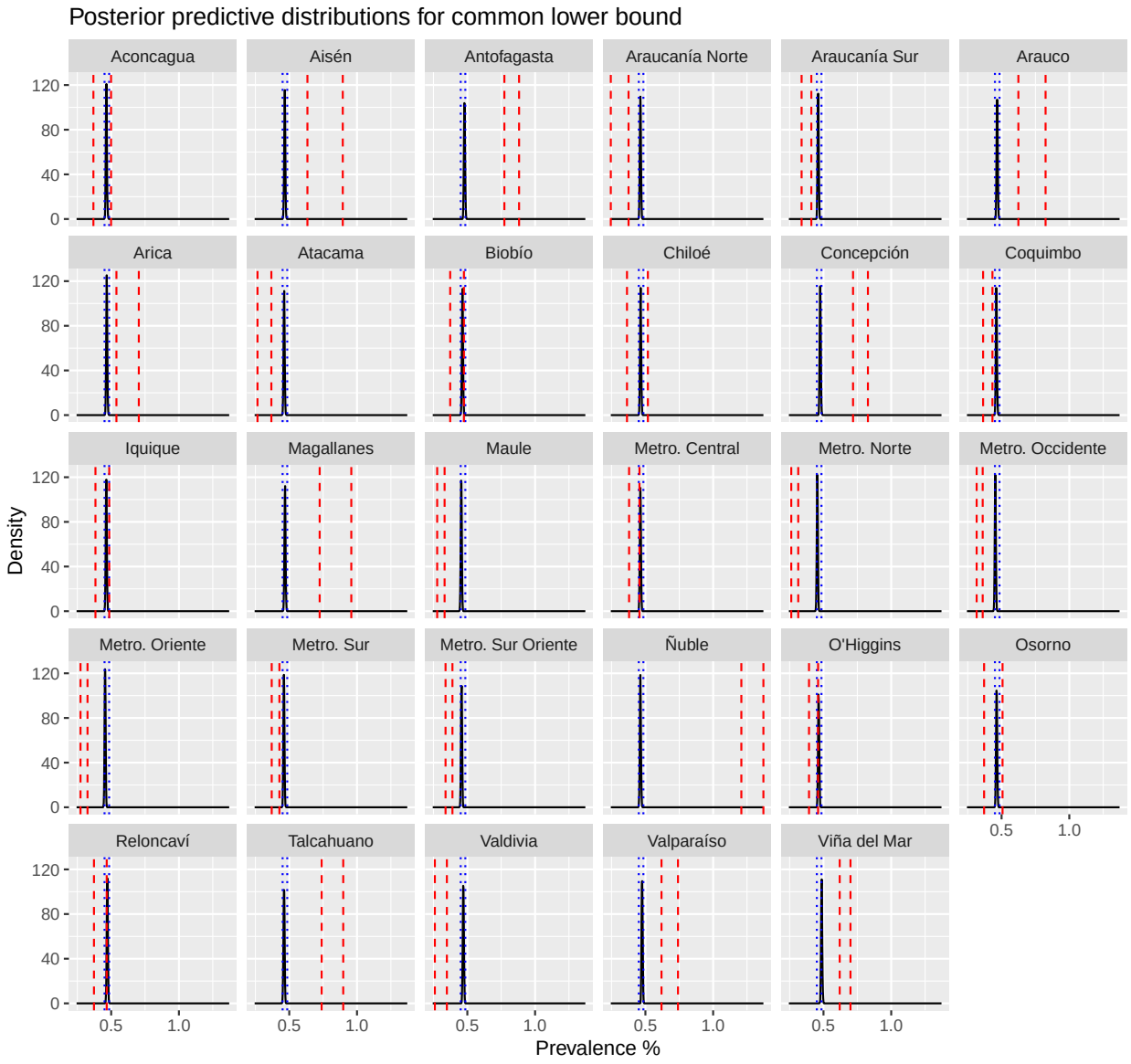


Figure 37: Posterior predictive distributions for autism prevalence using adjusted case counts from the school data with a random effect on student's health service. Modeling used a beta conjugate prior of age- and sex-adjusted national autism prevalence from school data. Red dashed lines show the adjusted sample prevalence 95% gamma confidence intervals and blue dotted lines show the posterior 95% credible interval.

Bayesian prevalence projections by health service with common national autism prevalence prior are shown in Figure 37. The differences in adjusted sample prevalence across health services are evident in the red bands and the posterior predictive distributions have been pulled towards the common prior. For regions like Ñuble which have adjusted sample prevalence in the school data higher than the national adjusted prevalence, these posterior densities are not plausible.

Bayesian prevalence projections by health service with health service specific priors are shown in Figure 38. As expected, the posterior 95% credible intervals were coincident with the adjusted sample prevalence 95% confidence intervals. The posterior prevalence peaks can be considered lower bounds for the true autism prevalence in each health service.

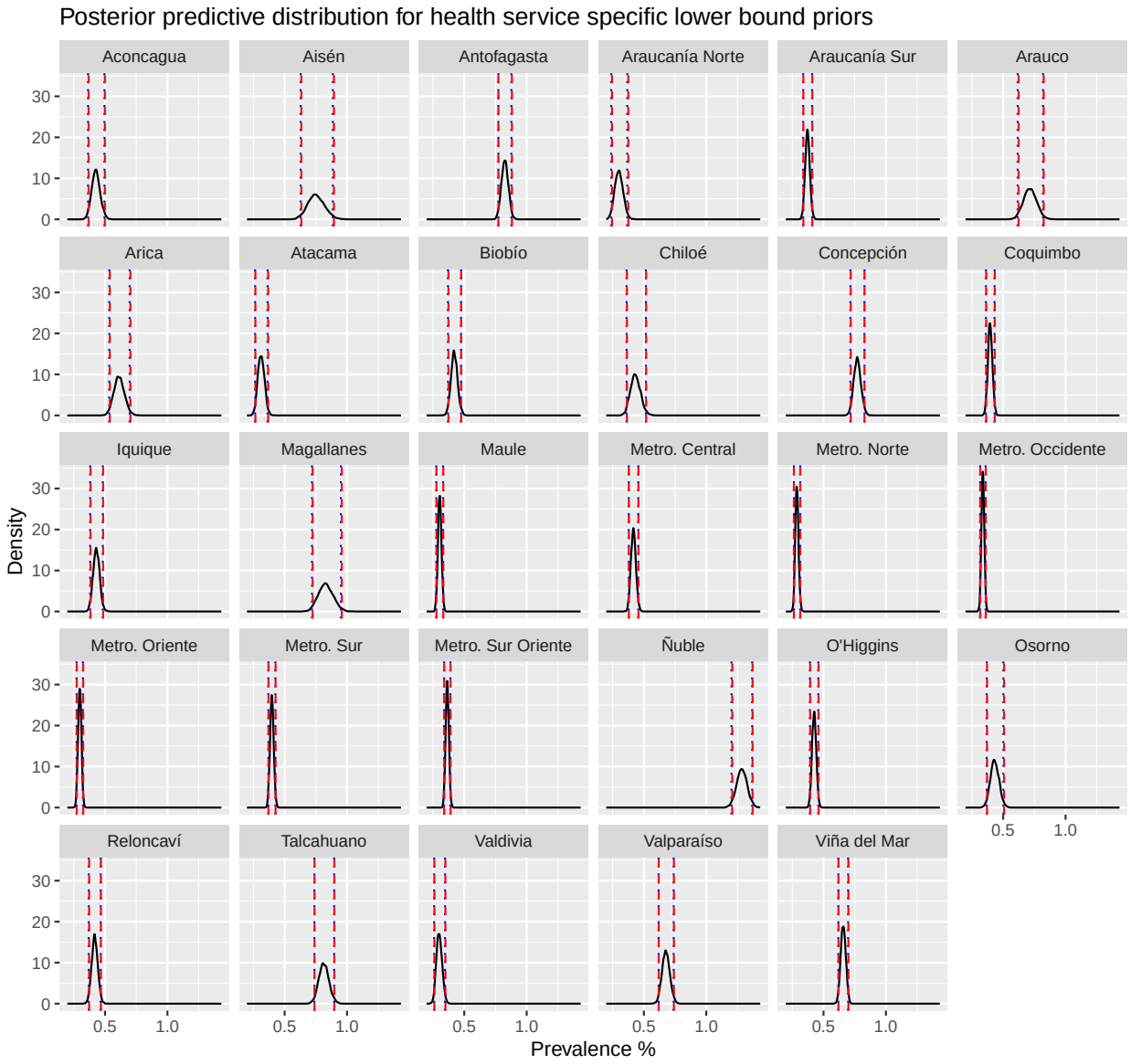


Figure 38: Posterior predictive distributions for autism prevalence using adjusted case counts from the school data with a random effect on student's health service. Modelling used beta conjugate priors of health service specific age- and sex-adjusted autism prevalence from school data. Red dashed lines show the adjusted sample prevalence 95% gamma confidence intervals and blue dotted lines show the posterior 95% credible interval.

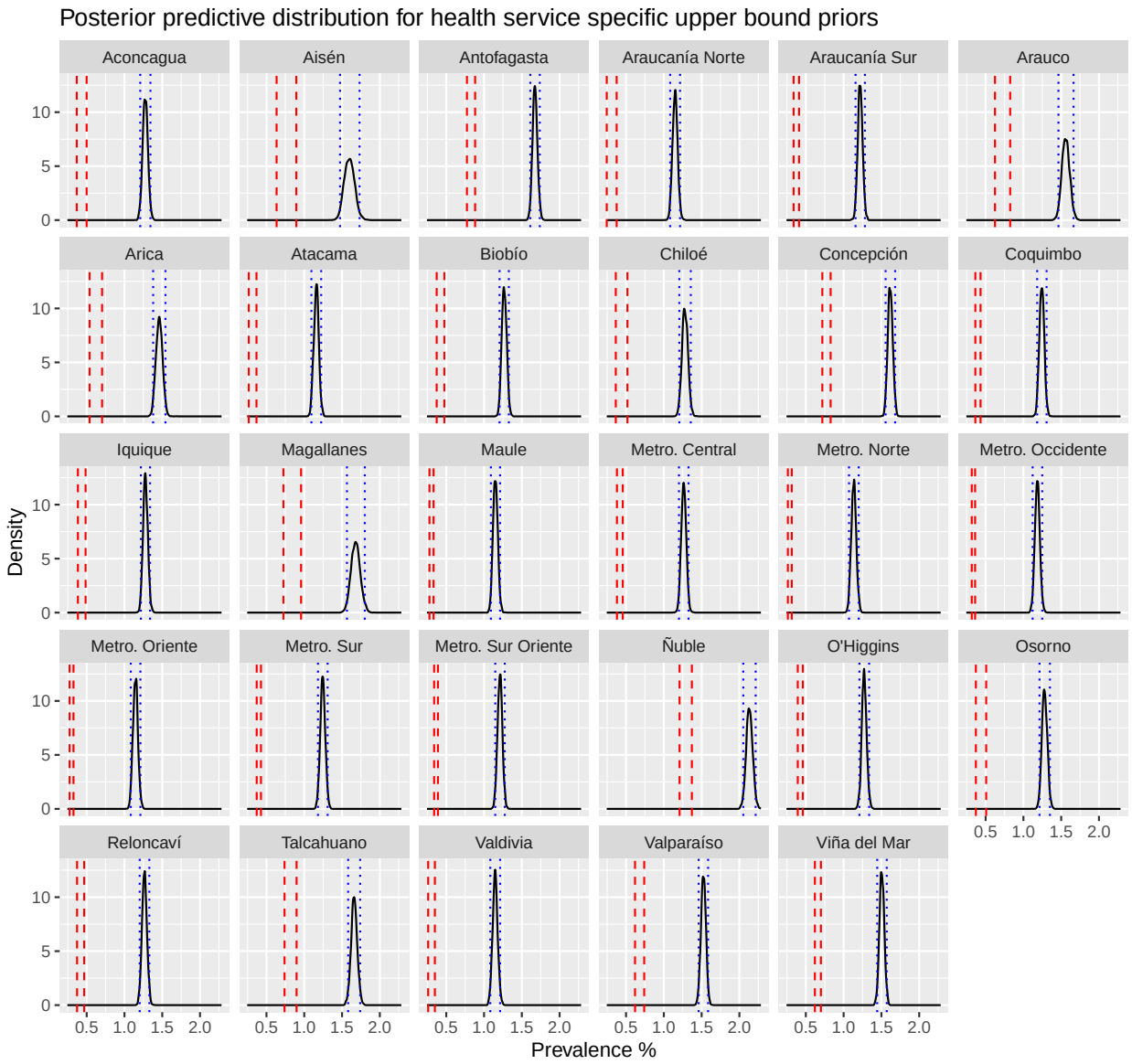


Figure 39: Posterior predictive distributions for autism prevalence using adjusted case counts from the school data with a random effect on student's health service. Modelling used beta conjugate priors of health service specific age- and sex-adjusted updated autism prevalence from data linkage. Red dashed lines show the adjusted sample prevalence 95% gamma confidence intervals and blue dotted lines show the posterior 95% credible interval.

With projected prevalence priors, the Bayesian prevalence projections shown in Figure 39 were pulled upward toward their priors by approximately delta. For regions with high sample prevalence such as Ñuble, this addition of delta resulted in posterior prevalence projections up to 1.5 percentage points higher than the national adjusted prevalence for the school data only which may not be plausible.

In Figure 40, the posterior distributions were reflective of their uniform priors with posterior credible intervals slightly within the prior bounds (Table 16). These predictive densities show a considerable departure from the adjusted sample prevalence from the school data and provide a window within which the true prevalence of recorded autism in each health service would plausibly fall.

6 Discussion

6.1 Findings

6.1.1 Aim 1: Use school data and frequentist method to find a lower bound on autism and ADHD prevalence, and assess socio-demographic variation

In Chilean 6–18-year-old students, a lower bound on autism prevalence is 0.46% (0.46- 0.47%). This is higher than estimates for Venezuela of 0.17% and Ecuador of 0.11-0.21%, but lower than the estimate for Argentina of 1.03% and lower than Yáñez et al's estimate for Chile of 1.95% (24,26–28). The observed crude prevalence of autism is highest in young students and decreases with age, and this pattern is preserved across demographic features. Yeargin-Allsop et al. also found autism prevalence decreased after age 8 and suggest changes to diagnostic criteria as a possible cause (65). Assuming true autism prevalence increases with age as individuals can be diagnosed well into adulthood, this result shows that it is easier for young children to access SEED funding in Chile. This could be the result of improved awareness of the SEED programme and policy changes in Chile to increase accessibility and participation for students with autism (14), and could also indicate that students with autism need fewer schooling interventions as they mature.

A lower bound on ADHD prevalence in Chile is 1.50% (1.48- 1.51%) which is much lower than de la Barra et al. and Vicente et al.'s prevalence values of 10.0% and 12.6% respectively (41,42). This is most likely due to the data providing at most one special needs code per student, the high rate of co-diagnosis of ADHD and autism, and autism being more likely to be recorded than ADHD in the case of co-diagnosis (38). More worked focused on ADHD will be needed to quantify how far this lower bound is from the true ADHD prevalence in Chile. The crude prevalence of ADHD in Chilean students is highest for early teenagers then decreases somewhat and this pattern is preserved across demographic features. This suggests ADHD is diagnosed later than autism which is somewhat consistent with others' work including Sainsbury et al. who note that autism is typically diagnosed at age 5 and ADHD at age 6 (66). The observed decrease in ADHD prevalence in later teenage years could indicate reduced ability to access schooling interventions with age, or reduced need for them.

This investigation appears to show autism and ADHD prevalence is very low in students with the highest socio-economic status, however this is most likely due to data artefacts as privately educated students are ineligible for SEED funding and therefore will not appear has having either condition in this data. No other clear differences in the prevalence of autism or of ADHD by socio-economic status have been found. For ADHD, this is consistent with de la Barra et al.'s findings in Chile (41).

Prevalence of autism and ADHD is lower among students of the Mapuche Indigenous group than among non-Indigenous students. For ADHD, prevalence is even lower among member of other Indigenous groups. For autism, prevalence in other Indigenous groups is lower than for non-Indigenous groups but higher than for Mapuche. This agrees with Lindblom's findings that Indigenous children with autism in Canada's British Colombia are underrepresented in school data (67). However it is important to note here that people Indigenous to different regions are not necessarily comparable and very little research has been done on autism and ADHD in Indigenous people in Latin America.

Autism and ADHD prevalence is higher in students at rural schools that urban schools which is somewhat surprising as urban areas generally have more resources to diagnose and provide adjustments for these

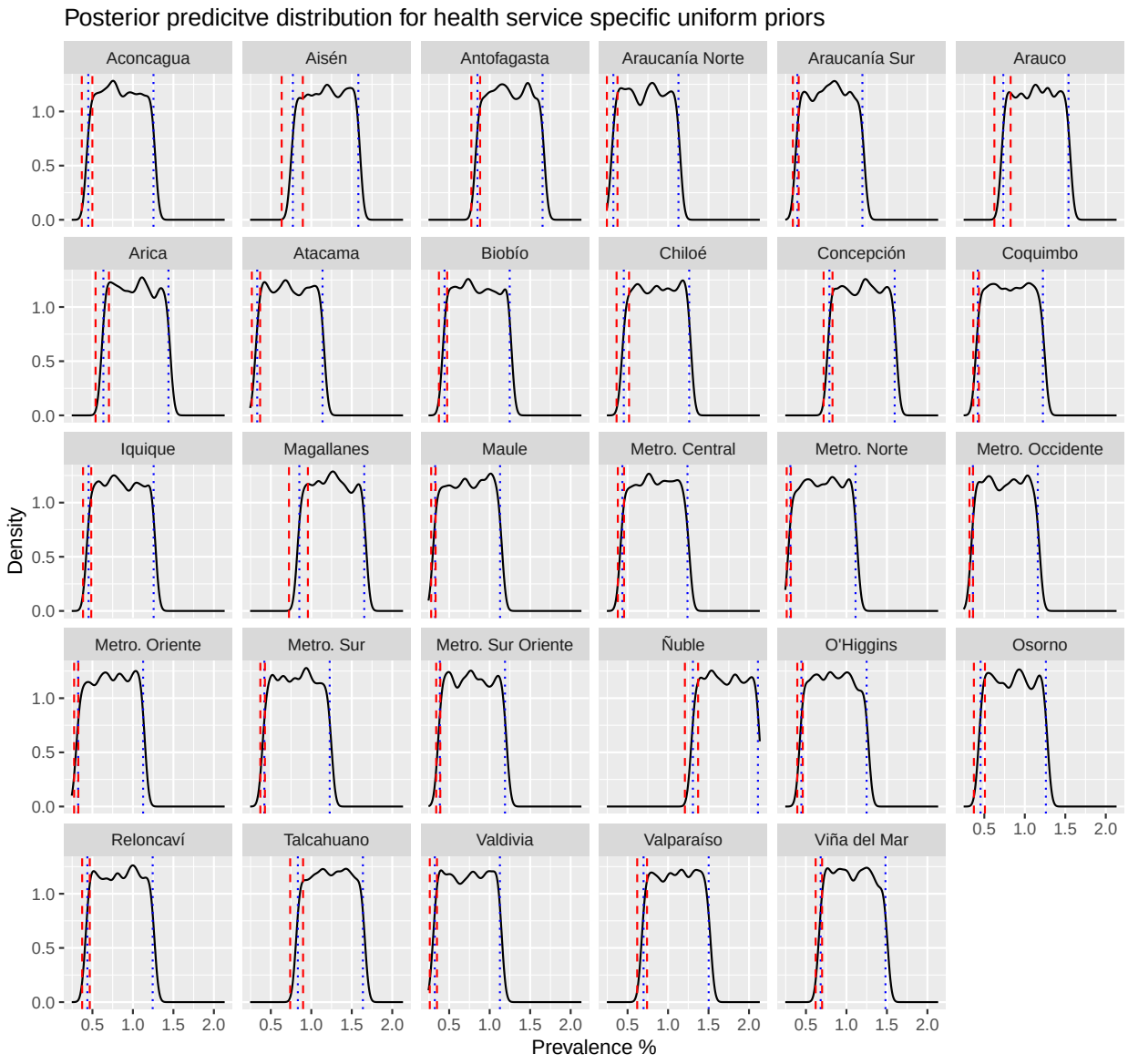


Figure 40: Posterior predictive distributions for autism prevalence using adjusted case counts from the school data with a random effect on student's health service. Modelling used uniform priors bounded below by health service specific age- and sex-adjusted autism prevalence from school data, and bounded above by health service specific age- and sex-adjusted updated autism prevalence from data linkage. Red dashed lines show the adjusted sample prevalence 95% gamma confidence intervals and blue dotted lines show the posterior 95% credible interval.

conditions, but may be the result of successful implementation of policy interventions described by Núñez and Manzano that aimed to improve healthcare for disadvantaged areas and population groups (6).

6.1.2 Aim 2: Use clinical data and machine learning to identify autism diagnosis characteristics

MCA revealed that age, ethnicity and commune of residence, and to some extent other disability status, foster care status and health insurance level, are important features for understanding the distribution of autism diagnosis. The clustering of patients based on whether information was available for other disability and foster care status shows that missingness in demographic data is important when understanding autism diagnoses. The variance explained by the available features is not large enough to draw strong conclusions about possible composite components associated with autism diagnosis, but it does indicate that further investigation with larger dataset and more demographic characteristics are likely to lead to interesting results.

6.1.3 Aim 3a: Use machine learning to link school and clinical records

Chi squared testing showed that students living in SSAS are different to students in other communes by age band, SES, ethnicity, rurality and special needs status. For the most part this is not surprising as La Araucanía region is known to be poor, rural and have a large Mapuche population (10,13,68). SSAS has a somewhat older student population than other regions and significantly fewer students access SEED funding for any condition, meaning the adjusted autism prevalence estimate for SSAS from the school data only of 0.37% is likely to be an underestimate.

Both manual and probabilistic linkage of SSAS school and patient data of individuals with autism were able to identify a reasonable number of matches. Probabilistic matching was more effective than manual matching here because matching on commune of residence was desirable but not essential, proxy SES features were known to be imprecise so requiring perfect matches would miss many correct matches, and probabilistic matching was able to identify more likely matches.

The proportions of matched and unmatched records in the SSAS school and patient datasets did differ by sex, commune of residence and proxy SES. The differences appear to be driven by differences in the frequency of these features' categories across the two datasets.

6.1.4 Aim 3b: Accurately estimate autism prevalence and unmet need, project estimates across health services using Bayesian prevalence prediction

The updated estimate for the prevalence of autism in SSAS is 1.22% (1.16 - 1.28%). It is likely to be an underestimate due to not having private sector diagnoses in the clinical data. Projected nationally, the estimated prevalence of autism in Chile is 1.31% (1.25 - 1.38). This estimate is somewhat smaller than Yáñez et al's estimate of 1.96% (0.81-4.63) in Santiago communes but well within their estimate's wide confidence interval (24). It is also lower than Roman-Urrestarazu et al.'s estimate of 1.76% (1.75%-1.77%) from similar school records for 120,000 English children (30). This suggests either that the data linkage conducted here has missed some autism cases, or that autism prevalence is lower or diagnosis less available in Chile than in England.

Using the updated prevalence estimate for SSAS, the male to female ratio is 4.18:1. This is consistent with the 4:1 ratio found by Yáñez et al (24), reasonably consistent with Roman-Urrestarazu et al.'s finding of a 4.32:1 ratio in England (30) and higher than the 3:1 ratio found by Loomes, Hull and Mandy though meta-analysis of studies in North America, Europe, East Asia, the Middle East, Australia, Aruba and Venezuela (31). This updated ratio is close to the ratio of 3.99:1 observed here in the patient data alone. It is much lower than the ratio for the whole school dataset of 6.00:1 which was itself lower than the 7.00:1 ratio observed for the SSAS school data before linkage. Taken together, this suggests females with autism are underrepresented in the school data and particularly underrepresented in the SSAS subset of the school data, which could indicate that there are barriers to females accessing SEED for autism.

Projecting the adjusted prevalence delta across all regions resulted in updated estimates for all health services that are comparable to Yáñez et al's estimate (24). Ñuble's standard adjusted prevalence estimate is considerably higher than for other communes which may indicate that Ñuble uses a different approach for

autism SEED funding assessment and this figure may be more indicative of Chile's true autism prevalence. Using the same adjusted prevalence delta for all regions may not have been appropriate given the regions differ considerably in prevalence observed from the school data. Additionally, SSAS is in the bottom third of adjusted prevalence estimates, meaning it is likely to have a larger difference between adjusted prevalence from the school data and true prevalence. If we can assume the adjusted updated prevalence for SSAS is the true autism prevalence of Chile then the delta to reach this will be larger for SSAS than for regions with higher original adjusted prevalence from the school data. Therefore using SSAS's delta in higher prevalence regions will produce overestimates. Ñuble's adjusted updated prevalence is 2.13% which is possible but there is a phenotypic upper bound on true autism prevalence and it is more likely that a larger proportion of people with autism in Ñuble are accessing SEED than in other regions, or that many students in Ñuble have been misdiagnosed with autism.

This investigation found an unmet need of 1,132 students in SSAS that did not access SEED for their autism in 2021. It also projected an unmet need of 25,903 students across Chile. This unmet need is likely an overestimate as it includes students with an autism diagnosis at non-subsidised private schools that were not eligible for SEED but may have received other school-based support, students that did receive SEED but for another condition, and students that did not need or did not want school-based interventions. However it may also underestimate the number of students with autism that did not access SEED because it does not include students diagnosed with autism in the private health sector. The accuracy of this estimate of unmet need relies on the assumption that the difference between prevalence values observed for SSAS in the school data with and without supplementation with clinical data is applicable to the national prevalence which may not be the case given we have shown above that the student population of SSAS does differ from other health services and is not nationally representative. Measured over time, the unmet need estimate could be used to assess the effectiveness of PIE interventions to integrate students with special needs in public schools (17).

The Bayes prevalence projection was able to circumvent issues with the applicability of the adjusted prevalence delta by modelling prevalence within plausible bounds provided by the school and clinical dataset. The theoretical approach presented by Ince et al. (46) was attempted and found to be ineffective here – the low prevalence of autism in these data caused calculation of the exact posterior predictive distribution to very quickly reach the limit of R's computational precision. Markov chain Monte Carlo sampling overcame this impediment and is demonstrably a more effective approach than Ince et al.'s (46) for prevalence estimation of low frequency conditions in large epidemiological datasets. Using uniform priors specific to the health services that spanned from the lower bounds found through frequentist methods to the possibly overestimated adjusted updated prevalence values found through linkage, the Bayesian modelling found that the true prevalence of autism in each health service is likely to be considerably higher than the prevalence observed for each using the school data alone. This analysis has determined a plausible range within which the true prevalence for each health service catchment area is quantifiably likely to fall.

6.2 Limitations

This investigation is limited by the quality of available data. Firstly, the school data is limited by having only one special needs code available per student as it is based on their SEED status and not all their extant special needs diagnoses. This limitation prompted the use of clinical data to supplement the available diagnoses, however these data are also somewhat limited as validated diagnoses and demographic information were only available for some patients. Additionally, it only includes patients treated in the public health system and the number of patients with autism that are treated privately is unknown. Autism prevalence estimates using clinical data may therefore be underestimates.

During MCA, the other disability and foster care status features were very unbalanced with few patients having a recorded disability and even fewer having experienced foster care. For both features, nearly two thirds of records had no information available and while reasonable imputation was used, any imputation adds bias to the results and here the bias is likely to be considerable as so many values were imputed. This means the MCA results should be taken as indicative and more research done on larger, more complete dataset to confirm them.

Linkage of school and patient data for SSAS used matching on proxy SES. In the school data, SES was

mapped from school fees paid and in the patient data it was mapped from health insurance contributions. These underlying features may not be comparable and the mapping to SES used for each was imprecise. To account for this, a high error rate was declared when allocating weights during probabilistic linkage, however some incorrect matches based on SES may have been made or correct matches been missed. Linkage also assumed that individuals did not move into or out of SSAS communes between collection of the school and patient data. This is unlikely to be true and a small number of correct matches are likely to have been missed due to individuals who moved not being included in one of the SSAS school data or the patient data being linked. Missed correct matches would cause the updated prevalence estimate for SSAS to be slightly too high and therefore projected updated estimates to also be slightly too high.

The Bayesian modelling did use information from the sample data, the full school dataset, twice during modelling, to varying extents for each prior. For all four analyses, random effects models were fit to health services. For the second prior, which used health service specific autism prevalence, this meant the sample data was directly used twice which generally would not be informative but here was valuable to evidence the lower bound on the true prevalence values. For the first prior, the sample data was used indirectly to calculate the common national prevalence and was valuable to demonstrate that the prevalence values observed from the school data can only be underestimates of the true prevalence. For the third prior, using the updated prevalence values found through data linkage, the sample data was used indirectly to give the prevalence values to which delta was added, and the fourth prior combined the second and third. Although it is usually not appropriate to use sample data twice in Bayesian modelling, it was considered acceptable here as the intention was to conduct exploratory analysis to project the bounds on prevalence rather than predict the true autism prevalence across the other health services.

6.3 Extensions

While this investigation has progressed the understanding of autism and ADHD prevalence in Chile, it has also uncovered new avenues for exploration. If clinical data on ADHD diagnoses were available for a region of Chile, the MCA analysis, data linkage, updated prevalence and unmet need estimation, and Bayesian projection conducted here for autism could be extended to ADHD. If clinical data on autism for other health service regions were available, the delta calculation could be validated and thus more accurate autism prevalence estimates could be found.

This investigation has considered autism and ADHD in isolation from each other. However, given 59.1% of people with autism have a co-diagnosis of ADHD (40), it would be valuable to consider these conditions together. MCA of autism patient data could be used to assess contributions from ADHD and other co-diagnoses such as depression and anxiety, physical disability and intellectual disability.

This investigation used MCA to assess the contribution of feature categories among patients with autism in SSAS which required casting the age feature to categorical age bands. However age is a continuous variable and the information contained in it would be better captured using a continuous variable analysis technique such as principle component analysis. The commune feature would need to be excluded because it cannot be meaningfully encoded as a continuous or ordered categorical variable. The remaining features could be one-hot encoded to pseudo-continuous variables before performing PCA. This would allow the contribution of patient age to be better characterised but one-hot encoding would somewhat reduce the power of inferences about the other features used. Cluster-based analysis of a larger dataset containing patients with and without autism would also be valuable and would allow further identification of the characteristics associated with autism diagnosis.

During data linkage, this investigation assumed that unmatched patients in the clinical data existed in the school data but did not have an autism diagnosis in the school data. It would be valuable to test this assumption by attempting to link the patient data for SSAS to all school records for SSAS, including those students that did not have an autism diagnosis. This would allow the number of patients that did not exist in the school data to be estimated which would provide a more accurate sample size for use as the denominator in prevalence calculations.

This investigation assumed the adjusted prevalence delta found for SSAS was directly applicable to other regions which may not be the case. Further analysis using Bayesian prevalence projection could use health

service specific deltas that are inversely proportional to the sample prevalence observed for each health service. This would decrease the delta for services with high observed prevalence and thus lessen the increase in the updated prevalence estimate, and would increase the updated prevalence estimate for health services with low observed prevalence. This would provide more biologically plausible prevalence estimates, however such Bayesian analysis would still be using information from the school data twice – as the sample data and to inform the health service specific prior through use of observed prevalence to scale delta.

Future work could replicate the prevalence estimation conducted here using the UK Department for Education's National Pupil Database which is similar to the school data used here. This would provide an interesting comparison of autism and ADHD prevalence in two large countries with different population structures and ethnic profiles.

7 Conclusions

This investigation has furthered the understanding of autism and ADHD prevalence in Chile. Autism prevalence is at least 0.46% (0.13% in females,0.79% in males and male to female ratio of 6.00:1) and ADHD prevalence is at least 1.5% (1.01% in females, 1.97% in males and male to female ratio of 1.94:1). Both autism and ADHD are more prevalent in non-Indigenous students and students at rural schools, and neither have a clear difference in prevalence by SES. In Servicio de Salud Araucanía Sur, age, ethnicity and commune of residence are important components for explaining the variance in data on patients with autism. School data and patient data on individuals with autism in SSAS were successfully linked. From linkage, a more accurate estimate of the prevalence of autism in this region is 1.22% (male to female ratio of 4.18:1) and the estimated prevalence in Chile is 1.31%. The unmet need of students with autism that did not receive SEED funding for support with it is 1,132 students in SSAS and a posited 25,903across Chile. Linkage also revealed females are likely underrepresented in school autism diagnoses. Bayesian prevalence projection extended prevalence projections across health services and found plausible prevalence bounds for each health service.

8 Supplementary materials

Table 17: Count and prevalence of autism cases by age band in Chile school data with normal confidence intervals.

Age band	Autism cases	Prevalence % (95% CI)
6-8	5162	0.69 (0.67, 0.71)
9-11	4212	0.55 (0.53, 0.57)
12-14	3038	0.41 (0.39, 0.42)
15-18	2137	0.27 (0.26, 0.28)

Table 18: Count and prevalence of ADHD cases by age band in Chile school data with normal confidence intervals.

Age band	ADHD cases	Prevalence % (95% CI)
6-8	5936	0.79 (0.77, 0.81)
9-11	15549	2.03 (1.99, 2.06)
12-14	14099	1.88 (1.85, 1.91)
15-18	10640	1.35 (1.32, 1.37)

Table 19: Count and prevalence of autism cases by age band in Chile school data for females and males with normal confidence intervals.

	Female		Male	
Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
6-8	774	0.21 (0.20, 0.23)	4388	1.15 (1.11, 1.18)
9-11	523	0.14 (0.13, 0.15)	3689	0.94 (0.91, 0.97)
12-14	391	0.11 (0.10, 0.12)	2647	0.69 (0.66, 0.72)
15-18	290	0.08 (0.07, 0.08)	1847	0.45 (0.43, 0.47)

Table 20: Count and prevalence of ADHD cases by age band in Chile school data for females and males with normal confidence intervals.

	Female		Male	
Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
6-8	774	0.21 (0.20, 0.23)	3944	1.03 (1.00, 1.06)
9-11	523	0.14 (0.13, 0.15)	10322	2.62 (2.57, 2.67)
12-14	391	0.11 (0.10, 0.12)	9714	2.53 (2.48, 2.58)
15-18	290	0.08 (0.07, 0.08)	7165	1.75 (1.71, 1.79)

Table 21: Count and prevalence of autism cases by health service and age band in Chile school data for females and males with normal confidence intervals.

		Female		Male	
Health service	Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
Aconcagua	6-8	<20	0.25 (0.12, 0.38)	59	0.99 (0.74, 1.24)
Aisén	6-8	<20	0.40 (0.14, 0.66)	52	2.29 (1.67, 2.90)
Antofagasta	6-8	54	0.39 (0.29, 0.50)	302	2.09 (1.85, 2.32)
Araucanía Norte	6-8	<20	0.07 (0.00, 0.15)	32	0.69 (0.45, 0.93)
Araucanía Sur	6-8	26	0.17 (0.10, 0.23)	139	0.86 (0.72, 1.00)
Arauco	6-8	<20	0.29 (0.12, 0.46)	67	1.79 (1.36, 2.21)
Arica	6-8	<20	0.34 (0.18, 0.49)	94	1.63 (1.31, 1.96)
Atacama	6-8	<20	0.26 (0.14, 0.39)	53	0.75 (0.55, 0.95)
Biobío	6-8	<20	0.14 (0.06, 0.23)	92	1.06 (0.85, 1.28)
Chiloé	6-8	<20	0.06 (0.00, 0.14)	38	1.07 (0.73, 1.41)
Concepción	6-8	44	0.34 (0.24, 0.43)	252	1.87 (1.64, 2.10)
Coquimbo	6-8	36	0.21 (0.14, 0.28)	207	1.14 (0.98, 1.29)
Iquique	6-8	<20	0.18 (0.09, 0.27)	100	1.11 (0.89, 1.33)
Magallanes	6-8	<20	0.19 (0.04, 0.35)	42	1.30 (0.91, 1.70)
Maule	6-8	41	0.19 (0.13, 0.24)	193	0.84 (0.72, 0.96)
Metro. Central	6-8	22	0.16 (0.09, 0.22)	162	1.07 (0.91, 1.23)
Metro. Norte	6-8	26	0.12 (0.07, 0.17)	174	0.77 (0.65, 0.88)
Metro. Occidente	6-8	62	0.16 (0.12, 0.20)	323	0.81 (0.72, 0.90)
Metro. Oriente	6-8	31	0.14 (0.09, 0.19)	154	0.69 (0.58, 0.80)
Metro. Sur	6-8	44	0.18 (0.13, 0.24)	261	1.04 (0.91, 1.16)
Metro. Sur Oriente	6-8	41	0.15 (0.10, 0.20)	277	0.95 (0.84, 1.07)
O'Higgins	6-8	38	0.20 (0.14, 0.26)	215	1.08 (0.94, 1.22)
Osorno	6-8	<20	0.20 (0.07, 0.33)	51	1.08 (0.78, 1.37)
Reloncaví	6-8	22	0.24 (0.14, 0.34)	96	1.02 (0.82, 1.23)
Talcahuano	6-8	25	0.40 (0.24, 0.55)	129	1.96 (1.63, 2.30)
Valdivia	6-8	<20	0.15 (0.06, 0.23)	71	0.90 (0.69, 1.10)
Valparaíso	6-8	22	0.23 (0.13, 0.33)	156	1.57 (1.33, 1.82)
Viña del Mar	6-8	55	0.27 (0.20, 0.34)	292	1.37 (1.21, 1.52)

Table 21: Count and prevalence of autism cases by health service and age band in Chile school data for females and males with normal confidence intervals. *(continued)*

		Female		Male	
Health service	Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
Ñuble	6-8	57	0.63 (0.46, 0.79)	305	3.21 (2.85, 3.56)
Aconcagua	9-11	<20	0.05 (0.00, 0.11)	59	0.95 (0.71, 1.20)
Aisén	9-11	<20	0.08 (0.00, 0.20)	26	1.05 (0.65, 1.45)
Antofagasta	9-11	45	0.30 (0.22, 0.39)	260	1.67 (1.47, 1.87)
Araucanía Norte	9-11	<20	0.09 (0.00, 0.18)	26	0.57 (0.35, 0.79)
Araucanía Sur	9-11	<20	0.10 (0.05, 0.15)	120	0.71 (0.58, 0.83)
Arauco	9-11	<20	0.23 (0.08, 0.38)	70	1.73 (1.33, 2.13)
Arica	9-11	<20	0.20 (0.08, 0.31)	57	0.95 (0.71, 1.20)
Atacama	9-11	<20	0.09 (0.02, 0.16)	45	0.59 (0.42, 0.76)
Biobío	9-11	<20	0.19 (0.09, 0.28)	68	0.76 (0.58, 0.94)
Chiloé	9-11	<20	0.06 (0.00, 0.13)	49	1.27 (0.92, 1.63)
Concepción	9-11	30	0.22 (0.14, 0.30)	216	1.49 (1.29, 1.69)
Coquimbo	9-11	<20	0.10 (0.05, 0.14)	145	0.78 (0.65, 0.90)
Iquique	9-11	<20	0.19 (0.09, 0.28)	80	0.86 (0.67, 1.04)
Magallanes	9-11	<20	0.17 (0.03, 0.31)	60	1.73 (1.30, 2.17)
Maule	9-11	<20	0.07 (0.04, 0.11)	140	0.60 (0.50, 0.70)
Metro. Central	9-11	<20	0.11 (0.06, 0.17)	136	0.86 (0.72, 1.00)
Metro. Norte	9-11	<20	0.07 (0.04, 0.11)	147	0.62 (0.52, 0.72)
Metro. Occidente	9-11	28	0.09 (0.05, 0.12)	234	0.67 (0.59, 0.76)
Metro. Oriente	9-11	24	0.11 (0.06, 0.15)	129	0.55 (0.46, 0.65)
Metro. Sur	9-11	38	0.15 (0.10, 0.20)	238	0.90 (0.79, 1.01)
Metro. Sur Oriente	9-11	24	0.08 (0.05, 0.12)	191	0.63 (0.54, 0.72)
O'Higgins	9-11	30	0.15 (0.10, 0.21)	193	0.92 (0.79, 1.05)
Osorno	9-11	<20	0.12 (0.02, 0.22)	54	1.05 (0.77, 1.32)
Reloncaví	9-11	<20	0.09 (0.03, 0.15)	81	0.80 (0.62, 0.97)
Talcahuano	9-11	25	0.38 (0.23, 0.52)	132	1.87 (1.55, 2.18)
Valdivia	9-11	<20	0.05 (0.00, 0.10)	38	0.45 (0.31, 0.59)
Valparaíso	9-11	<20	0.16 (0.08, 0.23)	122	1.23 (1.01, 1.44)
Viña del Mar	9-11	47	0.22 (0.16, 0.28)	294	1.32 (1.17, 1.48)
Ñuble	9-11	39	0.41 (0.28, 0.54)	279	2.82 (2.50, 3.15)
Aconcagua	12-14	<20	0.11 (0.02, 0.20)	38	0.67 (0.45, 0.88)
Aisén	12-14	<20	0.24 (0.05, 0.44)	25	0.94 (0.57, 1.31)
Antofagasta	12-14	28	0.19 (0.12, 0.27)	170	1.09 (0.93, 1.25)
Araucanía Norte	12-14	<20	0.13 (0.03, 0.24)	21	0.44 (0.25, 0.63)
Araucanía Sur	12-14	<20	0.07 (0.03, 0.12)	94	0.57 (0.45, 0.68)
Arauco	12-14	<20	0.14 (0.02, 0.25)	46	1.17 (0.84, 1.51)
Arica	12-14	<20	0.19 (0.07, 0.31)	44	0.80 (0.57, 1.04)
Atacama	12-14	<20	0.04 (0.00, 0.09)	28	0.37 (0.23, 0.51)
Biobío	12-14	<20	0.06 (0.01, 0.11)	58	0.63 (0.47, 0.79)
Chiloé	12-14	<20	0.03 (0.00, 0.08)	32	0.79 (0.52, 1.06)
Concepción	12-14	28	0.22 (0.14, 0.30)	155	1.14 (0.96, 1.32)
Coquimbo	12-14	<20	0.08 (0.04, 0.13)	88	0.51 (0.40, 0.62)
Iquique	12-14	<20	0.10 (0.03, 0.17)	52	0.58 (0.42, 0.74)
Magallanes	12-14	<20	0.20 (0.05, 0.35)	55	1.55 (1.15, 1.96)
Maule	12-14	<20	0.07 (0.04, 0.11)	86	0.37 (0.29, 0.45)
Metro. Central	12-14	<20	0.03 (0.00, 0.06)	96	0.62 (0.49, 0.74)
Metro. Norte	12-14	<20	0.07 (0.03, 0.11)	78	0.35 (0.27, 0.42)
Metro. Occidente	12-14	25	0.08 (0.05, 0.11)	186	0.56 (0.48, 0.64)
Metro. Oriente	12-14	21	0.10 (0.05, 0.14)	122	0.54 (0.45, 0.64)
Metro. Sur	12-14	<20	0.07 (0.04, 0.11)	137	0.54 (0.45, 0.63)
Metro. Sur Oriente	12-14	<20	0.06 (0.03, 0.09)	172	0.57 (0.49, 0.66)
O'Higgins	12-14	21	0.11 (0.06, 0.15)	116	0.57 (0.47, 0.67)
Osorno	12-14	<20	0.12 (0.02, 0.22)	27	0.53 (0.33, 0.72)
Reloncaví	12-14	<20	0.13 (0.06, 0.21)	63	0.61 (0.46, 0.76)
Talcahuano	12-14	<20	0.17 (0.07, 0.26)	87	1.23 (0.97, 1.49)
Valdivia	12-14	<20	0.04 (0.00, 0.08)	43	0.51 (0.36, 0.66)
Valparaíso	12-14	<20	0.17 (0.09, 0.26)	107	1.11 (0.90, 1.32)

Table 21: Count and prevalence of autism cases by health service and age band in Chile school data for females and males with normal confidence intervals. *(continued)*

		Female		Male	
Health service	Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
Viña del Mar	12-14	31	0.15 (0.10, 0.20)	230	1.08 (0.94, 1.22)
Ñuble	12-14	34	0.35 (0.23, 0.46)	191	1.87 (1.61, 2.13)
Aconcagua	15-18	<20	0.07 (0.00, 0.14)	22	0.36 (0.21, 0.51)
Aisén	15-18	<20	0.19 (0.02, 0.36)	24	0.85 (0.51, 1.19)
Antofagasta	15-18	25	0.17 (0.10, 0.23)	120	0.76 (0.62, 0.89)
Araucanía Norte	15-18	<20	0.02 (0.00, 0.06)	<20	0.34 (0.18, 0.51)
Araucanía Sur	15-18	<20	0.04 (0.01, 0.07)	74	0.42 (0.32, 0.51)
Arauco	15-18	<20	0.00 (0.00, 0.00)	21	0.50 (0.29, 0.71)
Arica	15-18	<20	0.07 (0.00, 0.14)	35	0.61 (0.41, 0.81)
Atacama	15-18	<20	0.04 (0.00, 0.09)	23	0.30 (0.18, 0.42)
Biobío	15-18	<20	0.07 (0.01, 0.12)	51	0.52 (0.38, 0.66)
Chiloé	15-18	<20	0.02 (0.00, 0.07)	<20	0.31 (0.15, 0.47)
Concepción	15-18	23	0.17 (0.10, 0.24)	111	0.76 (0.62, 0.91)
Coquimbo	15-18	<20	0.06 (0.02, 0.10)	63	0.35 (0.26, 0.43)
Iquique	15-18	<20	0.02 (0.00, 0.06)	44	0.48 (0.34, 0.62)
Magallanes	15-18	<20	0.38 (0.18, 0.57)	43	1.06 (0.74, 1.37)
Maule	15-18	<20	0.06 (0.03, 0.09)	51	0.21 (0.15, 0.27)
Metro. Central	15-18	<20	0.07 (0.03, 0.11)	67	0.40 (0.30, 0.49)
Metro. Norte	15-18	<20	0.05 (0.02, 0.08)	62	0.26 (0.19, 0.32)
Metro. Occidente	15-18	<20	0.04 (0.02, 0.07)	122	0.35 (0.29, 0.42)
Metro. Oriente	15-18	<20	0.04 (0.01, 0.06)	65	0.26 (0.20, 0.32)
Metro. Sur	15-18	<20	0.04 (0.01, 0.06)	86	0.33 (0.26, 0.40)
Metro. Sur Oriente	15-18	<20	0.05 (0.02, 0.07)	131	0.40 (0.33, 0.47)
O'Higgins	15-18	<20	0.08 (0.04, 0.12)	72	0.33 (0.25, 0.40)
Osorno	15-18	<20	0.04 (0.00, 0.09)	24	0.43 (0.26, 0.60)
Reloncaví	15-18	<20	0.02 (0.00, 0.05)	52	0.47 (0.34, 0.59)
Talcahuano	15-18	<20	0.12 (0.04, 0.19)	42	0.56 (0.39, 0.74)
Valdivia	15-18	<20	0.01 (0.00, 0.04)	35	0.39 (0.26, 0.52)
Valparaíso	15-18	<20	0.11 (0.05, 0.18)	91	0.85 (0.68, 1.02)
Viña del Mar	15-18	35	0.16 (0.11, 0.22)	172	0.73 (0.62, 0.84)
Ñuble	15-18	27	0.26 (0.16, 0.36)	113	1.02 (0.83, 1.21)

Table 22: Count and prevalence of ADHD cases by health service and age band in Chile school data for females and males with normal confidence intervals.

		Female		Male	
Health service	Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
Aconcagua	6-8	56	0.99 (0.74, 1.25)	71	1.19 (0.91, 1.46)
Aisén	6-8	<20	0.54 (0.23, 0.84)	38	1.67 (1.14, 2.20)
Antofagasta	6-8	66	0.48 (0.36, 0.59)	106	0.73 (0.59, 0.87)
Araucanía Norte	6-8	24	0.57 (0.34, 0.79)	52	1.13 (0.82, 1.43)
Araucanía Sur	6-8	99	0.63 (0.51, 0.76)	206	1.27 (1.10, 1.45)
Arauco	6-8	20	0.53 (0.30, 0.76)	49	1.31 (0.94, 1.67)
Arica	6-8	<20	0.32 (0.17, 0.47)	46	0.80 (0.57, 1.03)
Atacama	6-8	<20	0.10 (0.03, 0.18)	20	0.28 (0.16, 0.41)
Biobío	6-8	79	0.95 (0.74, 1.16)	133	1.54 (1.28, 1.79)
Chiloé	6-8	28	0.85 (0.54, 1.16)	66	1.86 (1.42, 2.31)
Concepción	6-8	101	0.77 (0.62, 0.92)	162	1.20 (1.02, 1.38)
Coquimbo	6-8	105	0.61 (0.50, 0.73)	234	1.29 (1.12, 1.45)
Iquique	6-8	35	0.42 (0.28, 0.56)	96	1.07 (0.85, 1.28)
Magallanes	6-8	<20	0.55 (0.29, 0.81)	57	1.77 (1.31, 2.22)
Maule	6-8	69	0.31 (0.24, 0.39)	182	0.79 (0.68, 0.91)
Metro. Central	6-8	87	0.62 (0.49, 0.75)	173	1.14 (0.97, 1.31)
Metro. Norte	6-8	114	0.53 (0.43, 0.62)	218	0.96 (0.84, 1.09)

Table 22: Count and prevalence of ADHD cases by health service and age band in Chile school data for females and males with normal confidence intervals. *(continued)*

		Female		Male	
Health service	Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
Metro. Occidente	6-8	167	0.43 (0.37, 0.50)	296	0.74 (0.66, 0.82)
Metro. Oriente	6-8	115	0.53 (0.43, 0.62)	202	0.90 (0.78, 1.03)
Metro. Sur	6-8	131	0.55 (0.45, 0.64)	240	0.95 (0.83, 1.07)
Metro. Sur Oriente	6-8	189	0.69 (0.59, 0.79)	355	1.22 (1.10, 1.35)
O'Higgins	6-8	124	0.65 (0.54, 0.77)	252	1.27 (1.11, 1.42)
Osorno	6-8	<20	0.36 (0.18, 0.53)	32	0.68 (0.44, 0.91)
Reloncaví	6-8	41	0.45 (0.31, 0.59)	76	0.81 (0.63, 0.99)
Talcahuano	6-8	62	0.99 (0.74, 1.23)	110	1.67 (1.36, 1.98)
Valdivia	6-8	41	0.54 (0.38, 0.71)	102	1.29 (1.04, 1.54)
Valparaíso	6-8	34	0.36 (0.24, 0.47)	89	0.90 (0.71, 1.08)
Viña del Mar	6-8	79	0.38 (0.30, 0.47)	171	0.80 (0.68, 0.92)
Ñuble	6-8	57	0.63 (0.46, 0.79)	110	1.16 (0.94, 1.37)
Aconcagua	9-11	144	2.45 (2.06, 2.85)	246	3.98 (3.49, 4.47)
Aisén	9-11	38	1.60 (1.10, 2.11)	94	3.79 (3.04, 4.55)
Antofagasta	9-11	143	0.97 (0.81, 1.12)	275	1.76 (1.56, 1.97)
Araucanía Norte	9-11	34	0.78 (0.52, 1.04)	130	2.85 (2.37, 3.33)
Araucanía Sur	9-11	240	1.48 (1.29, 1.66)	487	2.87 (2.62, 3.12)
Arauco	9-11	47	1.21 (0.86, 1.55)	107	2.64 (2.15, 3.13)
Arica	9-11	60	1.07 (0.80, 1.34)	118	1.97 (1.62, 2.32)
Atacama	9-11	28	0.37 (0.24, 0.51)	62	0.81 (0.61, 1.01)
Biobío	9-11	175	2.03 (1.73, 2.32)	329	3.67 (3.28, 4.06)
Chiloé	9-11	83	2.31 (1.82, 2.80)	183	4.76 (4.09, 5.43)
Concepción	9-11	333	2.41 (2.16, 2.67)	633	4.37 (4.03, 4.70)
Coquimbo	9-11	321	1.76 (1.57, 1.96)	545	2.93 (2.68, 3.17)
Iquique	9-11	108	1.25 (1.02, 1.48)	245	2.62 (2.30, 2.95)
Magallanes	9-11	83	2.39 (1.89, 2.90)	167	4.82 (4.11, 5.54)
Maule	9-11	227	1.03 (0.90, 1.17)	567	2.43 (2.24, 2.63)
Metro. Central	9-11	198	1.34 (1.15, 1.52)	430	2.72 (2.47, 2.97)
Metro. Norte	9-11	299	1.36 (1.20, 1.51)	507	2.15 (1.96, 2.33)
Metro. Occidente	9-11	372	1.14 (1.02, 1.25)	713	2.05 (1.90, 2.20)
Metro. Oriente	9-11	238	1.07 (0.94, 1.21)	478	2.05 (1.87, 2.23)
Metro. Sur	9-11	386	1.54 (1.39, 1.70)	656	2.48 (2.29, 2.67)
Metro. Sur Oriente	9-11	446	1.56 (1.41, 1.70)	813	2.70 (2.52, 2.88)
O'Higgins	9-11	318	1.60 (1.43, 1.78)	650	3.11 (2.88, 3.35)
Osorno	9-11	40	0.81 (0.56, 1.06)	95	1.84 (1.48, 2.21)
Reloncaví	9-11	80	0.83 (0.65, 1.01)	201	1.97 (1.70, 2.24)
Talcahuano	9-11	200	3.02 (2.60, 3.43)	316	4.47 (3.99, 4.95)
Valdivia	9-11	81	0.99 (0.77, 1.20)	179	2.12 (1.81, 2.42)
Valparaíso	9-11	99	1.03 (0.83, 1.23)	200	2.01 (1.74, 2.29)
Viña del Mar	9-11	231	1.09 (0.95, 1.23)	531	2.39 (2.19, 2.59)
Ñuble	9-11	175	1.83 (1.57, 2.10)	365	3.70 (3.32, 4.07)
Aconcagua	12-14	84	1.52 (1.20, 1.84)	187	3.27 (2.81, 3.74)
Aisén	12-14	39	1.58 (1.09, 2.07)	112	4.22 (3.45, 4.98)
Antofagasta	12-14	94	0.65 (0.52, 0.78)	248	1.59 (1.40, 1.79)
Araucanía Norte	12-14	44	0.97 (0.68, 1.25)	118	2.48 (2.04, 2.92)
Araucanía Sur	12-14	153	0.95 (0.80, 1.10)	389	2.34 (2.11, 2.57)
Arauco	12-14	33	0.89 (0.59, 1.19)	118	3.01 (2.48, 3.55)
Arica	12-14	35	0.67 (0.45, 0.89)	136	2.48 (2.07, 2.89)
Atacama	12-14	31	0.43 (0.28, 0.58)	59	0.78 (0.58, 0.98)
Biobío	12-14	140	1.61 (1.35, 1.87)	358	3.90 (3.50, 4.30)
Chiloé	12-14	95	2.49 (1.99, 2.98)	209	5.15 (4.47, 5.83)
Concepción	12-14	374	2.91 (2.62, 3.20)	641	4.72 (4.36, 5.08)
Coquimbo	12-14	286	1.71 (1.52, 1.91)	565	3.27 (3.01, 3.54)
Iquique	12-14	100	1.22 (0.98, 1.46)	209	2.33 (2.02, 2.65)
Magallanes	12-14	81	2.35 (1.84, 2.86)	178	5.02 (4.30, 5.74)
Maule	12-14	197	0.89 (0.77, 1.02)	544	2.36 (2.16, 2.56)
Metro. Central	12-14	171	1.16 (0.99, 1.33)	400	2.57 (2.32, 2.82)

Table 22: Count and prevalence of ADHD cases by health service and age band in Chile school data for females and males with normal confidence intervals. *(continued)*

		Female		Male	
Health service	Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
Metro. Norte	12-14	270	1.26 (1.11, 1.41)	518	2.29 (2.10, 2.49)
Metro. Occidente	12-14	278	0.89 (0.78, 0.99)	597	1.80 (1.65, 1.94)
Metro. Oriente	12-14	212	0.97 (0.84, 1.10)	458	2.04 (1.85, 2.22)
Metro. Sur	12-14	283	1.17 (1.04, 1.31)	557	2.18 (2.00, 2.36)
Metro. Sur Oriente	12-14	378	1.33 (1.19, 1.46)	687	2.29 (2.13, 2.46)
O'Higgins	12-14	263	1.35 (1.19, 1.51)	633	3.11 (2.87, 3.35)
Osorno	12-14	30	0.60 (0.39, 0.82)	106	2.07 (1.68, 2.46)
Reloncaví	12-14	64	0.66 (0.50, 0.82)	156	1.52 (1.28, 1.76)
Talcahuano	12-14	173	2.60 (2.22, 2.98)	367	5.19 (4.67, 5.70)
Valdivia	12-14	52	0.64 (0.47, 0.81)	139	1.65 (1.38, 1.92)
Valparaíso	12-14	76	0.82 (0.64, 1.00)	218	2.25 (1.96, 2.55)
Viña del Mar	12-14	183	0.88 (0.75, 1.01)	432	2.02 (1.83, 2.21)
Ñuble	12-14	166	1.70 (1.44, 1.95)	375	3.67 (3.31, 4.04)
Aconcagua	15-18	58	0.99 (0.74, 1.24)	128	2.11 (1.75, 2.47)
Aisén	15-18	35	1.35 (0.91, 1.80)	76	2.70 (2.10, 3.30)
Antofagasta	15-18	72	0.48 (0.37, 0.60)	191	1.21 (1.04, 1.38)
Araucanía Norte	15-18	34	0.74 (0.49, 0.98)	53	1.07 (0.79, 1.36)
Araucanía Sur	15-18	98	0.58 (0.47, 0.70)	208	1.17 (1.01, 1.33)
Arauco	15-18	37	0.92 (0.63, 1.22)	103	2.44 (1.98, 2.91)
Arica	15-18	30	0.55 (0.35, 0.74)	66	1.14 (0.87, 1.42)
Atacama	15-18	22	0.31 (0.18, 0.43)	60	0.78 (0.58, 0.97)
Biobío	15-18	128	1.40 (1.16, 1.64)	278	2.83 (2.50, 3.16)
Chiloé	15-18	84	2.00 (1.58, 2.42)	167	3.67 (3.13, 4.22)
Concepción	15-18	373	2.74 (2.46, 3.01)	605	4.16 (3.84, 4.49)
Coquimbo	15-18	231	1.36 (1.19, 1.54)	508	2.79 (2.55, 3.03)
Iquique	15-18	72	0.85 (0.66, 1.05)	179	1.96 (1.68, 2.25)
Magallanes	15-18	80	2.15 (1.69, 2.62)	198	4.87 (4.21, 5.53)
Maule	15-18	99	0.44 (0.35, 0.52)	291	1.19 (1.05, 1.32)
Metro. Central	15-18	139	0.89 (0.74, 1.03)	272	1.61 (1.42, 1.80)
Metro. Norte	15-18	256	1.15 (1.01, 1.29)	370	1.55 (1.39, 1.70)
Metro. Occidente	15-18	189	0.59 (0.51, 0.67)	418	1.21 (1.09, 1.32)
Metro. Oriente	15-18	187	0.79 (0.68, 0.90)	342	1.37 (1.22, 1.51)
Metro. Sur	15-18	203	0.83 (0.71, 0.94)	406	1.55 (1.40, 1.70)
Metro. Sur Oriente	15-18	292	0.95 (0.84, 1.06)	538	1.65 (1.51, 1.79)
O'Higgins	15-18	178	0.89 (0.76, 1.02)	366	1.67 (1.50, 1.84)
Osorno	15-18	31	0.59 (0.38, 0.80)	71	1.26 (0.97, 1.55)
Reloncaví	15-18	45	0.44 (0.31, 0.56)	152	1.36 (1.15, 1.58)
Talcahuano	15-18	162	2.33 (1.98, 2.69)	289	3.89 (3.45, 4.33)
Valdivia	15-18	32	0.38 (0.25, 0.51)	87	0.96 (0.76, 1.16)
Valparaíso	15-18	59	0.60 (0.44, 0.75)	161	1.50 (1.27, 1.73)
Viña del Mar	15-18	105	0.49 (0.40, 0.58)	292	1.24 (1.10, 1.38)
Ñuble	15-18	144	1.40 (1.17, 1.63)	290	2.62 (2.33, 2.92)

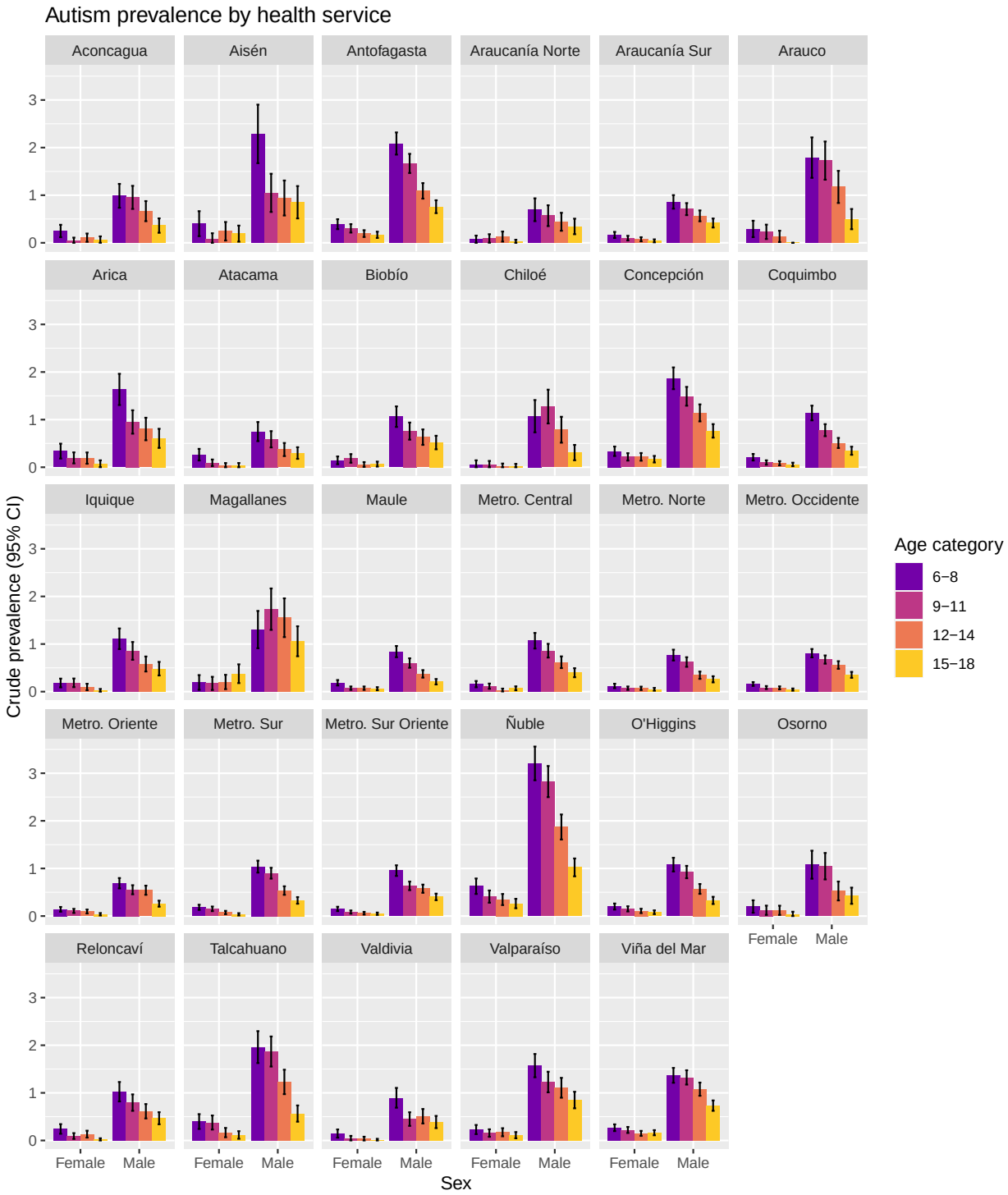


Figure 41: Crude prevalence of autism in school data by health service, age band and sex. Bars show 95% normal confidence intervals.

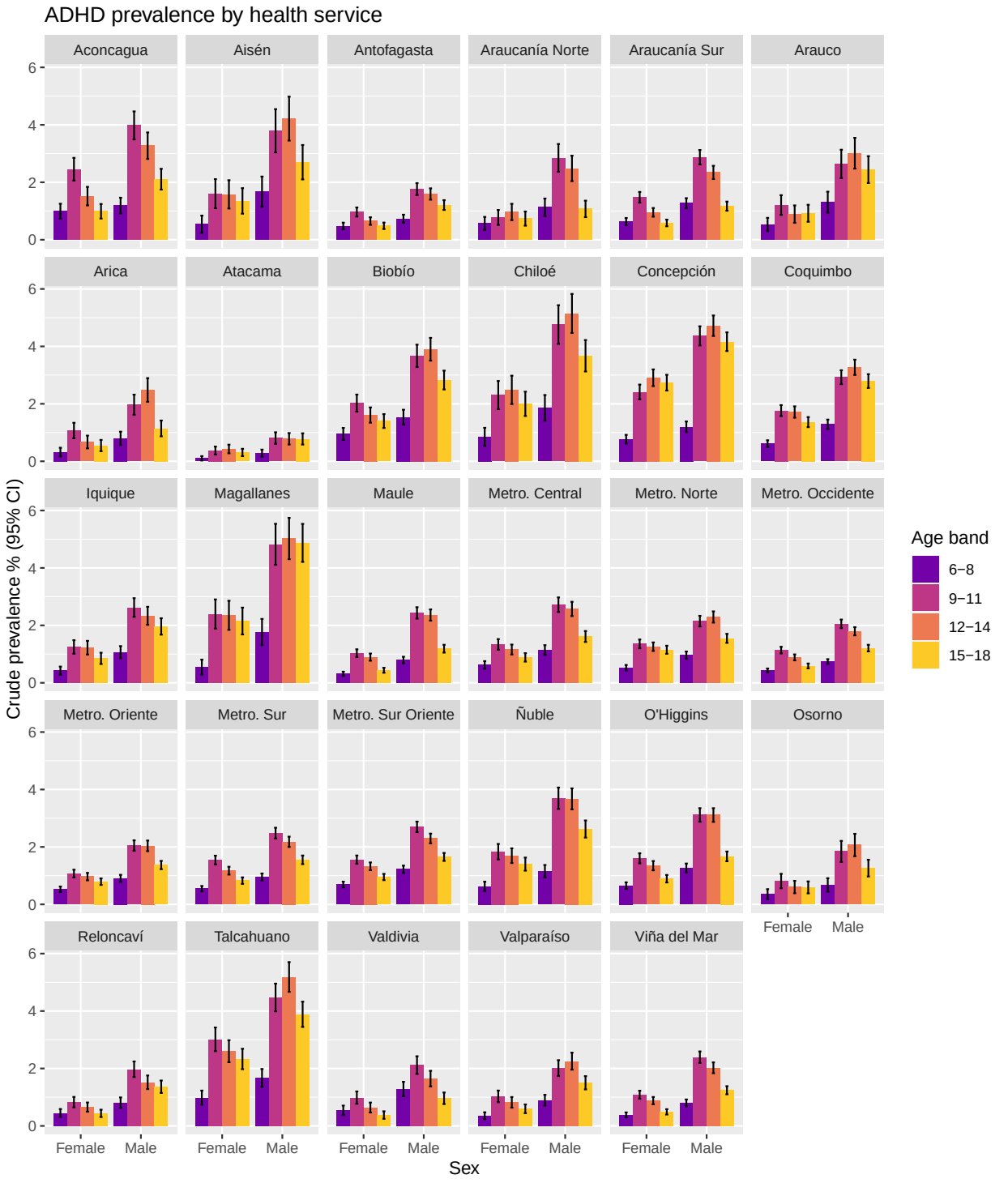


Figure 42: Crude prevalence of ADHD in school data by health service, age band and sex. Bars show 95% normal confidence intervals.

Table 23: Count and prevalence of Autism cases by school fee and age band in Chile school data for females and males with normal confidence intervals.

		Female		Male	
School fee	Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
Free	6-8	655	0.25 (0.23, 0.27)	3799	1.37 (1.33, 1.42)
$1,000-$10,000	6-8	<20	0.00 (0.00, 0.00)	<20	1.17 (0.00, 2.78)
$10,001-$25,000	6-8	<20	0.11 (0.01, 0.20)	20	0.44 (0.25, 0.64)
$25,001-$50,000	6-8	37	0.14 (0.10, 0.19)	168	0.70 (0.59, 0.80)
$50,001-$100,000	6-8	53	0.17 (0.13, 0.22)	296	0.94 (0.83, 1.04)
$100,001+	6-8	<20	0.02 (0.01, 0.04)	30	0.08 (0.05, 0.11)
No information	6-8	<20	0.23 (0.12, 0.35)	73	0.92 (0.71, 1.13)
Free	9-11	434	0.16 (0.15, 0.18)	3147	1.10 (1.06, 1.14)
$1,000-$10,000	9-11	<20	0.00 (0.00, 0.00)	<20	1.59 (0.00, 3.77)
$10,001-$25,000	9-11	<20	0.02 (0.00, 0.06)	23	0.52 (0.31, 0.73)
$25,001-$50,000	9-11	26	0.10 (0.06, 0.14)	152	0.61 (0.52, 0.71)
$50,001-$100,000	9-11	45	0.14 (0.10, 0.18)	255	0.77 (0.67, 0.86)
$100,001+	9-11	<20	0.02 (0.00, 0.03)	35	0.09 (0.06, 0.12)
No information	9-11	<20	0.19 (0.08, 0.30)	75	1.11 (0.86, 1.36)
Free	12-14	332	0.13 (0.11, 0.14)	2246	0.81 (0.77, 0.84)
$1,000-$10,000	12-14	<20	0.00 (0.00, 0.00)	<20	3.67 (0.14, 7.20)
$10,001-$25,000	12-14	<20	0.04 (0.00, 0.10)	<20	0.21 (0.07, 0.34)
$25,001-$50,000	12-14	21	0.08 (0.04, 0.11)	126	0.53 (0.44, 0.63)
$50,001-$100,000	12-14	25	0.08 (0.05, 0.11)	195	0.56 (0.48, 0.64)
$100,001+	12-14	<20	0.01 (0.00, 0.02)	31	0.09 (0.06, 0.12)
No information	12-14	<20	0.14 (0.04, 0.25)	36	0.60 (0.41, 0.80)
Free	15-18	236	0.09 (0.08, 0.10)	1547	0.53 (0.50, 0.55)
$1,000-$10,000	15-18	<20	0.00 (0.00, 0.00)	<20	0.00 (0.00, 0.00)
$10,001-$25,000	15-18	<20	0.06 (0.00, 0.13)	<20	0.23 (0.09, 0.37)
$25,001-$50,000	15-18	<20	0.05 (0.03, 0.08)	82	0.31 (0.24, 0.38)
$50,001-$100,000	15-18	29	0.08 (0.05, 0.11)	153	0.40 (0.33, 0.46)
$100,001+	15-18	<20	0.01 (0.00, 0.01)	25	0.06 (0.04, 0.09)
No information	15-18	<20	0.08 (0.00, 0.16)	30	0.44 (0.29, 0.60)

Table 24: Count and prevalence of ADHD cases by school fee and age band in Chile school data for females and males with normal confidence intervals.

		Female		Male	
School fee	Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
Free	6-8	1543	0.59 (0.56, 0.62)	3117	1.13 (1.09, 1.17)
$1,000-$10,000	6-8	<20	0.00 (0.00, 0.00)	<20	0.58 (0.00, 1.73)
$10,001-$25,000	6-8	32	0.70 (0.46, 0.94)	32	0.71 (0.46, 0.95)
$25,001-$50,000	6-8	167	0.65 (0.55, 0.75)	267	1.11 (0.97, 1.24)
$50,001-$100,000	6-8	200	0.65 (0.56, 0.74)	437	1.38 (1.26, 1.51)
$100,001+	6-8	20	0.05 (0.03, 0.07)	32	0.08 (0.06, 0.11)
No information	6-8	30	0.44 (0.28, 0.59)	58	0.73 (0.54, 0.92)
Free	9-11	4061	1.52 (1.47, 1.57)	8323	2.91 (2.85, 2.97)
$1,000-$10,000	9-11	<20	0.00 (0.00, 0.00)	<20	0.00 (0.00, 0.00)
$10,001-$25,000	9-11	47	1.01 (0.72, 1.29)	102	2.29 (1.85, 2.73)
$25,001-$50,000	9-11	417	1.58 (1.43, 1.73)	616	2.49 (2.29, 2.68)
$50,001-$100,000	9-11	567	1.76 (1.62, 1.91)	1024	3.08 (2.89, 3.26)
$100,001+	9-11	62	0.17 (0.12, 0.21)	112	0.30 (0.24, 0.35)
No information	9-11	73	1.26 (0.97, 1.54)	145	2.15 (1.80, 2.50)
Free	12-14	3342	1.28 (1.24, 1.33)	7634	2.74 (2.68, 2.80)
$1,000-$10,000	12-14	<20	0.00 (0.00, 0.00)	<20	0.00 (0.00, 0.00)
$10,001-$25,000	12-14	48	1.02 (0.74, 1.31)	79	1.80 (1.41, 2.20)
$25,001-$50,000	12-14	364	1.36 (1.22, 1.50)	638	2.70 (2.49, 2.90)
$50,001-$100,000	12-14	509	1.55 (1.42, 1.68)	1083	3.12 (2.94, 3.30)
$100,001+	12-14	72	0.20 (0.15, 0.24)	147	0.41 (0.34, 0.47)
No information	12-14	50	1.03 (0.74, 1.31)	133	2.23 (1.85, 2.60)
Free	15-18	2579	0.97 (0.93, 1.00)	5529	1.88 (1.83, 1.93)
$1,000-$10,000	15-18	<20	0.00 (0.00, 0.00)	<20	0.39 (0.00, 1.14)
$10,001-$25,000	15-18	21	0.43 (0.25, 0.62)	25	0.57 (0.35, 0.80)
$25,001-$50,000	15-18	343	1.17 (1.04, 1.29)	482	1.84 (1.67, 2.00)
$50,001-$100,000	15-18	416	1.12 (1.01, 1.22)	898	2.33 (2.18, 2.48)
$100,001+	15-18	79	0.21 (0.16, 0.26)	151	0.39 (0.33, 0.45)
No information	15-18	37	0.73 (0.50, 0.97)	79	1.17 (0.91, 1.43)

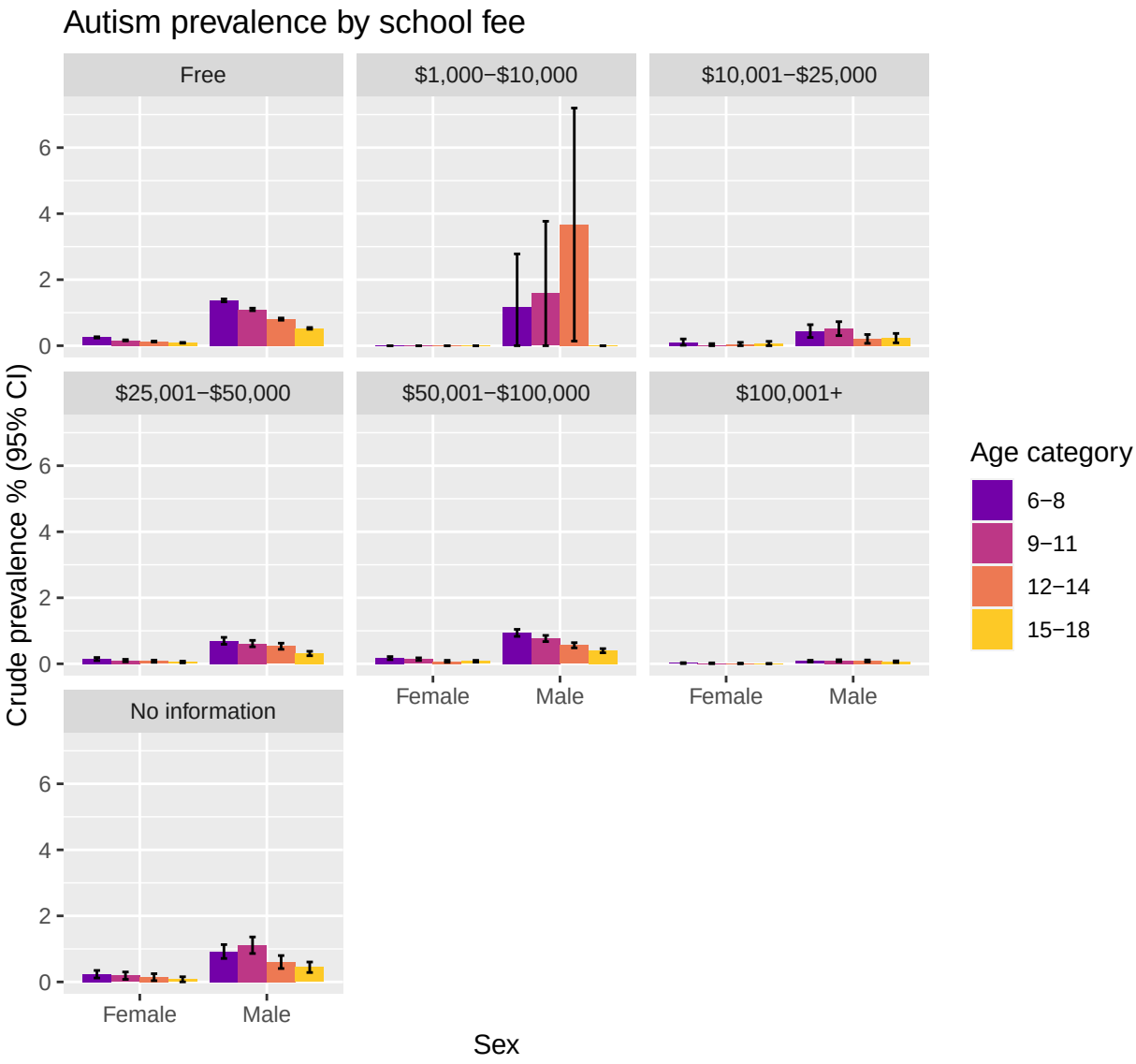


Figure 43: Crude prevalence of autism in school data by student's monthly school fee (Peso), age band and sex. Bars show 95% normal confidence intervals.

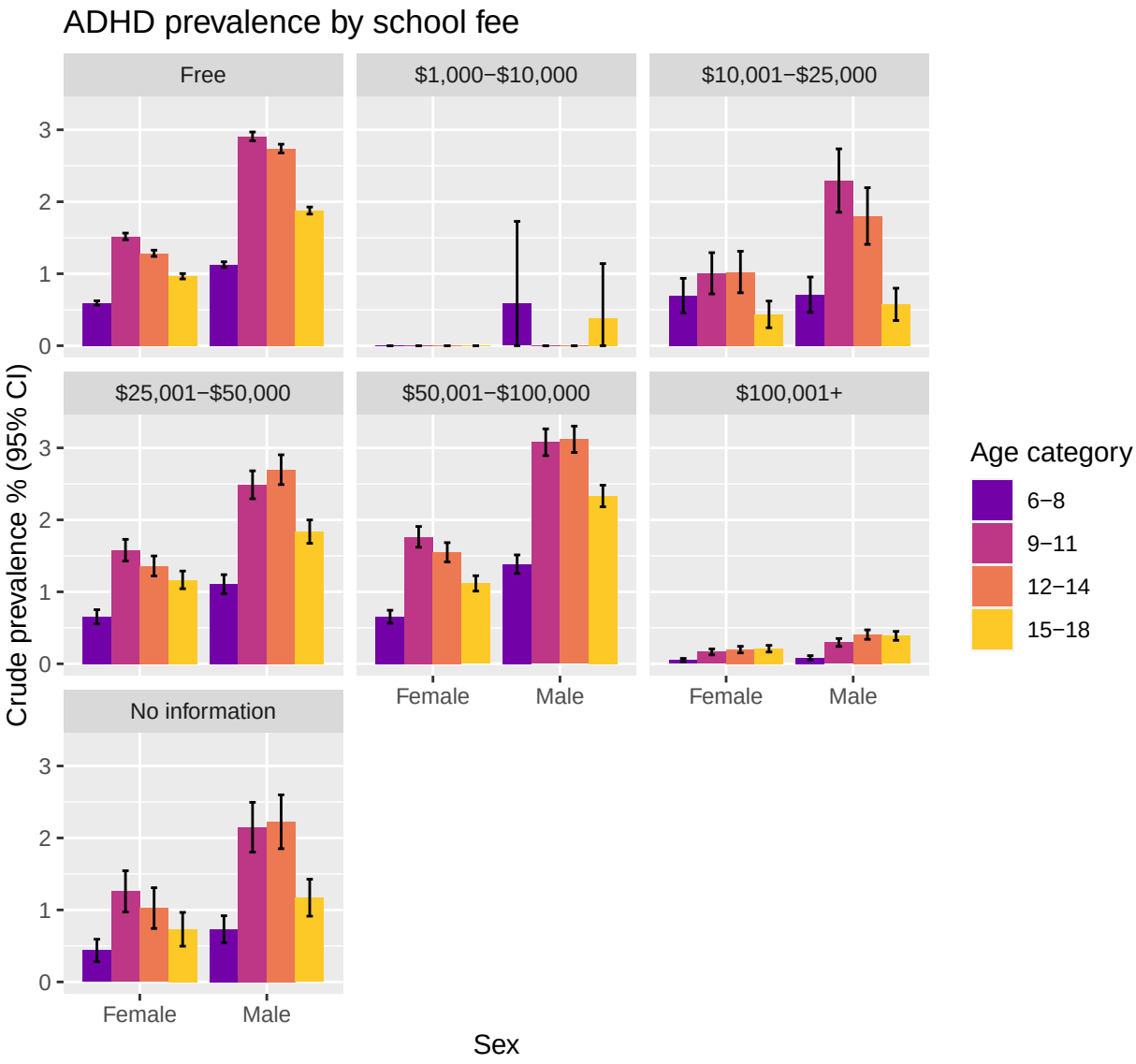


Figure 44: Crude prevalence of ADHD in school data by student's monthly school fee (Peso), age band and sex. Bars show 95% normal confidence intervals.

Table 25: Count and prevalence of Autism cases by ethnicity and age band in Chile school data for females and males with normal confidence intervals. Regions with high Mapuche populations only.

		Female		Male	
Ethnicity	Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
Mapuche	6-8	29	0.15 (0.10, 0.20)	186	0.93 (0.80, 1.07)
Other Indigenous group	6-8	<20	0.00 (0.00, 0.00)	<20	0.77 (0.16, 1.39)
No Indigenous group	6-8	377	0.18 (0.16, 0.20)	2220	1.02 (0.98, 1.06)
Mapuche	9-11	<20	0.05 (0.02, 0.09)	155	0.72 (0.61, 0.84)
Other Indigenous group	9-11	<20	0.17 (0.00, 0.50)	<20	1.36 (0.42, 2.30)
No Indigenous group	9-11	264	0.13 (0.11, 0.14)	1852	0.84 (0.80, 0.87)
Mapuche	12-14	<20	0.06 (0.03, 0.09)	110	0.52 (0.42, 0.62)
Other Indigenous group	12-14	<20	0.00 (0.00, 0.00)	<20	0.34 (0.00, 0.81)
No Indigenous group	12-14	192	0.09 (0.08, 0.11)	1385	0.64 (0.61, 0.67)
Mapuche	15-18	<20	0.02 (0.00, 0.03)	71	0.35 (0.27, 0.43)
Other Indigenous group	15-18	<20	0.00 (0.00, 0.00)	<20	0.19 (0.00, 0.56)
No Indigenous group	15-18	136	0.06 (0.05, 0.07)	969	0.41 (0.39, 0.44)

Table 26: Count and prevalence of ADHD cases by ethnicity and age band in Chile school data for females and males with normal confidence intervals. Regions with high Mapuche populations only.

		Female		Male	
Ethnicity	Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
Mapuche	6-8	113	0.58 (0.48, 0.69)	218	1.09 (0.95, 1.24)
Other Indigenous group	6-8	<20	0.39 (0.00, 0.84)	<20	0.90 (0.24, 1.57)
No Indigenous group	6-8	1227	0.59 (0.56, 0.62)	2342	1.07 (1.03, 1.12)
Mapuche	9-11	253	1.24 (1.08, 1.39)	557	2.59 (2.38, 2.81)
Other Indigenous group	9-11	<20	1.19 (0.31, 2.06)	<20	1.87 (0.78, 2.97)
No Indigenous group	9-11	3113	1.48 (1.43, 1.53)	5950	2.68 (2.62, 2.75)
Mapuche	12-14	206	1.00 (0.87, 1.14)	479	2.26 (2.06, 2.46)
Other Indigenous group	12-14	<20	0.72 (0.02, 1.42)	<20	2.54 (1.27, 3.81)
No Indigenous group	12-14	2660	1.29 (1.24, 1.34)	5614	2.59 (2.52, 2.66)
Mapuche	15-18	131	0.70 (0.58, 0.81)	278	1.38 (1.22, 1.54)
Other Indigenous group	15-18	<20	1.14 (0.23, 2.05)	<20	0.38 (0.00, 0.90)
No Indigenous group	15-18	2268	1.03 (0.99, 1.08)	4353	1.86 (1.80, 1.91)

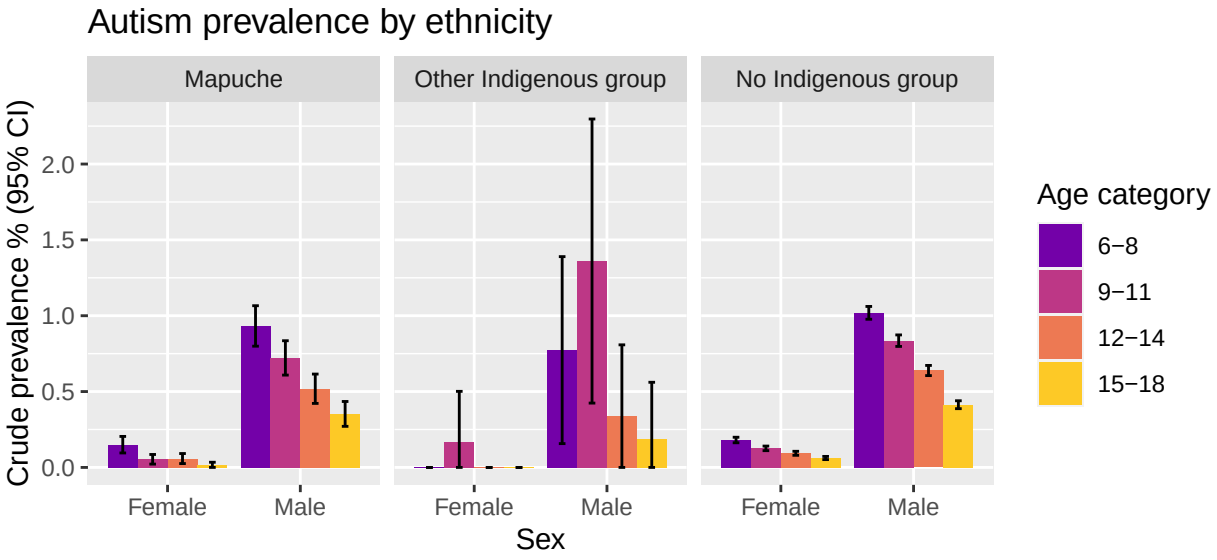


Figure 45: Crude prevalence of autism in school data by ethnicity, age band and sex. Bars show 95% normal confidence intervals. Regions with high Mapuche populations only.

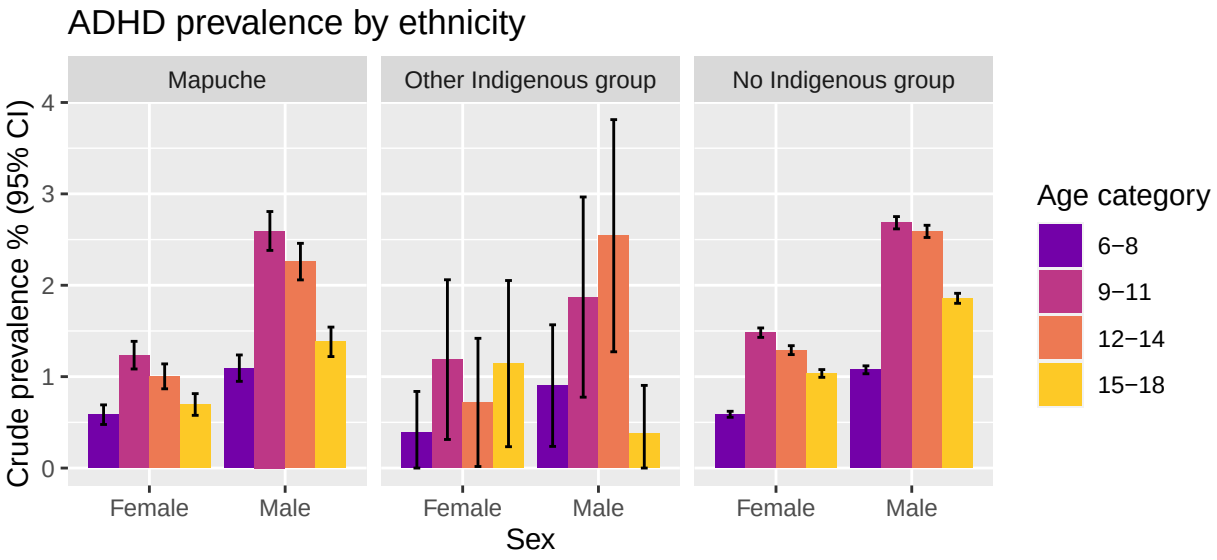


Figure 46: Crude prevalence of ADHD in school data by ethnicity, age band and sex. Bars show 95% normal confidence intervals. Regions with high Mapuche populations only.

Table 27: Count and prevalence of Autism cases by school's rurality and age band in Chile school data for females and males with normal confidence intervals.

		Female		Male	
School rurality	Age band	Autism cases	Prevalence % (95% CI)	Autism cases	Prevalence % (95% CI)
Rural	6-8	93	0.26 (0.21, 0.31)	519	1.33 (1.22, 1.45)
Urban	6-8	681	0.21 (0.19, 0.22)	3869	1.13 (1.09, 1.16)
Rural	9-11	61	0.16 (0.12, 0.20)	472	1.14 (1.03, 1.24)
Urban	9-11	462	0.14 (0.12, 0.15)	3217	0.91 (0.88, 0.95)
Rural	12-14	44	0.16 (0.12, 0.21)	291	0.93 (0.82, 1.03)
Urban	12-14	347	0.10 (0.09, 0.11)	2356	0.67 (0.64, 0.70)
Rural	15-18	<20	0.08 (0.03, 0.13)	92	0.61 (0.49, 0.73)
Urban	15-18	281	0.08 (0.07, 0.08)	1755	0.45 (0.42, 0.47)

Table 28: Count and prevalence of ADHD cases by school's rurality and age band in Chile school data for females and males with normal confidence intervals.

		Female		Male	
School rurality	Age band	ADHD cases	Prevalence % (95% CI)	ADHD cases	Prevalence % (95% CI)
Rural	6-8	221	0.62 (0.53, 0.70)	503	1.29 (1.18, 1.40)
Urban	6-8	1771	0.54 (0.51, 0.56)	3441	1.00 (0.97, 1.03)
Rural	9-11	532	1.41 (1.29, 1.53)	1304	3.14 (2.97, 3.31)
Urban	9-11	4695	1.40 (1.36, 1.44)	9018	2.56 (2.51, 2.62)
Rural	12-14	323	1.21 (1.08, 1.34)	946	3.02 (2.83, 3.21)
Urban	12-14	4062	1.20 (1.16, 1.23)	8768	2.49 (2.44, 2.54)
Rural	15-18	112	0.97 (0.79, 1.15)	287	1.90 (1.68, 2.12)
Urban	15-18	3363	0.91 (0.88, 0.94)	6878	1.74 (1.70, 1.79)

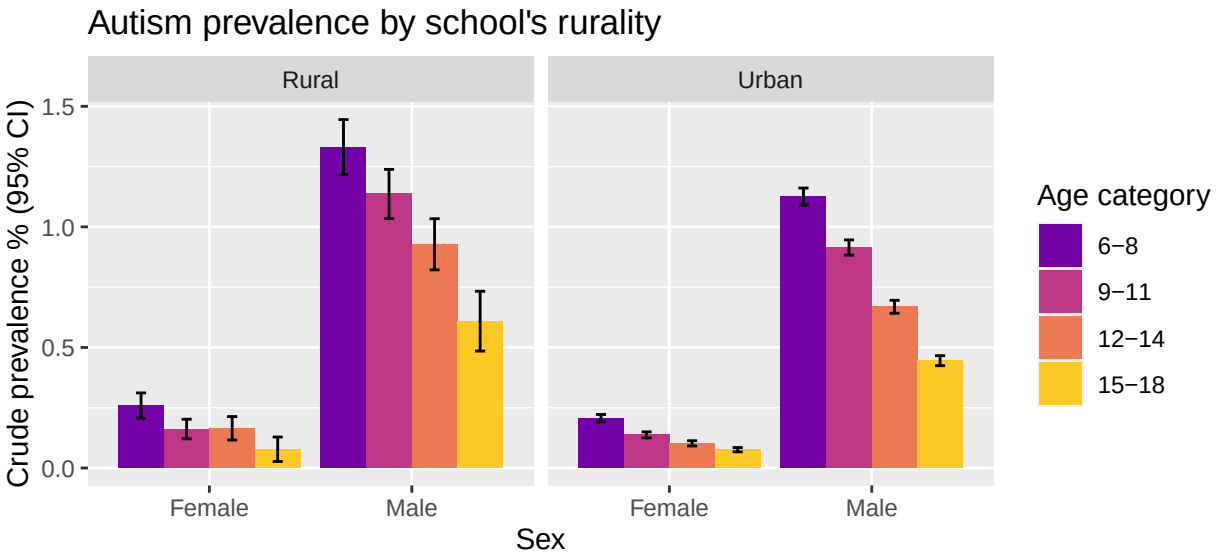


Figure 47: Crude prevalence of autism in school data by school's rurality, age band and sex. Bars show 95% normal confidence intervals.

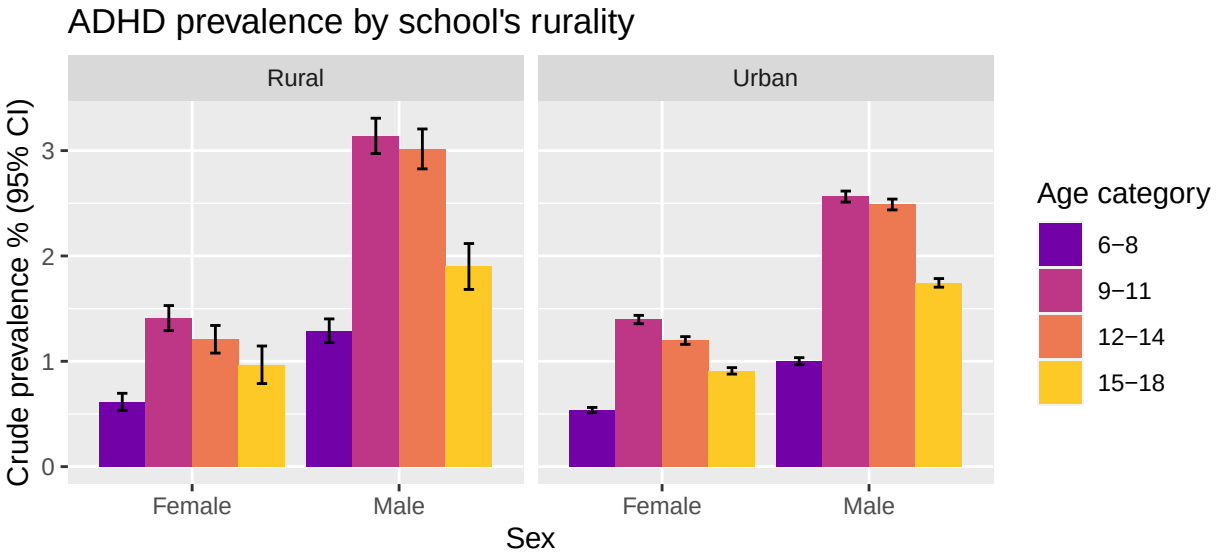


Figure 48: Crude prevalence of ADHD in school data by school's rurality, age band and sex. Bars show 95% normal confidence intervals.

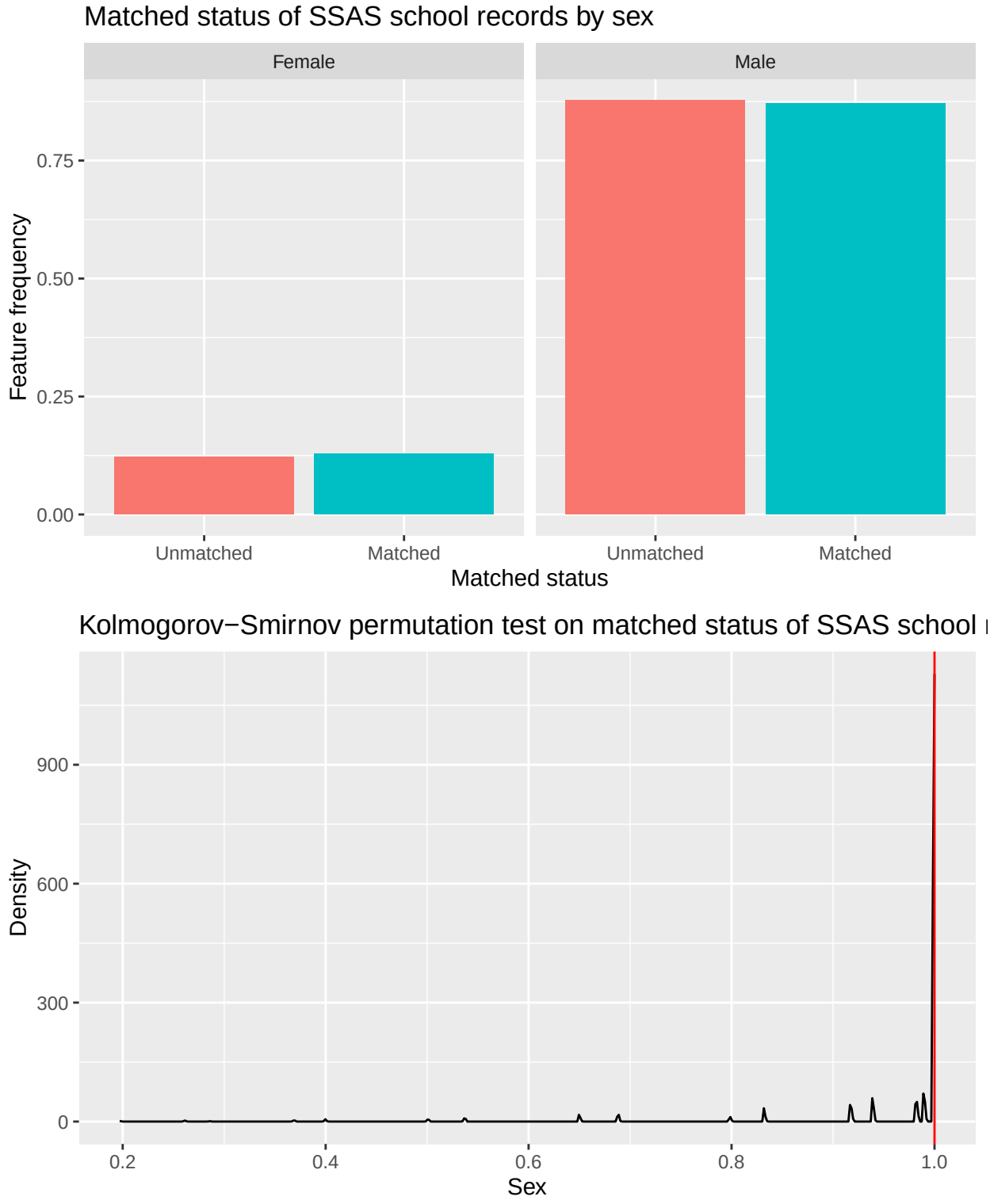
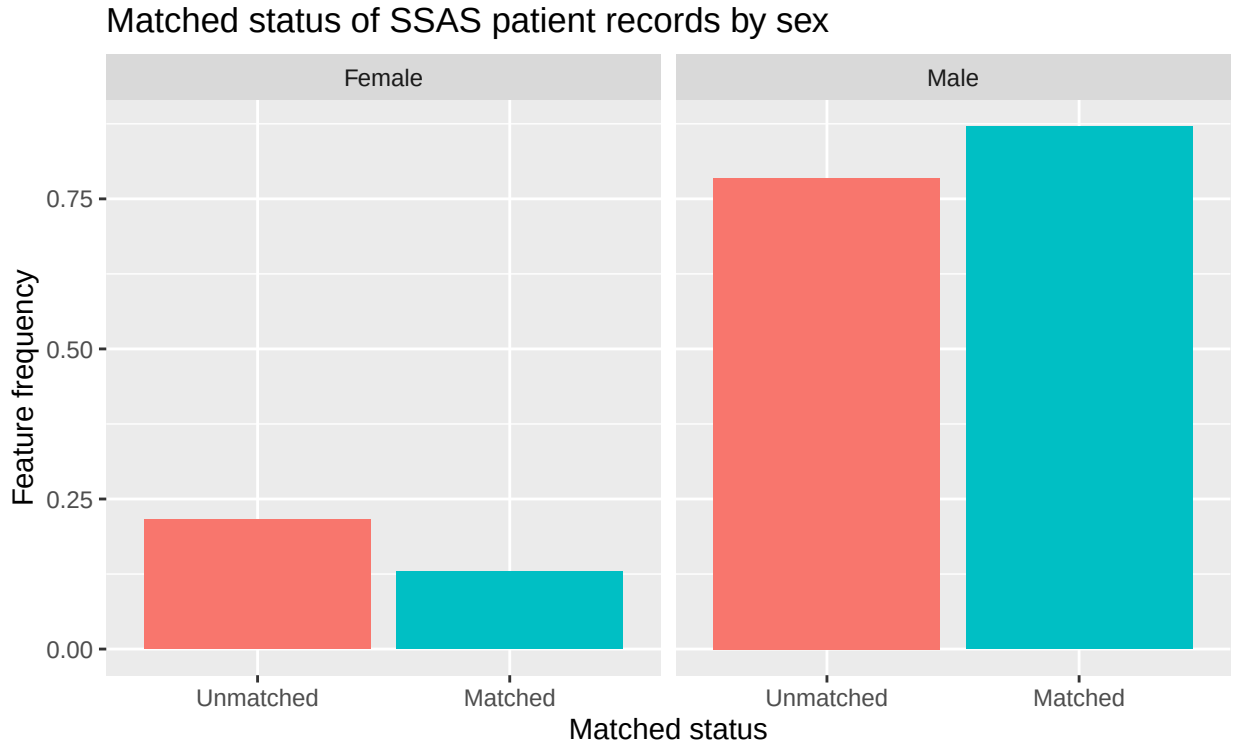


Figure 49: a) Difference in frequency of sexes among school records that matched to patient records and school records that did not match. b) Density of Kolmogorov-Smirnov p-values for 2000 permutations of matched status for school records by sex with observed p-value shown in red.



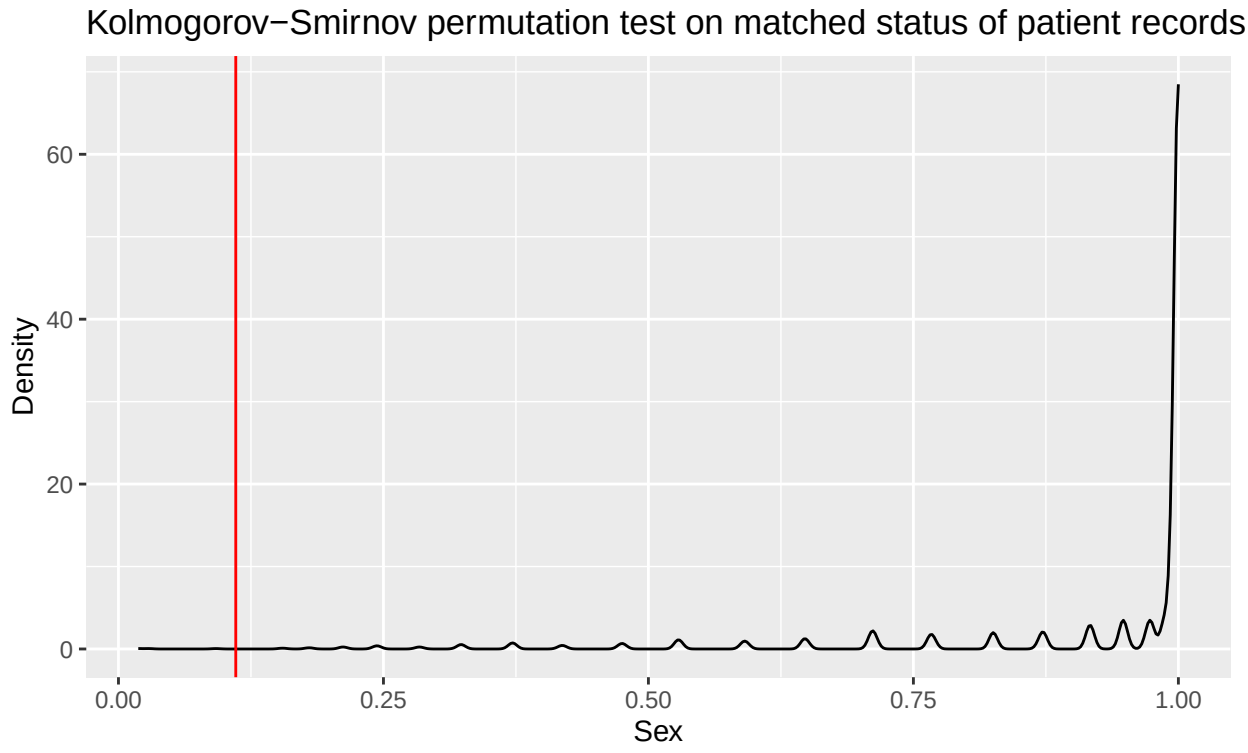


Figure 50: a) Difference in frequency of sexes among patient records that matched to school records and patient records that did not match. b) Density of Kolmogorov-Smirnov p-values for 2000 permutations of matched status for patient records by sex with observed p-value shown in red.

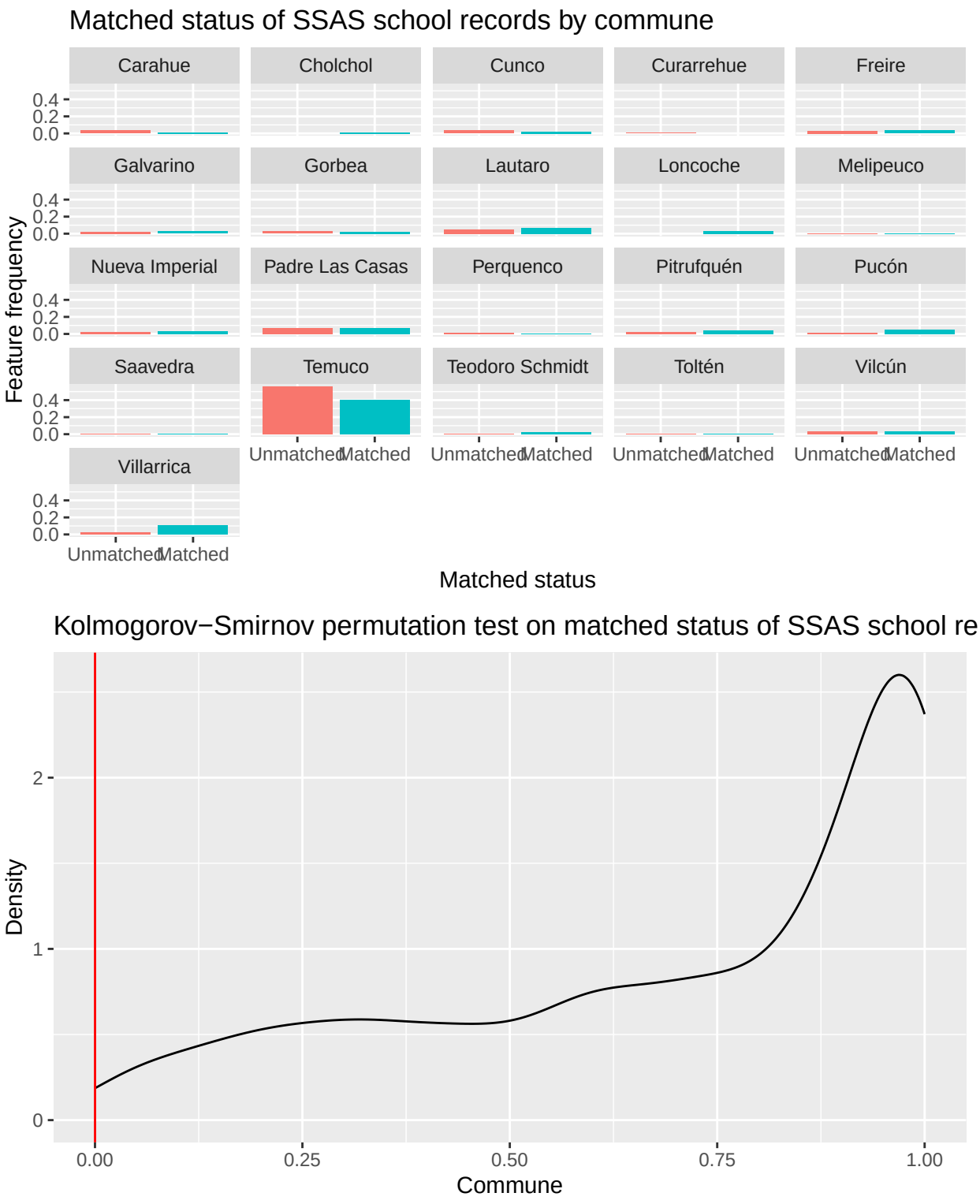


Figure 51: a) Difference in frequency of communes of residence among school records that matched to patient records and school records that did not match. b) Density of Kolmogorov-Smirnov p-values for 2000 permutations of matched status for school records by commune of residence with observed p-value shown in red.

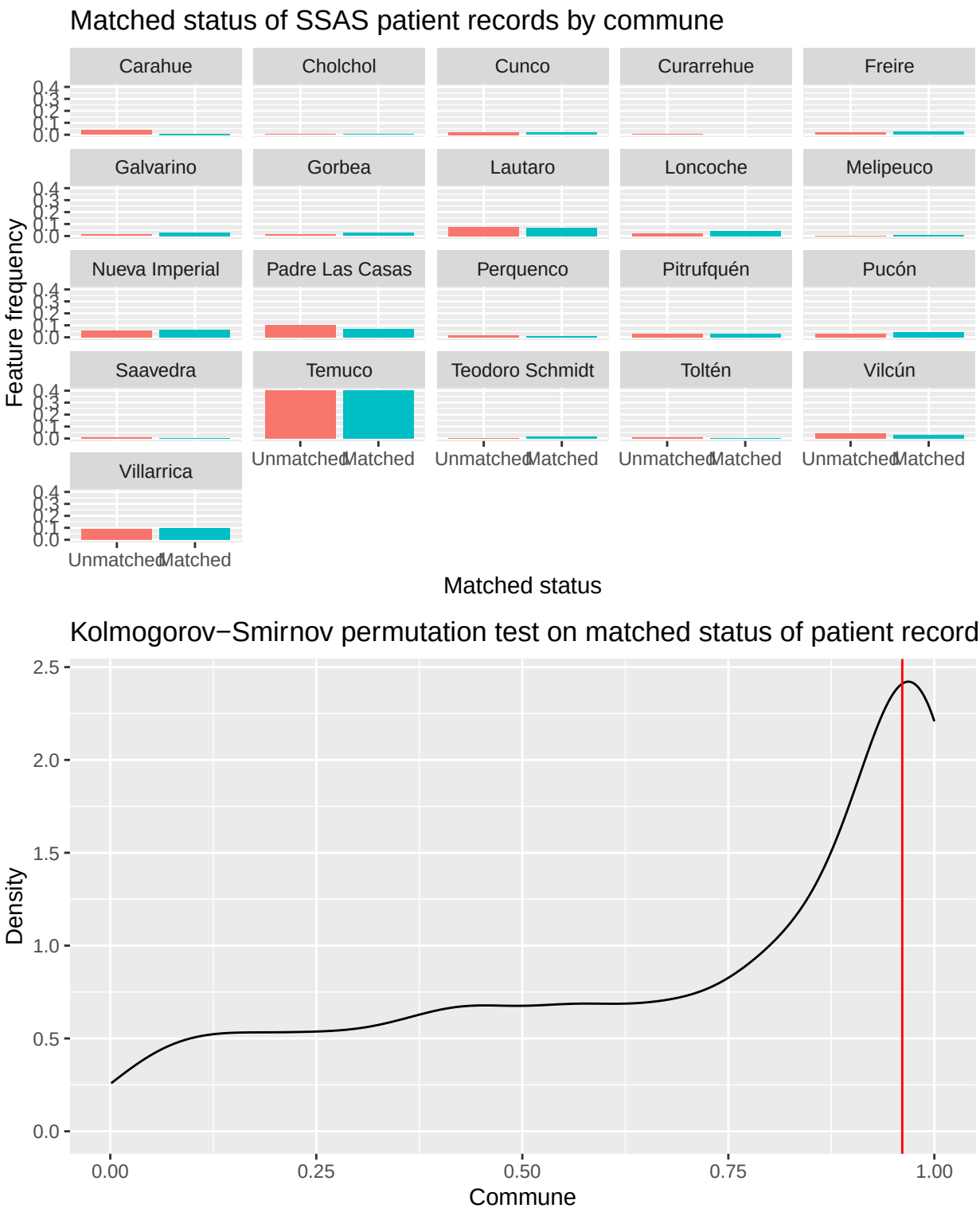


Figure 52: a) Difference in frequency of commune of residence among patient records that matched to school records and patient records that did not match. b) Density of Kolmogorov-Smirnov p-values for 2000 permutations of matched status for patient records by commune of residence with observed p-value shown in red.

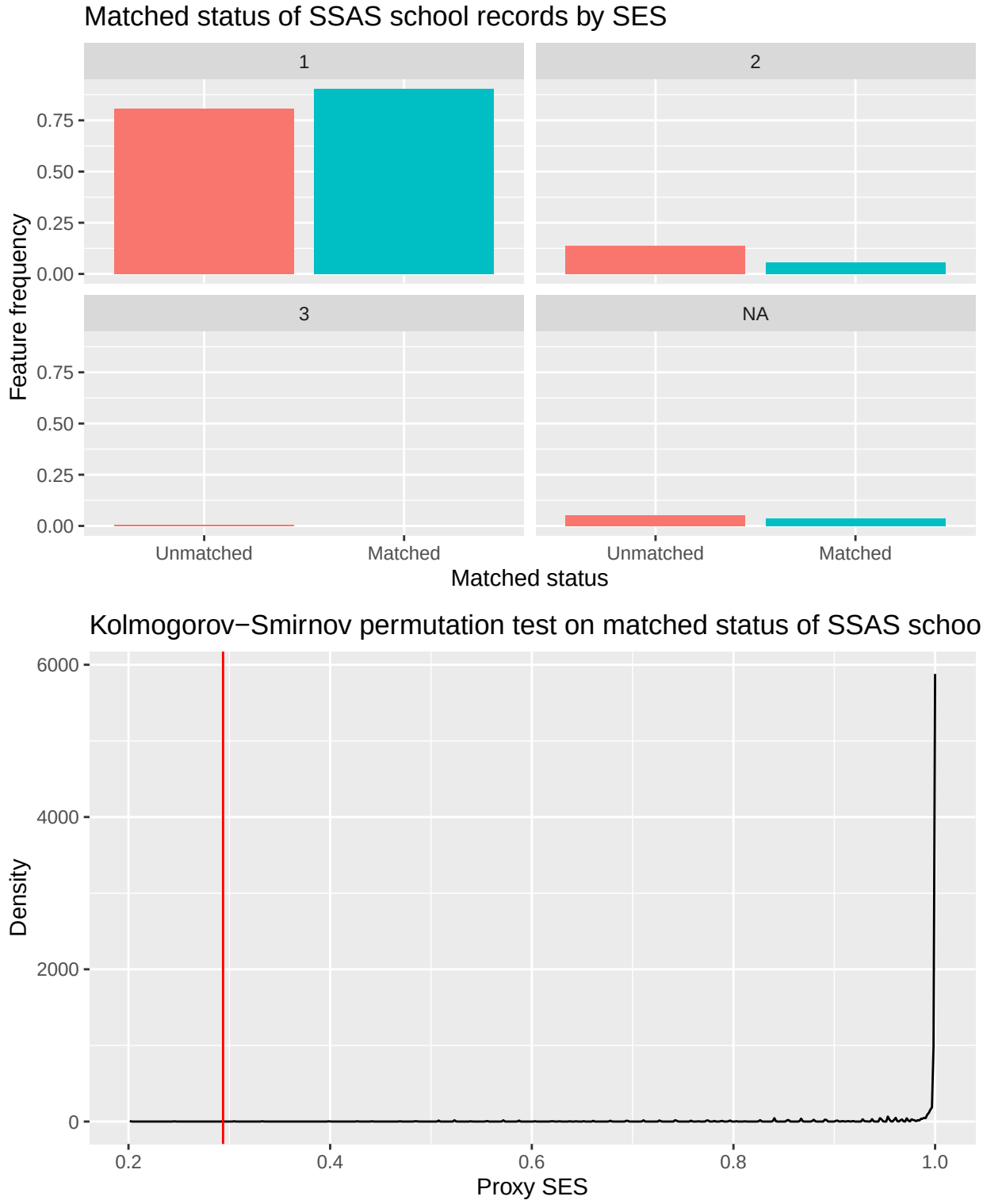


Figure 53: a) Difference in frequency of proxy SES among school records that matched to patient records and school records that did not match. b) Density of Kolmogorov-Smirnov p-values for 2000 permutations of matched status for school records by proxy SES with observed p-value shown in red.

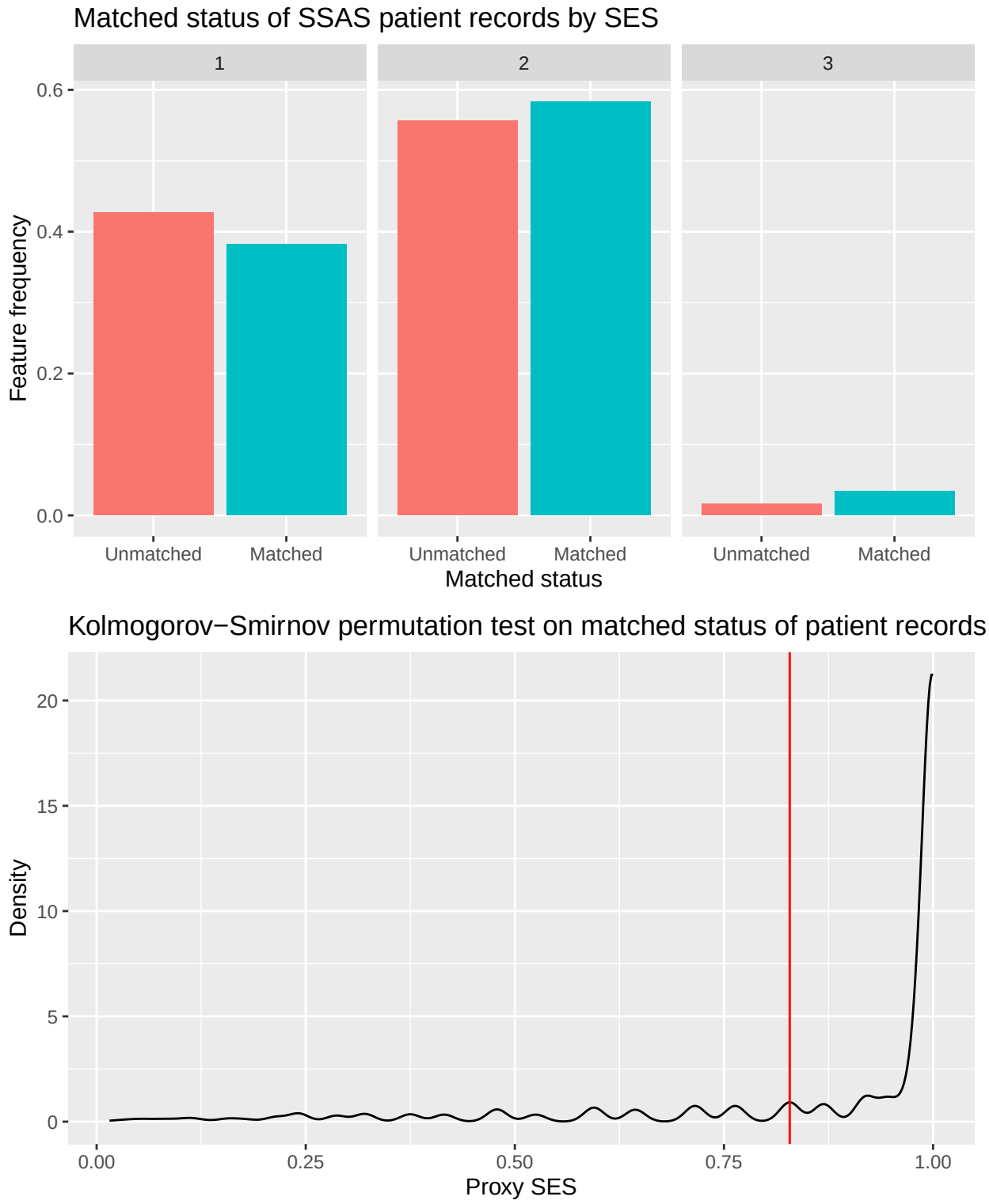


Figure 54: a) Difference in frequency of proxy SES among patient records that matched to school records and patient records that did not match. b) Density of Kolmogorov-Smirnov p-values for 2000 permutations of matched status for patient records by proxy SES with observed p-value shown in red.

9 References

1. World Bank. GDP per capita (current US$) - Chile [Internet]. [cited 2023 Jul 17]. Available from: https://data.worldbank.org
2. Organisation for Economic Co-operation and Development (OECD). Chile - OECD Data [Internet]. [cited 2023 Jul 14]. Available from: http://data.oecd.org/chile.htm
3. Cerda AA, García LY, Rivera-Arroyo J, Riquelme A, Teixeira JP, Jakovljevic M. Comparison of the healthcare system of Chile and Brazil: strengths, inefficiencies, and expenditures. Cost Eff Resour Alloc. 2022 Dec 16;20(1):71.
4. Ministerio de Salud. Ministerio de Salud – Gobierno de Chile. [cited 2023 Jul 14]. Ubicación Servicios de Salud. Available from: https://www.minsal.cl/conozcanos_servicios_salud/
5. Koch KJ, Cid Pedraza C, Schmid A. Out-of-pocket expenditure and financial protection in the Chilean health care system—A systematic review. Health Policy. 2017 May 1;121(5):481–94.
6. Núñez A, Manzano CA. Identifying local barriers to access to healthcare services in Chile using a communitarian approach. Health Expect Int J Public Particip Health Care Health Policy. 2022 Feb;25(1):254–63.
7. Castillo-Laborde C, Aguilera-Sanhueza X, Hirmas-Adauy M, Matute I, Delgado-Becerra I, Ferrari MND, et al. Health Insurance Scheme Performance and Effects on Health and Health Inequalities in Chile. MEDICC Rev. 2017 Jul;19:57–64.
8. World Bank. Gini Index - Chile [Internet]. [cited 2023 Jul 17]. Available from: https://data.worldbank.org
9. Paredes D, Iturra V, Lufin M. A Spatial Decomposition of Income Inequality in Chile. Reg Stud. 2016 May 3;50(5):771–89.
10. Instituto Nacional de Estadisticas. Default. [cited 2023 Jul 5]. Encuesta suplementaria de ingresos. Available from: http://www.ine.gob.cl/estadisticas/sociales/seguridad-publica-y-justicia/estadisticas-policiales-y-judiciales/encuesta-suplementaria-de-ingresos
11. Sandoval MH, Portaccio MEA. Death certificate: The urgent consideration of ethnic and racial origin in Chile. Lancet Reg Health – Am [Internet]. 2022 Dec 1 [cited 2023 Jul 14];16. Available from: https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(22)00219-8/fulltext
12. Sandoval MH, Alvear Portaccio ME, Albala C. Life expectancy by ethnic origin in Chile. Front Public Health [Internet]. 2023 [cited 2023 Jul 14];11. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1147542
13. Instituto Nacional de Estadisticas. Default. [cited 2023 Jul 14]. Censo de Población y Vivienda. Available from: http://www.ine.gob.cl/estadisticas/sociales/censos-de-poblacion-y-vivienda/censo-de-poblacion-y-vivienda
14. Istuany OE, Wood R. Perspectives on Educational Inclusion from a Small Sample of Autistic Pupils in Santiago, Chile. 2020 Jul 6;22(1):210–20.
15. Breinbauer C, Vidal V, Molina P, Trabucco C, Gutierrez L, Cordero M. Early Childhood Development policy in Chile: Progress and pitfalls supporting children with developmental disabilities toward school readiness. Front Public Health [Internet]. 2022 [cited 2023 Jul 18];10. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2022.983513
16. Ministerio de Educación. Educación Especial. [cited 2023 Jun 28]. Incremento de la Subvención Especial Diferencial. Available from: https://especial.mineduc.cl/incremento-subvencion-educacion-especial/incremento-la-subvencion-especial-diferencial/
17. Centro de Innovación en Educación Fundación Chile. Análisis de la implementación de los programas de integración escolar (PIE) en establecimientos que han incorporado estudiantes con necesidades

educativas especiales transitorias (NEET) [Internet]. 2013 [cited 2023 Jul 18]. Available from: https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/18386/EV13-0012.pdf?sequence=1

18. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders [Internet]. 5th ed. Arlington, VA; 2013 [cited 2023 Jul 15]. Available from: https://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596

19. World Health Organisation. International Statistical Classification of Diseases and Related Health Problems (10th Revision) [Internet]. 2019. Available from: https://icd.who.int/browse10/2019/en#/F84

20. Lai MC, Lombardo MV, Baron-Cohen S. Autism. The Lancet. 2014 Mar 8;383(9920):896–910.

21. Paula CS, Cukier S, Cunha GR, Irarrázaval M, Montiel-Nava C, Garcia R, et al. Challenges, priorities, barriers to care, and stigma in families of people with autism: Similarities and differences among six Latin American countries. Autism. 2020 Nov 1;24(8):2228–42.

22. Roman-Urrestarazu A, Yáñez C, López-Garí C, Elgueta C, Allison C, Brayne C, et al. Autism screening and conditional cash transfers in Chile: Using the Quantitative Checklist (Q-CHAT) for early autism detection in a low resource setting. Autism. 2021 May 1;25(4):932–45.

23. Brett D, Warnell F, McConachie H, Parr JR. Factors Affecting Age at ASD Diagnosis in UK: No Evidence that Diagnosis Age has Decreased Between 2004 and 2014. J Autism Dev Disord. 2016;46:1974–84.

24. Yáñez C, Maira P, Elgueta C, Brito M, Crockett MA, Troncoso L, et al. [Prevalence estimation of Autism Spectrum disorders in Chilean urban population]. Andes Pediatr Rev Chil Pediatr. 2021 Aug;92(4):519–25.

25. Catherine Rice. Centers for Disease Control and Prevention. 2006 [cited 2023 Jul 18]. Prevalence of Autism Spectrum Disorders. Available from: https://www.cdc.gov/mmwr/preview/mmwrhtml/ss5810a1.htm

26. Elsabbagh M, Divan G, Koh YJ, Kim YS, Kauchali S, Marcín C, et al. Global Prevalence of Autism and Other Pervasive Developmental Disorders. Autism Res. 2012;5(3):160–79.

27. Puga C, Pagotto V, Giunta D, Vicens J, Leist M, Vaucheret Paz E, et al. [Prevalence and incidence of disability based on the Unique Certificate of Disability at a teaching hospital in the Metropolitan Area of Buenos Aires]. Arch Argent Pediatr. 2019 Jun 1;117(3):183–7.

28. Dekkers LMS, Groot NA, Díaz Mosquera EN, Andrade Zúñiga IP, Delfos MF. Prevalence of Autism Spectrum Disorders in Ecuador: A Pilot Study in Quito. J Autism Dev Disord. 2015 Dec 1;45(12):4165–73.

29. Russell G, Stapley S, Newlove-Delgado T, Salmon A, White R, Warren F, et al. Time trends in autism diagnosis over 20 years: a UK population-based cohort study. J Child Psychol Psychiatry. 2022;63(6):674–82.

30. Roman-Urrestarazu A, van Kessel R, Allison C, Matthews FE, Brayne C, Baron-Cohen S. Association of Race/Ethnicity and Social Disadvantage With Autism Prevalence in 7 Million School Children in England. JAMA Pediatr. 2021 Jun 7;175(6):e210054–e210054.

31. Loomes R, Hull L, Mandy WPL. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. J Am Acad Child Adolesc Psychiatry. 2017 Jun;56(6):466–74.

32. Giarelli E, Wiggins LD, Rice CE, Levy SE, Kirby RS, Pinto-Martin J, et al. Sex differences in the evaluation and diagnosis of autism spectrum disorders among children. Disabil Health J. 2010 Apr;3(2):107–16.

33. Shenouda J, Barrett E, Davidow AL, Halperin W, Silenzio VMB, Zahorodny W. Prevalence of autism spectrum disorder in a large, diverse metropolitan area: Variation by sociodemographic factors. Autism Res. 2022;15(1):146–55.

34. Thomas P, Zahorodny W, Peng B, Kim S, Jani N, Halperin W, et al. The association of autism diagnosis with socioeconomic status. Autism. 2012 Mar 1;16(2):201–13.

35. Delobel-Ayoub M, Ehlinger V, Klapouszczak D, Maffre T, Raynaud JP, Delpierre C, et al. Socioeconomic Disparities and Prevalence of Autism Spectrum Disorders and Intellectual Disability. PLOS ONE. 2015 Nov 5;10(11):e0141964.

36. Bailey B, Arciuli J. Indigenous Australians with autism: A scoping review. Autism. 2020 Jul 1;24(5):1031–46.

37. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: A systematic review update. Autism Res. 2022;15(5):778–90.

38. Antshel KM, Russo N. Autism Spectrum Disorders and ADHD: Overlapping Phenomenology, Diagnostic Issues, and Treatment Considerations. Curr Psychiatry Rep. 2019 Mar 22;21(5):34.

39. Fioravante I, Lozano-Lozano JA, Martella D. Attention deficit hyperactivity disorder: A pilot study for symptom assessment and diagnosis in children in Chile. Front Psychol [Internet]. 2022 [cited 2023 Jul 15];13. Available from: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.946273

40. Salazar F, Baird G, Chandler S, Tseng E, O'sullivan T, Howlin P, et al. Co-occurring Psychiatric Disorders in Preschool and Elementary School-Aged Children with Autism Spectrum Disorder. J Autism Dev Disord. 2015 Aug 1;45(8):2283–94.

41. de la Barra FE, Vicente B, Saldivia S, Melipillan R. Epidemiology of ADHD in Chilean children and adolescents. ADHD Atten Deficit Hyperact Disord. 2013 Mar 1;5(1):1–8.

42. Vicente B, de la Barra F, Saldivia S, Kohn R, Rioseco P, Melipillan R. Prevalence of child and adolescent psychiatric disorders in Santiago, Chile: a community epidemiological study. Soc Psychiatry Psychiatr Epidemiol. 2012 Jul 1;47(7):1099–109.

43. Rowland AS, Skipper BJ, Rabiner DL, Qeadan F, Campbell RA, Naftel AJ, et al. Attention-Deficit/Hyperactivity Disorder (ADHD): Interaction between socioeconomic status and parental history of ADHD determines prevalence. J Child Psychol Psychiatry. 2018;59(3):213–22.

44. Borges e Azevêdo PV, Caixeta LF, Rabelo Taveira DL, Peixoto Giglio MR, Rosário MC do, Rohde LA. Suggestive diagnosis of attention-deficit/hyperactivity disorder in indigenous children and adolescents from the Brazilian Amazon. Eur Child Adolesc Psychiatry. 2020 Mar 1;29(3):373–84.

45. Pham HD, Nguyen HBH, Tran DT. Prevalence of ADHD in primary school children in Vinh Long, Vietnam. Pediatr Int. 2015;57(5):856–9.

46. Ince RA, Paton AT, Kay JW, Schyns PG. Bayesian inference of population prevalence. Serences JT, Behrens TE, Huth A, Ling S, editors. eLife. 2021 Oct 6;10:e62461.

47. Downing BC, Hickman M, Jones NR, Larney S, Sweeting MJ, Xu Y, et al. Prevalence of opioid dependence in New South Wales, Australia, 2014–16: Indirect estimation from multiple data sources using a Bayesian approach. Addiction [Internet]. [cited 2023 Jul 16];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/add.16268

48. Depaoli S, Winter SD, Visser M. The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. Front Psychol. 2020 Nov 24;11:608045.

49. Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunananthan S, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. J Clin Epidemiol. 2010 Jun 1;63(6):638–46.

50. Costa PS, Santos NC, Cunha P, Cotter J, Sousa N. The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing. J Aging Res. 2013;2013:302163.

51. Enamorado T, Fifield B, Imai K. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. Am Polit Sci Rev. 2019 May;113(2):353–71.

52. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969 Dec;64(328):1183–210.

53. Sariyar M, Borg A. The RecordLinkage Package: Detecting Errors in Data. R J. 2010;2(2):61.

54. Jaro MA. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1989;84(406):414–20.

55. Contiero P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P, et al. The EpiLink record linkage software: presentation and results of linkage test on cancer registry files. Methods Inf Med. 2005;44(1):66–71.

56. Hejblum BP, Weber GM, Liao KP, Palmer NP, Churchill S, Shadick NA, et al. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. Sci Data. 2019 Jan 8;6(1):180298.

57. Roman-Urrestarazu A, Yang JC, van Kessel R, Warrier V, Dumas G, Jongsma H, et al. Autism incidence and spatial analysis in more than 7 million pupils in English schools: a retrospective, longitudinal, school registry study. Lancet Child Adolesc Health. 2022 Dec 1;6(12):857–68.

58. Biblioteca del Congreso Nacional de Chile. www.bcn.cl/leychile. 2023 [cited 2023 Jul 12]. LEY 21545 Establece la promoción de la inclusión, la atención integral, y la protección de los derechos de las persons con trastorno del espectro autista en el ámbito social, de salud y educatión. Available from: https://www.bcn.cl/leychile

59. Instituto Nacional de Estadisticas. Default. [cited 2023 Jul 5]. Proyecciones de Población. Available from: http://www.ine.gob.cl/estadisticas/sociales/seguridad-publica-y-justicia/estadisticas-policiales-y-judiciales/proyecciones-de-poblaciÃşn

60. Humanitarian Data Exchange. Chile - Subnational Administrative Boundaries [Internet]. [cited 2023 Jul 5]. Available from: https://data.humdata.org/dataset/cod-ab-chl

61. Fay MP, Feuer EJ. Confidence intervals for directly standardised rates: a method based on the gamma distribution. Stat Med. 1997 Apr 15;16(7):791–801.

62. Sadinle M. Bayesian Estimation of Bipartite Matchings for Record Linkage. 2016 [cited 2023 Jun 28]; Available from: https://arxiv.org/abs/1601.06630

63. Stringham T. Fast Bayesian Record Linkage With Record-Specific Disagreement Parameters. J Bus Econ Stat. 2022 Oct 2;40(4):1509–22.

64. Pita R, Mendonça E, Reis S, Barreto M, Denaxas S. A Machine Learning Trainable Model to Assess the Accuracy of Probabilistic Record Linkage. In: Bellatreche L, Chakravarthy S, editors. Big Data Analytics and Knowledge Discovery [Internet]. Cham: Springer International Publishing; 2017 [cited 2023 Jun 28]. p. 214–27. (Lecture Notes in Computer Science; vol. 10440). Available from: http://link.springer.com/10.1007/978-3-319-64283-3_16

65. Yeargin-Allsopp M, Rice C, Karapurkar T, Doernberg N, Boyle C, Murphy C. Prevalence of Autism in a US Metropolitan Area. JAMA. 2003 Jan 1;289(1):49–55.

66. Sainsbury WJ, Carrasco K, Whitehouse AJO, Waddington H. Parent-reported Early Atypical Development and Age of Diagnosis for Children with Co-occurring Autism and ADHD. J Autism Dev Disord. 2023 Jun 1;53(6):2173–84.

67. Lindblom A. Under-detection of autism among First Nations children in British Columbia, Canada. Disabil Soc. 2014 Sep 14;29(8):1248–59.

68. Rojas F. Poverty determinants of acute respiratory infections among Mapuche indigenous peoples in Chile's Ninth Region of Araucania, using GIS and spatial statistics to identify health disparities. Int J Health Geogr. 2007 Jul 2;6(1):26.

10 Appendix A | R code

```
library(janitor)
library(gridExtra)
library(readxl)
library(writexl)
library(png)
library(viridis)
library(rgbif)
library(chilemapas)
library(leaflet)
library(plotly)
library(knitr)
library(kableExtra)
library(sf)
library(sp)
library(broom)
library(psych)
library(Hmisc)
library(poolr)
library(epitools)
library(mltools)
library(ggrepel)
library(rjags)
library(rstan)
library(posterior)
library(tidybayes)
library(bayesplot)
library(reclin2)
library(lubridate)
library(RecordLinkage)
library(dgof)
library(fdm2id)
library(ppclust)
library(factoextra)
library(FactoMineR)
library(tidyverse)

### Import geographic data

# Import region and commune boundaries, https://data.humdata.org/dataset/cod-ab-chl
chile.adm1 <- st_read("04_Data/CHL_adm_humdata/chl_admbnda_adm1_bcn_20211008.shp", quiet = TRUE)
chile.adm3 <- st_read("04_Data/CHL_adm_humdata/chl_admbnda_adm3_bcn_20211008.shp", quiet = TRUE) %>%
  mutate(commune_code = str_sub(ADM3_PCODE, start = 3, end = -1),
         commune_name = ADM3_ES)

# Get communes with health services
chile_communes_raw <- read_excel("04_Data/commune_by_health_service.xlsx") %>%
  clean_names()

chile_communes <- chile_communes_raw %>%
  mutate(commune_name = ifelse(comuna == "La Calera", "Calera",
                        ifelse(comuna == "Coihaique", "Coyhaique",
```

```
                        ifelse(comuna == "Paiguano", "Paihuano",
                        ifelse(comuna == "Pedro Aguirre Cerda", "Pedro Aguirre Cerda",
                        comuna))))) %>%
rename("health_service_name" = "servicio_de_salud") %>%
  select(commune_name, health_service_name)

# Create a number for each health service
health_service_lookup <- chile_communes %>%
  group_by(health_service_name) %>%
  summarise() %>%
  mutate(health_service_name_shortish =
           ifelse(grepl(" Del ", health_service_name),
                  str_sub(health_service_name, start = 23, end = -1),
                  str_sub(health_service_name, start = 19, end = -1)),
         health_service_name_short =
           ifelse(health_service_name_shortish == "Libertador B.O'Higgins", "O'Higgins",
           ifelse(health_service_name_shortish == "Metropolitano Central", "Metro. Central",
           ifelse(health_service_name_shortish == "Metropolitano Norte", "Metro. Norte",
           ifelse(health_service_name_shortish == "Metropolitano Occidente", "Metro. Occidente",
           ifelse(health_service_name_shortish == "Metropolitano Oriente", "Metro. Oriente",
           ifelse(health_service_name_shortish == "Metropolitano Sur", "Metro. Sur",
           ifelse(health_service_name_shortish == "Metropolitano Sur Oriente", "Metro. Sur Oriente",
           ifelse(health_service_name_shortish == "Valparaíso San Antonio", "Valparaíso",
           ifelse(health_service_name_shortish == "Viña del Mar Quillota", "Viña del Mar",
           health_service_name_shortish)))))))))) %>%
  arrange(health_service_name_short) %>%
  rowid_to_column("health_service_code")

# Get region codes with region name abbreviations
region_abr_lookup <- data.frame(
  ADM1_PCODE = c(paste0("CL0", c(1:9)), paste0("CL", c(10:16))),
  region_name_abr = c("TPCA", "ANTOF", "ATCMA", "COQ", "VALPO", "LGBO", "MAULE", "BBIO",
                      "ARAUC", "LAGOS", "AYSEN", "MAG", "RM", "RIOS", "AYP", "NUBLE"))

# Put region, health service and communes together
region_service_commune_lookup <- chile.adm3 %>%
  merge(chile_communes, by = "commune_name", all = TRUE) %>%
  mutate(commune_code = ifelse(commune_name == "Antártica", "12202", commune_code),
         ADM3_PCODE = ifelse(commune_name == "Antártica", "CL12202", ADM3_PCODE),
         ADM1_ES = ifelse(commune_name == "Antártica",
                          "Región de Magallanes y Antártica Chilena", ADM1_ES),
         ADM1_PCODE = ifelse(commune_name == "Antártica", "CL12", ADM1_PCODE),
         commune_name = ifelse(ADM3_PCODE == "CL01401", "Pozo Almonte", commune_name),
           # For the Tocopilla in Tarapaca, use it's old name to avoid duplication issues
         health_service_name =
           ifelse(ADM3_PCODE == "CL01401", "Servicio de Salud Iquique",
           ifelse(ADM3_PCODE == "CL11201", "Servicio de Salud Aisén",
           ifelse(ADM3_PCODE == "CL05201", "Servicio de Salud Valparaíso San Antonio",
            # Isla del Pascua is actually served by Metro Oriente but maps to Valparaiso
           ifelse(ADM3_PCODE == "CL08206", "Servicio de Salud Arauco",
           ifelse(ADM3_PCODE == "CL08301", "Servicio de Salud Biobío",
           ifelse(ADM3_PCODE == "CL06204", "Servicio de Salud Del Libertador B.O'Higgins",
           ifelse(ADM3_PCODE == "CL13121", "Servicio de Salud Metropolitano Sur",
```

```
           ifelse(ADM3_PCODE == "CL16206", "Servicio de Salud Ñuble",
                  health_service_name))))))),
         region_code = str_sub(ADM1_PCODE, start = 3, end = -1),
         commune_name_upper = toupper(commune_name)) %>%
  rename(region_name = ADM1_ES) %>%
  filter(!is.na(ADM3_PCODE)) %>%
    # Get rid of the communes that didn't match properly with health service
    #lookup and were corrected above
  merge(region_abr_lookup, by = "ADM1_PCODE")

region_service_commune_lookup <- merge(region_service_commune_lookup,
                                       health_service_lookup,
                                       by = "health_service_name", all = TRUE) %>%
  rename(region_name_long = region_name,
         health_service_name_long = health_service_name,
         health_service_name = health_service_name_short) %>%
  mutate(region_name = ifelse(region_name_long == "Región Metropolitana de Santiago",
                              "Metropolitana de Santiago",
                       ifelse(grepl("Región de ", region_name_long),
                              substr(region_name_long, start = 11,
                                     stop = nchar(region_name_long)),
                       ifelse(grepl("Región del ", region_name_long),
                              substr(region_name_long, start = 12,
                                     stop = nchar(region_name_long)), NA)))) %>%
  select(region_name_long, region_name, region_name_abr, region_code, ADM1_PCODE,
         commune_name, commune_name_upper, commune_code,
         health_service_name_long, health_service_name_shortish,
         health_service_name, health_service_code,
         geometry)

# Aggregate to region level (no )
region_lookup <- region_service_commune_lookup %>%
  as.tibble() %>%
  select(-geometry) %>%
  group_by(region_code, ADM1_PCODE, region_name, region_name_abr) %>%
  summarise()

region_service_lookup <- region_service_commune_lookup %>%
  as.tibble() %>%
  select(-geometry) %>%
  group_by(region_code, region_name, health_service_code,
           health_service_name_shortish, health_service_name) %>%
  summarise()

# Extract communes for the Araucania Sur and Norte health services
araucnorte_communes <- region_service_commune_lookup %>%
  filter(str_detect(health_service_name, "a Norte")) %>%
  as.tibble() %>%
  select(commune_code, commune_name)
araucsur_communes <- region_service_commune_lookup %>%
  filter(str_detect(health_service_name, "a Sur")) %>%
  as.tibble() %>%
  select(commune_code, commune_name)
```

```
### Import standard population data

chile_stdpop_raw <- read_excel("04_Data/pop_chile_2021_single_age.xlsx") %>%
  clean_names()

chile_stdpop <- chile_stdpop_raw %>%
  filter(sex != 9) %>%
  rename("std_pop" = "pop_2021") %>%
  mutate(pop_prop = std_pop / sum(std_pop))

### Import Chile demographic data

# https://www.ine.gob.cl/estadisticas/sociales/demografia-y-vitales/proyecciones-de-poblacion
chile_regionpop_raw <- read_excel("04_Data/ine_estimaciones-y-proyecciones-2002-2035_base-2017_region_a
                                  sheet = "Region summary 2021") %>%
  clean_names()

chile_regionpop <- chile_regionpop_raw %>%
  mutate(region_name = ifelse(region == "Aysén", "Aysén del Gral. Ibañez del Campo",
                         ifelse(region == "Biobío", "Bío-Bío",
                         ifelse(region == "Magallanes", "Magallanes y Antártica Chilena",
                         ifelse(region == "Metropolitana", "Metropolitana de Santiago",
                         ifelse(region == "OHiggins", "Libertador Bernardo O'Higgins",
                                region)))))) %>%
  select(-region)

# https://www.ine.gob.cl/estadisticas/sociales/ingresos-y-gastos/encuesta-suplementaria-de-ingresos
chile_regionincome_raw <- read_excel("04_Data/ingreso-medio-mensual-por-region-2010---2021.xlsx",
                                     skip = 2) %>%
  clean_names()

chile_regionincome <- chile_regionincome_raw %>%
  filter(ano == 2021,
         region != "Nacional") %>%
  mutate(region_short = ifelse(grepl("Región de ", region),
                               substr(region, start = 11, stop = nchar(region)),
                        ifelse(grepl("Región del ", region),
                               substr(region, start = 12, stop = nchar(region)),
                        ifelse(region == "Región Metropolitana", "Metropolitana", NA))),
         region_name = ifelse(region_short == "Aysén", "Aysén del Gral. Ibañez del Campo",
                         ifelse(region_short == "Biobío", "Bío-Bío",
                         ifelse(region_short == "Magallanes", "Magallanes y Antártica Chilena",
                         ifelse(region_short == "Metropolitana", "Metropolitana de Santiago",
                         ifelse(region_short == "O'Higgins", "Libertador Bernardo O'Higgins",
                                region_short)))))) %>%
  select(region_name, ingreso_medio_nominal)

# http://www.ine.gob.cl/estadisticas/sociales/censos-de-poblacion-y-vivienda/censo-de-poblacion-y-vivie
chile_regionrural_raw <- read_excel("04_Data/1_2_poblacion.xls", sheet = "Sheet1") %>%
  clean_names()

chile_regionrural <- chile_regionrural_raw %>%
  mutate(pop = urban + rural,
```

```
         urban_perc = urban / pop * 100,
         rural_perc = rural / pop * 100)

### Import school data

chile_merged_raw <- read.csv("04_Data/Data_Chile_Merge.csv") %>% clean_names()

chile_merged <- chile_merged_raw %>%
  rename(sex_desc = sex,
         year = agno,
         school_code = rbd,
         school_check_code = dgv_rbd,
         school_name = nom_rbd,
         school_region_code = cod_reg_rbd,
         school_region_name_abr = nom_reg_rbd_a,
         school_province_code = cod_pro_rbd,
         school_commune_code = cod_com_rbd_2,
         school_commune_name = nom_com_rbd_2,
         school_dept_code = cod_deprov_rbd,
         school_dept_name = nom_deprov_rbd,
         school_dependency_code = cod_depe, # has categories 1-6, no1 and no2 here are no1 in grouped
         school_dependency_code_grouped = cod_depe2, # has categories 1-5
         school_rurality_code = rural_rbd,
         school_operation_status = estado_estab,
         teaching_code1 = cod_ense, # min = 10, max = 910, eg preschool, special education hearing impa
         teaching_code2 = cod_ense2, # subject matter coding, 1-8
         teaching_code3 = cod_ense3, # age based coding, 1-7
         grade_code1 = cod_grado, # grade of schooling, 1-10, 21-25, 31-34, nests in teaching_code1
         grade_code2 = cod_grado2, # equivalent grade of schooling for adult special education, 1-8, 99
         grade_letter = let_cur, # refers to the class within the grade, close to start of alphabet is
         course_timing = cod_jor, # time of day, morning, afternoon, both, night, no info
         course_type = cod_tip_cur, # 0 = simple course, 1-4 = combined course, 99 = no info
         course_descr = cod_des_cur, # Description of course (TP secondary education only). 0: Does not
         student_id = mrun,
         sex = gen_alu, # 0 = no info, 1 = male, 2 = female
         dob = fec_nac_alu_2, # The second one has DD
         age_june30 = edad_alu, # age at 30th June 2021
         special_needs_status = int_alu, # integrated student indicator, 0 = no, 1 = yes. Mostly no
         special_needs_code = cod_int_alu, # ADHD, blindness, etc. 0 = none. 105 = autism, 203 = ADHD.
         student_region_code = cod_reg_alu,
         student_commune_code = cod_com_alu,
         student_commune_name = nom_com_alu,
         economic_sector_code = cod_sec,
         economic_specialty_code = cod_espe,
         economic_branch_code = cod_rama,
         economic_profspec_code = cod_men,
         teaching_code_new = ens)

### Restructure data

chile_bayes <- chile_merged %>%
  filter(age_june30 >= 6 & age_june30 <= 18,
         sex != 0) %>%
```

```
  mutate(autism = ifelse(special_needs_code == 105, 1, 0),
         adhd = ifelse(special_needs_code == 203, 1, 0),
         age_cat_name = factor(ifelse(age_june30 <= 8, "6-8",
                                      ifelse(age_june30 <= 11, "9-11",
                                      ifelse(age_june30 <= 14, "12-14", "15-18"))),
                                      levels = c("6-8", "9-11", "12-14", "15-18")),
         ethnic_2_group = factor(ifelse(ethnicity == "Mapuche", "Mapuche",
                                 ifelse(ethnicity == "No Native Group" | ethnicity == "No registry",
                                        "No Indigenous group",
                                        "Other Indigenous group")),
                                 levels = c("Mapuche", "Other Indigenous group", "No Indigenous group")),
         school_rurality = ifelse(school_rurality_code == 0, "Urban", "Rural"),
         school_fee = factor(ifelse(school_fee == "", "No information",
                                ifelse(school_fee == "GRATUITO", "Free",
                                ifelse(school_fee == "$1.000 A $10.000", "$1,000-$10,000",
                                ifelse(school_fee == "$10.001 A $25.000", "$10,001-$25,000",
                                ifelse(school_fee == "$25.001 A $50.000", "$25,001-$50,000",
                                ifelse(school_fee == "$50.001 A $100.000", "$50,001-$100,000",
                                ifelse(school_fee == "MAS DE $100.000", "$100,001+",
                                ifelse(school_fee == "SIN INFORMACION", "No information", NA)))))))),
                                levels = c("Free", "$1,000-$10,000", "$10,001-$25,000",
                                           "$25,001-$50,000", "$50,001-$100,000",
                                           "$100,001+", "No information")),
         commune_code_temp = ifelse(student_commune_code == "0", school_commune_code,
                                    student_commune_code),
                                # If there is no commune code for the student, use their school's code
         commune_code_old = ifelse(nchar(as.character(commune_code_temp)) == 4,
                                   paste0("0", as.character(commune_code_temp)),
                                   as.character(commune_code_temp)),
         # Fix Biobio to Nuble communes
         commune_code = ifelse(commune_code_old == "08401", "16101",
                        ifelse(commune_code_old == "08402", "16102",
                        ifelse(commune_code_old == "08403", "16202",
                        ifelse(commune_code_old == "08404", "16203",
                        ifelse(commune_code_old == "08405", "16302",
                        ifelse(commune_code_old == "08406", "16103",
                        ifelse(commune_code_old == "08407", "16104",
                        ifelse(commune_code_old == "08408", "16204",
                        ifelse(commune_code_old == "08409", "16303",
                        ifelse(commune_code_old == "08410", "16105",
                        ifelse(commune_code_old == "08411", "16106",
                        ifelse(commune_code_old == "08412", "16205",
                        ifelse(commune_code_old == "08413", "16107",
                        ifelse(commune_code_old == "08414", "16201",
                        ifelse(commune_code_old == "08415", "16206",
                        ifelse(commune_code_old == "08416", "16301",
                        ifelse(commune_code_old == "08417", "16304",
                        ifelse(commune_code_old == "08418", "16108",
                        ifelse(commune_code_old == "08419", "16305",
                        ifelse(commune_code_old == "08420", "16207",
                        ifelse(commune_code_old == "08421", "16109",
                               commune_code_old))))))))))))))))))))))) %>%
  left_join(region_service_commune_lookup %>%
```

```
              as.tibble() %>%
              select(-geometry), by = "commune_code") %>%
  mutate(ssas = ifelse(commune_code %in% araucsur_communes$commune_code, 1, 0)) %>%
  select(region_code,
    region_name,
    region_name_abr,
    commune_code, # student, see chile_merged
    student_commune_name, # upper case, from chile_merged
    commune_name, # Sentence case, from lookup table
    health_service_code,
    health_service_name,
    sex,
    sex_desc,
    dob,
    age_june30,
    age_cat_name,
    school_rurality_code,
    school_rurality,
    pago_matricula,
    pago_mensual,
    school_fee,
    ethnicity,
    mapuche,
    nationality,
    ethnic_3_group,
    ethnic_2_group,
    special_needs_status,
    special_needs_code,
    autism,
    adhd,
    ssas
  )

chile_bayes_aut <- chile_bayes %>% select(-adhd)
chile_bayes_adhd <- chile_bayes %>% select(-autism)

mapuche_regions <- c("09", "11", "08", "10", "14", "12", "13")
chile_bayes_aut_ethnic <- chile_bayes_aut %>%
  filter(region_code %in% mapuche_regions)
chile_bayes_adhd_ethnic <- chile_bayes_adhd %>%
  filter(region_code %in% mapuche_regions)

### Import clinical data

clinical_large_raw <- read_excel("04_Data/dataset_ssas_2015_2021.xlsx") %>%
  clean_names

clinical_large <- clinical_large_raw %>%
  select(c(-procedence, -ethnicity, -education_level, -disability, -foster_care)) %>%
  # Fix the date columns
  mutate(dob_eng = ifelse(str_detect(date_of_birth, "/"), 1,
                   ifelse(str_detect(date_of_birth, "-"), 0, NA)),
         apt_eng = ifelse(str_detect(date_appointment, "/"), 1,
```

```
                  ifelse(str_detect(date_appointment, "-"), 0, NA)),
       dob_day = ifelse(dob_eng == 1, as.integer(str_extract(date_of_birth, "^\\d+")),
                 ifelse(dob_eng == 0, as.integer(str_extract(date_of_birth, "^\\d+")), NA)),
       dob_month = ifelse(dob_eng == 1, as.integer(str_extract(date_of_birth, "(?<=/)\\d+(?=/)")),
                   ifelse(dob_eng == 0, str_extract(date_of_birth, "(?<=-)\\w+(?=-)"), NA)),
       dob_year = ifelse(dob_eng == 1, as.integer(str_extract(date_of_birth, "\\d+$")),
                  ifelse(dob_eng == 0, as.integer(str_extract(date_of_birth, "\\d+$")) + 2000, NA)),
       dob_month_eng = as.integer(ifelse(dob_month == "ene", 1,
                                  ifelse(dob_month == "abr", 4,
                                  ifelse(dob_month == "ago", 8,
                                  ifelse(dob_month == "sept", 9,
                                  ifelse(dob_month == "dic", 12, dob_month)))))),
       dob = make_date(year = dob_year, month = dob_month_eng, day = dob_day),
       apt_day = ifelse(apt_eng == 1, as.integer(str_extract(date_appointment, "^\\d+")),
                 ifelse(apt_eng == 0, as.integer(str_extract(date_appointment, "^\\d+")), NA)),
       apt_month = ifelse(apt_eng == 1, as.integer(str_extract(date_appointment, "(?<=/)\\d+(?=/)")),
                   ifelse(apt_eng == 0, str_extract(date_appointment, "(?<=-)\\w+(?=-)"), NA)),
       apt_year = ifelse(apt_eng == 1, as.integer(str_extract(date_appointment, "\\d+$")),
                  ifelse(apt_eng == 0, as.integer(str_extract(date_appointment, "\\d+$")) + 2000, NA)),
       apt_month_eng = as.integer(ifelse(apt_month == "ene", 1,
                                  ifelse(apt_month == "abr", 4,
                                  ifelse(apt_month == "ago", 8,
                                  ifelse(apt_month == "sept", 9,
                                  ifelse(apt_month == "dic", 12, apt_month)))))),
       apt_date = make_date(year = apt_year, month = apt_month_eng, day = apt_day),
       age_june30 = trunc(time_length(interval(ymd(dob), ymd("2021-06-30")), unit = "year")),
       commune_name_upper = ifelse(comuna == "CHOL CHOL", "CHOLCHOL",
                            ifelse(comuna == "CURACAUTIN", "CURACAUTÍN",
                            ifelse(comuna == "PITRUFQUEN", "PITRUFQUÉN",
                            ifelse(comuna == "PUCON", "PUCÓN",
                            ifelse(comuna == "TOLTEN", "TOLTÉN",
                            ifelse(comuna == "VILCUN", "VILCÚN", comuna)))))),
       ses_status = ifelse(socio_economic_level == "FONASA - A", 1,
                    ifelse(socio_economic_level == "FONASA - B", 2,
                    ifelse(socio_economic_level == "FONASA - C", 2,
                    ifelse(socio_economic_level == "FONASA - D", 2,
                    ifelse(socio_economic_level == "Private Health Insurance", 3,
                    ifelse(socio_economic_level %in% c("COLMENA GOLDEN CROSS",
                                                       "RIO BLANCO",
                                                       "CARABINEROS (DIPRECA)",
                                                       "BANMEDICA S.A.",
                                                       "PARTICULAR (SIN PREVISION)",
                                                       "VIDA TRES"), 3, NA)))))),
       autism = 1,
       intdisab = 0,
       aut_rank = 1
       ) %>%
left_join(region_service_commune_lookup, by = "commune_name_upper") %>%
select(id, gender,
       commune_code, commune_name, commune_name_upper,
       health_service_name, region_name,
       socio_economic_level, ses_status,
       dob, age_june30,
```

```
         apt_date, hospital, medical_specialty, type_appointment,
         autism, intdisab, aut_rank)

aut_codes <- unique(clinical_large_raw$codigo)

clinical_small_raw <- read_excel("04_Data/Dataset_Vill_2014_2021.xlsx", col_names = TRUE) %>%
  clean_names()

clinical_small <- clinical_small_raw %>%
  rename("dob" = "fecha_nacimiento",
         "apt_date" = "fecha_ejecutada",
         "type_appointment" = "appoinment",
         "diagnosis" = "diagnostico_1") %>%
  mutate(gender = str_to_title(gender),
         autism = ifelse(cod_dg_1 %in% aut_codes |
                           cod_dg_2 %in% aut_codes |
                           cod_dg_3 %in% aut_codes, 1, 0),
         aut_rank = ifelse(cod_dg_1 %in% aut_codes, 1,
                    ifelse(cod_dg_2 %in% aut_codes, 2,
                    ifelse(cod_dg_3 %in% aut_codes, 3, NA))),
         age_june30 = trunc(time_length(interval(ymd(dob), ymd("2021-06-30")), unit = "year")),
         commune_name_upper = ifelse(comuna == "CHOL CHOL", "CHOLCHOL",
                       ifelse(comuna == "CURACAUTIN", "CURACAUTÍN",
                       ifelse(comuna == "PITRUFQUEN", "PITRUFQUÉN",
                       ifelse(comuna == "PUCON", "PUCÓN",
                       ifelse(comuna == "TOLTEN", "TOLTÉN",
                       ifelse(comuna == "VILCUN", "VILCÚN",
                       ifelse(comuna == "DIEGO DE ALMAGRO  (#)", "DIEGO DE ALMAGRO",
                       ifelse(comuna == "MACHALI", "MACHALÍ",
                       ifelse(comuna == "TEMUCO (##)", "TEMUCO", comuna)))))))),
         ses_status = ifelse(socio_economic_level == "FONASA - A", 1,
                      ifelse(socio_economic_level == "FONASA - B", 2,
                      ifelse(socio_economic_level == "FONASA - C", 2,
                      ifelse(socio_economic_level == "FONASA - D", 2,
                      ifelse(socio_economic_level == "Private Health Insurance", 3,
                      ifelse(socio_economic_level %in% c("COLMENA GOLDEN CROSS",
                                                         "RIO BLANCO",
                                                         "CARABINEROS (DIPRECA)",
                                                         "BANMEDICA S.A.",
                                                         "PARTICULAR (SIN PREVISION)",
                                                         "VIDA TRES"), 3, NA)))))) %>%
  left_join(region_service_commune_lookup, by = "commune_name_upper") %>%
  filter(autism == 1) %>%
  select(id, gender, commune_code, commune_name, commune_name_upper,
         health_service_name, region_name, socio_economic_level, ses_status,
         dob, age_june30, apt_date, hospital, medical_specialty, type_appointment,
         cod_dg_1, cod_dg_2, cod_dg_3, diagnosis, autism, aut_rank)

intdisab_codes <- unique(c(clinical_small_raw$cod_dg_1, clinical_small_raw$cod_dg_2,
                           clinical_small_raw$cod_dg_3)) %>%
  str_subset("F7") %>%
  sort()
```

```r
clinical_small <- clinical_small %>%
  mutate(intdisab = ifelse(cod_dg_1 %in% intdisab_codes |
                             cod_dg_2 %in% intdisab_codes |
                             cod_dg_3 %in% intdisab_codes, 1, 0)) %>%
  select(c(-cod_dg_1, -cod_dg_2, -cod_dg_3, -diagnosis))

clinical <- rbind(clinical_large, clinical_small)

clinical_communes <- clinical %>%
  group_by(commune_code) %>%
  summarise() %>%
  arrange() %>%
  mutate(commune_in_school_data = ifelse(commune_code %in% unique(chile_merged$commune_code)
                                         , 1, 0))

### Define functions to help with Bayesian analysis

get_grouped_prev_plot <- function(x, grouping_vars, disease) {
  # Calculates sample prevalence and its confidence intervals for supplied feature grouping
  # x = chile_bayes_aut or chile_bayes_adhd, needs columns called autism or adhd, and count
  # grouping_vars = variables in x to group by
  # disease = autism or adhd

  x_grouped <- x %>%
    group_by(across(all_of(grouping_vars))) %>%
    summarise(count = n()) %>%
    pivot_wider(names_from = disease, values_from = count) %>%
    rename("n_nodisease" = "0", "n_disease" = "1") %>% #, "age" = "age_june30") %>%
    mutate(n_disease = ifelse(is.na(n_disease), 0, n_disease),
            # If there are no cases of autism in the group, input 0
           sample_pop_size = n_nodisease + n_disease,
            # Total sample population is autism cases + not cases
           sample_prevalence = n_disease / sample_pop_size,
            # Prevalence of autism in the group
           ci_lower = sample_prevalence - (1.96 * sqrt(sample_prevalence *
                                                         (1 - sample_prevalence) /
                                                         sample_pop_size)),
           ci_upper = sample_prevalence + (1.96 * sqrt(sample_prevalence *
                                                         (1 - sample_prevalence) /
                                                         sample_pop_size)),
           ci_lower = ifelse(ci_lower < 0, 0, ci_lower),
           prev_ci = paste0(sprintf("%.2f", round(sample_prevalence * 100, 2)),
                            " (",
                            sprintf("%.2f", round(ci_lower * 100, 2)),
                            ", ",
                            sprintf("%.2f", round(ci_upper * 100, 2)),
                            ")")) %>%
    ungroup()
  return(x_grouped)
}

get_grouped_prev <- function(x, stdpop, grouping_vars, disease) {
  # Calculates sample prevalence, age- and sex-standardised prevalence and
```

```
  # group weighting for supplied feature grouping
  # x = chile_bayes_aut, needs columns called autism, count
  # stdpop = standard population with age and sex counts
  # grouping_vars = variables in x to group by
  # disease = autism or adhd

  n_stdpop <- sum(stdpop$std_pop)

  x_grouped <- x %>%
    group_by(across(all_of(grouping_vars))) %>%
    summarise(count = n()) %>%
    pivot_wider(names_from = disease, values_from = count) %>%
    rename("n_nodisease" = "0", "n_disease" = "1", "age" = "age_june30") %>%
    mutate(n_disease = ifelse(is.na(n_disease), 0, n_disease),
             # If there are no cases of autism in the group, input 0
           sample_pop_size = n_nodisease + n_disease,
             # Total sample population is autism cases + not cases
           sample_prevalence = n_disease / sample_pop_size) %>%
             # Prevalence of autism in the group
    left_join(stdpop, by = c("age", "sex")) %>%
    mutate(disease_prev_std = n_disease / sample_pop_size * pop_prop,
             # Prevalence of autism in the group, standardised to standard population
           w = std_pop / (sample_pop_size * n_stdpop),
             # Weight of the group using standard population
           w2 = pop_prop / sample_pop_size,
           #sum_std_pop = sum(std_pop)
           ) %>%
    ungroup()
  return(x_grouped)
}

get_adjusted_prev <- function(x, grouping_vars) {
  # Turns grouped prevalence into age- and sex- adjusted prevalence with
  # Fay and Feuer Gamma confidence intervals
  # x = output from get_grouped_prev
  x_adj <- x %>%
  group_by(across(all_of(grouping_vars))) %>%
  summarise(sum_sample_pop_size = sum(sample_pop_size),
            crude_rate = sum(n_disease) / sum(sample_pop_size),
            crude_count = sum(n_disease),
            adjusted_rate = sum(n_disease / sample_pop_size * pop_prop),
            adjusted_count = round(adjusted_rate * sum_sample_pop_size, 0),
            var = sum(pop_prop^2 * n_disease / sample_pop_size^2),
            w_M = max(w),
            crude_ci_lower = crude_rate - (1.96 * sqrt(crude_rate * (1 - crude_rate) /
                                                       sum_sample_pop_size)),
            crude_ci_upper = crude_rate + (1.96 * sqrt(crude_rate * (1 - crude_rate) /
                                                       sum_sample_pop_size)),
            adjusted_ci_lower = ifelse(var == 0, 0, var / (2*adjusted_rate) *
               qchisq(p = 0.05/2, df = 2*adjusted_rate^2 / var)),
            adjusted_ci_upper = (var + w_M^2) / (2*(adjusted_rate + w_M)) *
               qchisq(p = 1-0.05/2, df = 2*(adjusted_rate+w_M)^2 / (var+w_M^2)),
            crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
```

```
            adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower),
            crude_prev_ci = paste0(sprintf("%.2f", round(crude_rate * 100, 2)),
                                   " (",
                                   sprintf("%.2f", round(crude_ci_lower * 100, 2)),
                                   ", ",
                                   sprintf("%.2f", round(crude_ci_upper * 100, 2)),
                                   ")"),
            adjusted_prev_ci = paste0(sprintf("%.2f", round(adjusted_rate * 100, 2)),
                                   " (",
                                   sprintf("%.2f", round(adjusted_ci_lower * 100, 2)),
                                   ", ",
                                   sprintf("%.2f", round(adjusted_ci_upper * 100, 2)),
                                   ")")) %>%
  arrange(across(all_of(grouping_vars)))
}

do_jags_rand_model <- function(x, feat, model, theta_mu, theta_sigma, pars,
                               nBurn = 1000, nIter = 1000, convergence_checks = FALSE) {
  # x = output from get_adjusted_prev.
  # x needs to have columns sum_sample_pop_size, adjusted_count
  # feat = feature being used as random effect
  # model = JAGS random effects model
  # theta_mu, theta_sigma = mean and sd of beta prior distribution
  # pars = model parameters to report
  # nBurn = number of burn-in samples
  # nIter = number of posterior iterations

  nFeat <- length(unique(x[[feat]]))
  FeatNames <- sort(unique(x[[feat]]))

  # Define beta prior
  theta_a <- theta_mu * (theta_mu * (1-theta_mu) / theta_sigma^2 - 1)
  theta_b <- (1 - theta_mu) * (theta_mu * (1-theta_mu) / theta_sigma^2 - 1)

  # Initial values for model chains
  rand_ini <- list(list(theta = rep(0.001, nFeat)),
                   list(theta = rep(0.01, nFeat)))

  # Run JAGS model
  rand_data <- list(theta_a = theta_a,
                    theta_b = theta_b,
                    nObs = x$sum_sample_pop_size,
                    disease_sample = x$adjusted_count,
                    nFeat = nFeat)
  rand_jag <- jags.model(textConnection(model),
                         data = rand_data,
                         inits = rand_ini,
                         n.chains = 2,
                         quiet = TRUE)
  update(rand_jag, n.iter = nBurn)
  rand_sam <- coda.samples(model = rand_jag,
                           variable.names = pars,
                           n.iter = nIter)
```

```
  # Convergence checks
  if(convergence_checks) {
    print(mcmc_trace(rand_sam, paste0("theta[", 1:nFeat, "]")))
    print(mcmc_trace(rand_sam, paste0("disease_pred[", 1:nFeat, "]")))
    rand_summ <- summary(subset_draws(as_draws(rand_sam), pars),
                         ~quantile(.x, probs=c(0.025, 0.5, 0.975)),
                         ~mcse_quantile(.x, probs=c(0.025, 0.5, 0.975)),
                         "rhat") %>%
      arrange(desc(rhat))
    print(rand_summ)
  }

  # Extract posterior density
  prev_post <- as_tibble(as_draws_matrix(rand_sam), rownames = "Iteration") %>%
    select(c("Iteration", contains("theta["))) %>%
    pivot_longer(cols = contains("theta["),
                 names_to = "Feat",
                 values_to = "predicted_prev") %>%
    mutate(Feat_names = factor(Feat, levels = c(paste0("theta[",1:nFeat,"]")), labels = FeatNames)) %>%
    select(Iteration, Feat_names, predicted_prev)

  return(prev_post)
}


plot_post_density <- function(jags_post, sample_data, feat, theta_mu, theta_sigma,
                              disease, title_text = "") {
  # Plots posterior densities and their 95% credible intervals, and sample
  # prevalence confidence intervals
  # jags_post = output from do_jags_rand_model, ie posterior densities
  # sample data = output from get_adjusted_prev, ie sample prevalence with
  # confidence intervals
  # feat = the same feature used as the random effect in do_jags_rand_model
  # theta_mu, theta_sigma = mean and sd of beta prior distribution used in do_jags_rand_model
  # disease = autism or ADHD
  # title_text = any additional text for the title, eg grouing variables

  # calcuate posterior credible intervals
  post_ci <- jags_post %>%
  group_by(across(all_of(feat))) %>%
  summarise(post_lower = quantile(predicted_prev, 0.025),
            post_upper = quantile(predicted_prev, 0.975))

  print(ggplot() +
          geom_density(data = jags_post, aes(x = predicted_prev)) +
          geom_vline(data = post_ci, aes(xintercept = post_lower),
                     color = "blue", linetype = "dotted") +
          geom_vline(data = post_ci, aes(xintercept = post_upper),
                     color = "blue", linetype = "dotted") +
          geom_vline(data = sample_data, aes(xintercept = adjusted_ci_lower),
                     color = "red", linetype = "dashed") +
          geom_vline(data = sample_data, aes(xintercept = adjusted_ci_upper),
                     color = "red", linetype = "dashed") +
```

```
          facet_wrap(as.formula(paste0("~", feat))) +
          labs(title = paste0(disease, " prevalence, prior mean = ",
                              signif(theta_mu, 3), ", prior sd = ",
                              signif(theta_sigma, 3), title_text),
               x = "Posterior predictive distribution",
               y = "Density"))
}

### Define functions to help with data restructuring

# Define some functions to help with restructuring clinical data into patient level data
get.min.na <- function(x) ifelse( !all(is.na(x)), min(x, na.rm = TRUE), NA)

get.max.na <- function(x) ifelse( !all(is.na(x)), max(x, na.rm = TRUE), NA)

get.mode.na <- function(x) {
  ux <- na.omit(unique(x))
  ux[which.max(tabulate(match(x, ux)))]
}

get.mapuche.ethnicity <- function(x) {
  y <- NA
  if(any(x == "Mapuche")) {
    y <- "Mapuche"
  } else if(any(x == "Chilean")) {
    y <- "Chilean"
  } else if(any(x == "Foreign")) {
    y <- "Foreign"
  } else {
    y <- "No ethnicity information"
  }
  return(y)
}

get.yes.disability <- function(x) {
  y <- NA
  if(any(x == "Yes")) {
    y <- "Yes disability"
  } else if(any(x == "No")) {
    y <- "No disability"
  } else {
    y <- "No disability information"
  }
  return(y)
}

get.yes.fostercare <- function(x) {
  y <- NA
  if(any(x == "Yes")) {
    y <- "Yes foster care"
  } else if(any(x == "No")) {
    y <- "No foster care"
  } else {
```

```
    y <- "No foster care information"
  }
  return(y)
}

### Define plotting colours

plasma_colours <- viridis(10, option = "C")
viridis_colours <- viridis(20, option = "D")
rocket_colours <- viridis(20, option = "F")
turbo_colours <- viridis(17, option = "H")

### Regions to health services

kbl(region_service_lookup %>% as.tibble() %>% select(region_name, health_service_name_shortish) %>% arra
    col.names = c("Region", "Health services"),
    align = "ll",
    booktabs = TRUE,
    format = "latex",
    linesep = c("\\addlinespace", "\\addlinespace", "\\addlinespace", "\\addlinespace",
                "", "", "", "\\addlinespace", "\\addlinespace", "", "\\addlinespace",
                "\\addlinespace",  "", "", "\\addlinespace", "\\addlinespace",
                "\\addlinespace", "\\addlinespace", "", "", "", "", "",
                "\\addlinespace", "\\addlinespace", "", "", "\\addlinespace"),
    caption = "Chilean health services by region")

### Demographics maps

# Get region-level geometries with demographic data for plotting
demog_geom <- chile.adm1 %>%
  left_join(region_lookup, by = "ADM1_PCODE") %>%
  left_join(chile_regionpop, by = "region_name") %>%
  left_join(chile_regionincome, by = "region_name") %>%
  left_join(select(chile_regionrural, -region), by = "region_code")

# Income
ggplot(demog_geom) +
  geom_sf(mapping = aes(geometry = geometry, fill = ingreso_medio_nominal),
          linewidth = 0.001, color = "#555555") +
  geom_sf_text(mapping = aes(geometry = geometry, label = region_name), nudge_x = 9, size = 3.5) +
  scale_fill_viridis(option = "viridis", direction = 1, name = "Median income (Peso)") +
  labs(title = "Net income from main job") +
  theme_void() +
  xlab("Longitude") +
  ylab("Latitude") +
  coord_sf(xlim = c(-75, -55), ylim = c(-56, -18))

# Rurality
ggplot(demog_geom) +
  geom_sf(mapping = aes(geometry = geometry, fill = rural_perc),
          linewidth = 0.001, color = "#555555") +
  geom_sf_text(mapping = aes(geometry = geometry, label = region_name), nudge_x = 9, size = 3.5) +
  scale_fill_viridis(option = "viridis", direction = 1, name = "Rural population (%)") +
```

```
  labs(title = "Percentage of population living in rural areas") +
  theme_void() +
  xlab("Longitude") +
  ylab("Latitude") +
  coord_sf(xlim = c(-75, -55), ylim = c(-56, -18))

# ARAUC communes
arauc <- region_service_commune_lookup %>%
  filter(ADM1_PCODE == "CL09")

ggplot(arauc) +
  geom_sf(mapping = aes(geometry = geometry, fill = health_service_name), color = "grey") +
  geom_sf_text(mapping = aes(geometry = geometry, label = commune_name), size = 3.5) +
  coord_sf(xlim = c(-73.5, -70.8), ylim = c(-39.7, -37.6)) +
  scale_fill_manual(name = "Health service", values = c("#A6FDFD", "#D6FD75")) +
  theme_void() +
  theme(legend.position = "bottom") +
  labs(title = "La Araucanía communes by health service") +
  xlab("Longitude") +
  ylab("Latitude")

### School data content table

school.table <- chile_bayes %>%
  group_by(sex_desc) %>%
  summarise(variable = "Sex",
            count = n(),
            perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
  rename(values = sex_desc) %>%
  rbind(chile_bayes %>%
          group_by(age_cat_name) %>%
          summarise(variable = "Age band",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = age_cat_name),
        chile_bayes %>%
          group_by(health_service_name) %>%
          summarise(variable = "Health service",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = health_service_name),
        chile_bayes %>%
          group_by(school_fee) %>%
          summarise(variable = "School fee",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = school_fee),
        chile_bayes %>%
          group_by(ethnic_2_group) %>%
          summarise(variable = "Ethnicity",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = ethnic_2_group),
```

```
        chile_bayes %>%
          group_by(school_rurality) %>%
          summarise(variable = "Rurality",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = school_rurality),
        chile_bayes %>%
          group_by(special_needs_status) %>%
          summarise(variable = "Accesses SEED",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = special_needs_status) %>%
          mutate(values = factor(ifelse(values == 0, "No SEED", "Yes SEED"),
                                 levels = c("Yes SEED", "No SEED"))) %>%
          arrange(values),
        chile_bayes %>%
          group_by(autism) %>%
          summarise(variable = "Autism",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = autism) %>%
          mutate(values = factor(ifelse(values == 0, "No autism", "Yes autism"),
                                 levels = c("Yes autism", "No autism"))) %>%
          arrange(values),
        chile_bayes %>%
          group_by(adhd) %>%
          summarise(variable = "ADHD",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(chile_bayes)*100, 2))) %>%
          rename(values = adhd) %>%
          mutate(values = factor(ifelse(values == 0, "No ADHD", "Yes ADHD"),
                                 levels = c("Yes ADHD", "No ADHD"))) %>%
          arrange(values)) %>%
  mutate(count_perc = paste0(format(count, big.mark = ",", trim = TRUE), " (", perc, "%)"),
         variables_short = c("Sex", "",
                             "Age band", "", "", "",
                             "Health service", "", "", "", "", "", "", "", "", "",
                             "", "", "", "", "", "", "", "", "", "", "", "", "", "",
                             "", "", "", "", "",
                             "School fee", "", "", "", "", "", "",
                             "Ethnicity", "", "",
                             "Rurality", "",
                             "Accesses SEED", "",
                             "Autism", "",
                             "ADHD", "")) %>%
  select(variables_short, values, count_perc)

kbl(school.table,
    col.names = c("Feature", "Available values", "Count (%)"),
    align = "llr",
    booktabs = TRUE,
    longtable = TRUE,
    format = "latex",
```

```
    linesep = c("", "\\addlinespace", # Sex
                "", "", "", "\\addlinespace", # Age
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "",  "\\addlinespace", # Health service
                "", "", "", "", "", "", "\\addlinespace", # SES
                "", "", "\\addlinespace", # Ethnicity
                "", "\\addlinespace", # Rurality
                "", "\\addlinespace", # SEED
                "", "\\addlinespace", # Autism
                "", "\\addlinespace" # ADHD
                ),
    caption = "Count and percentage of features' values in the school dataset. Metro. is short for Metr
  kable_styling(latex_options = c("repeat_header"))

### Adjsuted prevalence

# Autism
aut_prev <- get_grouped_prev(x = chile_bayes_aut, stdpop = chile_stdpop,
                             grouping_vars = c("age_june30", "age_cat_name",
                                               "sex", "sex_desc", "autism"),
                             disease = "autism")

aut_prev_adj <- aut_prev %>%
  summarise(sum_sample_pop_size = sum(sample_pop_size),
            crude_rate = sum(n_disease) / sum(sample_pop_size),
            crude_count = sum(n_disease),
            adjusted_rate = sum(n_disease / sample_pop_size * pop_prop),
            adjusted_count = round(adjusted_rate * sum_sample_pop_size, 0),
            var = sum(pop_prop^2 * n_disease / sample_pop_size^2),
            w_M = max(w),
            crude_ci_lower = crude_rate - (1.96 * sqrt(crude_rate *
                                                       (1 - crude_rate) / sum_sample_pop_size)),
            crude_ci_upper = crude_rate + (1.96 * sqrt(crude_rate *
                                                       (1 - crude_rate) / sum_sample_pop_size)),
            adjusted_ci_lower = ifelse(var == 0, 0, var / (2*adjusted_rate) *
               qchisq(p = 0.05/2, df = 2*adjusted_rate^2 / var)),
            adjusted_ci_upper = (var + w_M^2) / (2*(adjusted_rate + w_M)) *
               qchisq(p = 1-0.05/2, df = 2*(adjusted_rate+w_M)^2 / (var+w_M^2))) %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# ADHD
adhd_prev <- get_grouped_prev(x = chile_bayes_adhd, stdpop = chile_stdpop,
                             grouping_vars = c("age_june30", "age_cat_name",
                                               "sex", "sex_desc", "adhd"),
                             disease = "adhd")

adhd_prev_adj <- adhd_prev %>%
  summarise(sum_sample_pop_size = sum(sample_pop_size),
```

```
            crude_rate = sum(n_disease) / sum(sample_pop_size),
            crude_count = sum(n_disease),
            adjusted_rate = sum(n_disease / sample_pop_size * pop_prop),
            adjusted_count = round(adjusted_rate * sum_sample_pop_size, 0),
            var = sum(pop_prop^2 * n_disease / sample_pop_size^2),
            w_M = max(w),
            crude_ci_lower = crude_rate - (1.96 * sqrt(crude_rate *
                                    (1 - crude_rate) / sum_sample_pop_size)),
            crude_ci_upper = crude_rate + (1.96 * sqrt(crude_rate *
                                    (1 - crude_rate) / sum_sample_pop_size)),
            adjusted_ci_lower = ifelse(var == 0, 0, var / (2*adjusted_rate) *
              qchisq(p = 0.05/2, df = 2*adjusted_rate^2 / var)),
            adjusted_ci_upper = (var + w_M^2) / (2*(adjusted_rate + w_M)) *
              qchisq(p = 1-0.05/2, df = 2*(adjusted_rate+w_M)^2 / (var+w_M^2))) %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

### Adjusted prevalence by age band

# Autism
aut_prev.agecat <- get_grouped_prev_plot(x = chile_bayes_aut,
                                         grouping_vars = c("age_cat_name", "autism"),
                                         disease = "autism") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = aut_prev.agecat) +
  geom_col(aes(x = age_cat_name, y = sample_prevalence*100, group = age_cat_name,
               fill = age_cat_name), position = position_dodge()) +
  geom_errorbar(aes(x = age_cat_name, ymin = ci_lower*100, ymax = ci_upper*100,
                    group = age_cat_name), width = 0.2, position = position_dodge(width = 0.9)) +
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5],
                               plasma_colours[7], plasma_colours[9])) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  labs(title = "Autism prevalence",
       x = "Age band",
       y = "Crude prevalence % (95% CI)",
       fill = "Age band")

# ADHD
adhd_prev.agecat <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                         grouping_vars = c("age_cat_name", "adhd"),
                                         disease = "adhd") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = adhd_prev.agecat) +
  geom_col(aes(x = age_cat_name, y = sample_prevalence*100, group = age_cat_name,
               fill = age_cat_name), position = position_dodge()) +
  geom_errorbar(aes(x = age_cat_name, ymin = ci_lower*100, ymax = ci_upper*100,
                    group = age_cat_name), width = 0.2, position = position_dodge(width = 0.9)) +
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5],
                               plasma_colours[7], plasma_colours[9])) +
  theme(axis.text.x = element_blank(),
```

```
        axis.ticks.x = element_blank()) +
  labs(title = "ADHD prevalence",
       x = "Age band",
       y = "Crude prevalence % (95% CI)",
       fill = "Age band")

### Adjusted prevalence by sex

# Seperate sexes so that standardisation is correct
chile_stdpop_f <- chile_stdpop %>%
  filter(sex == 2) %>%
  mutate(pop_prop = std_pop / sum(std_pop))
chile_stdpop_m <- chile_stdpop %>%
  filter(sex == 1) %>%
  mutate(pop_prop = std_pop / sum(std_pop))

# Autism
aut_prev_f <- chile_bayes_aut %>%
  filter(sex == 2) %>%
  get_grouped_prev(stdpop = chile_stdpop_f,
                   grouping_vars = c("age_june30", "sex", "autism"), disease = "autism")
aut_prev_adj_f <- get_adjusted_prev(aut_prev_f, grouping_vars = c()) %>%
  mutate(sex_desc = "Female") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

aut_prev_m <- chile_bayes_aut %>%
  filter(sex == 1) %>%
  get_grouped_prev(stdpop = chile_stdpop_m,
                   grouping_vars = c("age_june30", "sex", "autism"), disease = "autism")
aut_prev_adj_m <- get_adjusted_prev(aut_prev_m, grouping_vars = c()) %>%
  mutate(sex_desc = "Male") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

aut_prev_sex_adj <- rbind(aut_prev_adj_m, aut_prev_adj_f)

# ADHD
adhd_prev_f <- chile_bayes_adhd %>%
  filter(sex == 2) %>%
  get_grouped_prev(stdpop = chile_stdpop_f,
                   grouping_vars = c("age_june30", "sex", "adhd"), disease = "adhd")
adhd_prev_adj_f <- get_adjusted_prev(adhd_prev_f, grouping_vars = c()) %>%
  mutate(sex_desc = "Female") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

adhd_prev_m <- chile_bayes_adhd %>%
  filter(sex == 1) %>%
  get_grouped_prev(stdpop = chile_stdpop_m,
                   grouping_vars = c("age_june30", "sex", "adhd"), disease = "adhd")
adhd_prev_adj_m <- get_adjusted_prev(adhd_prev_m, grouping_vars = c()) %>%
  mutate(sex_desc = "Male") %>%
```

```
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

adhd_prev_sex_adj <- rbind(adhd_prev_adj_m, adhd_prev_adj_f)

### Adjusted prevalence by sex, plots

# Autism
aut_prev.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                             grouping_vars = c("age_cat_name",
                                                               "sex_desc",
                                                               "autism"),
                                             disease = "autism") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = aut_prev.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name,
               fill = age_cat_name), position = position_dodge()) +
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100,
                    group = age_cat_name), width = 0.2,
                position = position_dodge(width = 0.9)) +
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5],
                               plasma_colours[7], plasma_colours[9])) +
  labs(title = "Autism prevalence",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age band")

# ADHD
adhd_prev.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                              grouping_vars = c("age_cat_name",
                                                                "sex_desc",
                                                                "adhd"),
                                              disease = "adhd") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = adhd_prev.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name,
               fill = age_cat_name), position = position_dodge()) +
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100,
                    group = age_cat_name), width = 0.2,
                position = position_dodge(width = 0.9)) +
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5],
                               plasma_colours[7], plasma_colours[9])) +
  labs(title = "ADHD prevalence",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age band")

### Adjusted prevalence by health service

# Autism
aut_prev_health <- get_grouped_prev(x = chile_bayes_aut, stdpop = chile_stdpop,
```

```
                                    grouping_vars = c("health_service_code",
                                                      "age_june30",
                                                      "age_cat_name",
                                                      "sex",
                                                      "sex_desc",
                                                      "autism"),
                                    disease = "autism")

aut_prev_health_adj <- get_adjusted_prev(aut_prev_health,
                                         grouping_vars = "health_service_code") %>%
  merge(health_service_lookup, by = "health_service_code") %>%
  rename(health_service_name_long = health_service_name,
         health_service_name = health_service_name_short) %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# ADHD
adhd_prev_health <- get_grouped_prev(x = chile_bayes_adhd, stdpop = chile_stdpop,
                                    grouping_vars = c("health_service_code",
                                                      "age_june30",
                                                      "age_cat_name",
                                                      "sex",
                                                      "sex_desc",
                                                      "adhd"),
                                    disease = "adhd")

adhd_prev_health_adj <- get_adjusted_prev(adhd_prev_health,
                                          grouping_vars = "health_service_code") %>%
  merge(health_service_lookup, by = "health_service_code") %>%
  rename(health_service_name_long = health_service_name,
         health_service_name = health_service_name_short) %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Autism table
aut_prev_health_adj.table <- aut_prev_health_adj %>%
  select(health_service_name, crude_prev_ci, adjusted_prev_ci)
kbl(aut_prev_health_adj.table,
    col.names = c("Health service", "Crude prevalence (95% CI)", "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted autism prevalence by health service in Chile school data

# ADHD table
adhd_prev_health_adj.table <- adhd_prev_health_adj %>%
  select(health_service_name, crude_prev_ci, adjusted_prev_ci)
kbl(adhd_prev_health_adj.table,
    col.names = c("Health service", "Crude prevalence (95% CI)", "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
```

```
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted ADHD prevalence by health service in Chile school data. C

### Adjusted prevalence by SES

# Autism
aut_prev_econA <- get_grouped_prev(x = chile_bayes_aut, stdpop = chile_stdpop,
                                   grouping_vars = c("school_fee",
                                                     "age_june30",
                                                     "age_cat_name",
                                                     "sex",
                                                     "sex_desc",
                                                     "autism"),
                                   disease = "autism")

aut_prev_econA_adj <- get_adjusted_prev(aut_prev_econA, grouping_vars = "school_fee") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# ADHD
adhd_prev_econA <- get_grouped_prev(x = chile_bayes_adhd, stdpop = chile_stdpop,
                                   grouping_vars = c("school_fee",
                                                     "age_june30",
                                                     "age_cat_name",
                                                     "sex",
                                                     "sex_desc",
                                                     "adhd"),
                                   disease = "adhd")

adhd_prev_econA_adj <- get_adjusted_prev(adhd_prev_econA, grouping_vars = "school_fee")  %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Autsim table
aut_prev_econA_adj.table <- aut_prev_econA_adj %>%
  select(school_fee, crude_prev_ci, adjusted_prev_ci)
kbl(aut_prev_econA_adj.table,
    col.names = c("School fee", "Crude prevalence (95% CI)",
                  "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted autism prevalence by monthly school fee (Peso) in Chile s

# ADHD table
adhd_prev_econA_adj.table <- adhd_prev_econA_adj %>%
  select(school_fee, crude_prev_ci, adjusted_prev_ci)
kbl(adhd_prev_econA_adj.table,
    col.names = c("School fee", "Crude prevalence (95% CI)",
                  "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
```

```
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted ADHD prevalence by monthly school fee (Peso) in Chile sc

### Adjusted prevalence by ethnicity

# Autism
aut_prev_ethnic <- get_grouped_prev(x = chile_bayes_aut_ethnic, stdpop = chile_stdpop,
                                    grouping_vars = c("ethnic_2_group",
                                                      "age_june30",
                                                      "age_cat_name",
                                                      "sex",
                                                      "sex_desc",
                                                      "autism"),
                                    disease = "autism")

aut_prev_ethnic_adj <- get_adjusted_prev(aut_prev_ethnic,
                                         grouping_vars = "ethnic_2_group") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# ADHD
adhd_prev_ethnic <- get_grouped_prev(x = chile_bayes_adhd_ethnic, stdpop = chile_stdpop,
                                    grouping_vars = c("ethnic_2_group",
                                                      "age_june30",
                                                      "age_cat_name",
                                                      "sex",
                                                      "sex_desc",
                                                      "adhd"),
                                    disease = "adhd")

adhd_prev_ethnic_adj <- get_adjusted_prev(adhd_prev_ethnic,
                                          grouping_vars = "ethnic_2_group")  %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Autism table
aut_prev_ethnic_adj.table <- aut_prev_ethnic_adj %>%
  select(ethnic_2_group, crude_prev_ci, adjusted_prev_ci)
kbl(aut_prev_ethnic_adj.table,
    col.names = c("Ethnicity", "Crude prevalence (95% CI)",
                  "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted autism prevalence by ethnicity in Chile school data. Cru

# ADHD table
adhd_prev_ethnic_adj.table <- adhd_prev_ethnic_adj %>%
  select(ethnic_2_group, crude_prev_ci, adjusted_prev_ci)
kbl(adhd_prev_ethnic_adj.table,
    col.names = c("Ethnicity", "Crude prevalence (95% CI)",
```

```
                  "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted ADHD prevalence by ethnicity in Chile school data. Crude

### Adjusted prevalence by rurality

# Autism
aut_prev_rural <- get_grouped_prev(x = chile_bayes_aut, stdpop = chile_stdpop,
                                   grouping_vars = c("school_rurality",
                                                     "age_june30",
                                                     "age_cat_name",
                                                     "sex",
                                                     "sex_desc",
                                                     "autism"),
                                   disease = "autism")

aut_prev_rural_adj <- get_adjusted_prev(aut_prev_rural,
                                        grouping_vars = "school_rurality") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# ADHD
adhd_prev_rural <- get_grouped_prev(x = chile_bayes_adhd, stdpop = chile_stdpop,
                                   grouping_vars = c("school_rurality",
                                                     "age_june30",
                                                     "age_cat_name",
                                                     "sex",
                                                     "sex_desc",
                                                     "adhd"),
                                   disease = "adhd")

adhd_prev_rural_adj <- get_adjusted_prev(adhd_prev_rural,
                                         grouping_vars = "school_rurality")  %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Autism table
aut_prev_rural_adj.table <- aut_prev_rural_adj %>%
  select(school_rurality, crude_prev_ci, adjusted_prev_ci)
kbl(aut_prev_rural_adj.table,
    col.names = c("School rurality", "Crude prevalence (95% CI)",
                  "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # NO extra white space
    caption = "Crude and age- and sex-adjusted autism prevalence by school's rurality in Chile school da

# ADHD table
adhd_prev_rural_adj.table <- adhd_prev_rural_adj %>%
```

```
  select(school_rurality, crude_prev_ci, adjusted_prev_ci)
kbl(adhd_prev_rural_adj.table,
    col.names = c("School rurality", "Crude prevalence (95% CI)",
                  "Adjusted prevalence (95% CI)"),
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Crude and age- and sex-adjusted ADHD prevalence by school's rurality in Chile school data
```

```
### Clinical data for MCA

patients_mca_small <- clinical_small_raw %>%
  rename("rurality" = "procedence",
         "dob" = "fecha_nacimiento",
         "apt_date" = "fecha_ejecutada",
         "type_appointment" = "appoinment") %>%
  mutate(gender = str_to_title(gender),
         autism = ifelse(cod_dg_1 %in% aut_codes |
                           cod_dg_2 %in% aut_codes |
                           cod_dg_3 %in% aut_codes, "Yes autism", "No autism"),
         aut_rank = ifelse(cod_dg_1 %in% aut_codes, 1,
                    ifelse(cod_dg_2 %in% aut_codes, 2,
                    ifelse(cod_dg_3 %in% aut_codes, 3, NA))),
         intdisab = ifelse(cod_dg_1 %in% intdisab_codes |
                           cod_dg_2 %in% intdisab_codes |
                           cod_dg_3 %in% intdisab_codes, "Yes intellectual disability",
                           "No intellectual disability"),
         age_june30 = trunc(time_length(interval(ymd(dob), ymd("2021-06-30")), unit = "year")),
         commune_name_upper = ifelse(comuna == "CHOL CHOL", "CHOLCHOL",
                       ifelse(comuna == "CURACAUTIN", "CURACAUTÍN",
                       ifelse(comuna == "PITRUFQUEN", "PITRUFQUÉN",
                       ifelse(comuna == "PUCON", "PUCÓN",
                       ifelse(comuna == "TOLTEN", "TOLTÉN",
                       ifelse(comuna == "VILCUN", "VILCÚN",
                       ifelse(comuna == "DIEGO DE ALMAGRO  (#)", "DIEGO DE ALMAGRO",
                       ifelse(comuna == "MACHALI", "MACHALÍ",
                       ifelse(comuna == "TEMUCO (##)", "TEMUCO", comuna)))))))),
         ses_status = ifelse(socio_economic_level %in% c("COLMENA GOLDEN CROSS",
                                                         "RIO BLANCO",
                                                         "CARABINEROS (DIPRECA)",
                                                         "BANMEDICA S.A.",
                                                         "PARTICULAR (SIN PREVISION)",
                                                         "VIDA TRES"),
                             "Private Health Insurance", socio_economic_level),
         rurality = ifelse(rurality == "URBAN", "Urban", "Rural"),
         ethnicity = ifelse(ethnicity == "CHILENO", "Chilean",
                     ifelse(ethnicity == "MAPUCHE", "Mapuche",
                     ifelse(ethnicity == "EXTRANJERO", "Foreign", "No ethnicity information"))),
         ethnicity = ifelse(is.na(ethnicity), "No ethnicity information", ethnicity), # for some reason
         disability = ifelse(is.na(disability), "No disability information", disability),
         foster_care = ifelse(is.na(foster_care), "No foster care information", foster_care),
         medical_specialty_english = ifelse(medical_specialty == "Physiatry", "Psychiatry",
                                     ifelse(medical_specialty == "PEDIATRIA", "Paediatrics",
```

```
                                    ifelse(medical_specialty == "PSIQUIATRIA", "Psychiatry",
                                    ifelse(medical_specialty == "NEUROLOGIA", "Neurology",
                                           medical_specialty)))),
       medical_specialty_grouped = ifelse(medical_specialty_english %in% c("Psychiatry",
                                                                           "Child Psychiatry"),
                                           "Psychiatry",
                                    ifelse(medical_specialty_english %in% c("Neurology",
                                                                           "Pediatric Neurology"),
                                           "Neurology",
                                    ifelse(medical_specialty_english == "Paediatrics",
                                           "Paediatrics", "Other specialty"))),
       ) %>%
left_join(region_service_commune_lookup, by = "commune_name_upper") %>%
filter(commune_name %in% araucsur_communes$commune_name,
       autism == "Yes autism") %>%
group_by(id, gender, age_june30) %>%
summarise(ses_status = get.mode.na(ses_status),
          autism = get.max.na(autism),
          intdisab = get.max.na(intdisab),
          medical_specialty_grouped = get.mode.na(medical_specialty_grouped),
          hospital = hospital[which.max(apt_date)],
          commune_name = commune_name[which.max(apt_date)],
          rurality = rurality[which.max(apt_date)],
          ethnicity = get.mapuche.ethnicity(ethnicity),
          disability = get.yes.disability(disability),
          foster_care = get.yes.fostercare(foster_care)
          ) %>%
ungroup() %>%
rename("sex_desc" = "gender",
       "age" = "age_june30") %>%
mutate(sex_desc = as.factor(sex_desc),
       age_group = factor(ifelse(age <= 2, "Age 0-2",
                           ifelse(age >= 3 & age <= 5, "Age 3-5",
                           ifelse(age >=6 & age <= 8, "Age 6-8",
                           ifelse(age >=9 & age <=11, "Age 9-11",
                           ifelse(age >=12 & age <= 14, "Age 12-14",
                           ifelse(age >= 15 & age <= 18, "Age 15-18", "Adult")))))),
                           levels = c("Age 0-2", "Age 3-5", "Age 6-8", "Age 9-11",
                                      "Age 12-14", "Age 15-18", "Adult")),
                           # Shouldn't be any adults in this dataset
       age = as.factor(age),
       commune_name = as.factor(commune_name),
       ses_status = as.factor(ses_status),
       rurality = as.factor(rurality),
       ethnicity = factor(ethnicity, levels = c("Mapuche", "Chilean",
                                                "Foreign",
                                                "No ethnicity information")),
       disability_imp = as.factor(ifelse(disability == "No disability information",
                                         "No disability", disability)),
       disability = factor(disability, levels = c("Yes disability",
                                                  "No disability",
                                                  "No disability information")),
       foster_care_imp = as.factor(ifelse(foster_care == "No foster care information",
```

```
                                          "No foster care", foster_care)),
         foster_care = factor(foster_care, levels = c("Yes foster care",
                                                      "No foster care",
                                                      "No foster care information")),
         autism = factor(autism, levels = c("Yes autism", "No autism")),
         intdisab = factor(intdisab, levels = c("Yes intellectual disability",
                                                "No intellectual disability")),
         medical_specialty_grouped = as.factor(medical_specialty_grouped),
         hospital = as.factor(hospital))

### MCA patients table

patients_mca_small.table <- patients_mca_small %>%
  group_by(sex_desc) %>%
  summarise(variable = "Sex",
            count = n(),
            perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
  rename(values = sex_desc) %>%
  rbind(patients_mca_small %>%
          group_by(age_group) %>%
          summarise(variable = "Age band",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
          rename(values = age_group),
        patients_mca_small %>%
          group_by(commune_name) %>%
          summarise(variable = "Commune",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
          rename(values = commune_name),
        patients_mca_small %>%
          group_by(ses_status) %>%
          summarise(variable = "Private health level",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
          rename(values = ses_status),
        patients_mca_small %>%
          group_by(ethnicity) %>%
          summarise(variable = "Ethnicity",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
          rename(values = ethnicity),
        patients_mca_small %>%
          group_by(rurality) %>%
          summarise(variable = "Rurality",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
          rename(values = rurality),
        patients_mca_small %>%
          group_by(disability) %>%
          summarise(variable = "Disability",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
```

```
          rename(values = disability),
        patients_mca_small %>%
          group_by(foster_care) %>%
          summarise(variable = "Foster care",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_mca_small)*100, 2))) %>%
          rename(values = foster_care)) %>%
  mutate(count_suppress = ifelse(count < 20, "<20", count),
         count_perc = ifelse(count < 20, count_suppress,
                             paste0(count_suppress, " (", perc, "%)")),
         variables_short = c("Sex", "",
                             "Age band", "", "", "", "", "",
                             "Commune", "", "", "", "", "", "", "", "", "", "", "",
                             "Private health level", "", "", "", "",
                             "Ethnicity", "", "", "",
                             "Rurality", "",
                             "Disability", "", "",
                             "Foster care", "", "")) %>%
  select(variables_short, values, count_perc)

kbl(patients_mca_small.table,
    col.names = c("Feature", "Available values", "Count (%)"),
    align = "llr",
    booktabs = TRUE,
    format = "latex",
    linesep = c("", "\\addlinespace", # Sex
                "", "", "", "", "", "\\addlinespace", # Age band
                "", "", "", "", "",
                "", "", "", "", "", "\\addlinespace", # Commune
                "", "", "", "", "\\addlinespace", # SES
                "", "", "", "\\addlinespace", # Ethnicity
                "", "\\addlinespace", # Rurality
                "", "", "\\addlinespace", # Disability
                "", "", "\\addlinespace"), # Foster care
    caption = "Count and percentage of features' values in the validated clinical dataset.") %>%
  kable_styling(latex_options = c("repeat_header"))

### MCA no imputation

patients_mca_noimp <- patients_mca_small %>%
   select(sex_desc,
          age_group,
          commune_name,
          ses_status,
          ethnicity,
          rurality,
          disability,
          foster_care) %>%
  rename("Sex" = sex_desc,
         "Age band" = age_group,
         "Commune" = commune_name,
         "Health insurance" = ses_status,
         "Rurality" = rurality,
```

```
         "Ethnicity" = ethnicity,
         "Disability" = disability,
         "Foster care" = foster_care)

res_mca_patients_noimp <- FactoMineR::MCA(patients_mca_noimp,
                                          ncp = 8, # Needs to match number of features
                                          graph = FALSE)

# Plot features
fviz_mca_var(res_mca_patients_noimp,
             choice = "mca.cor",
             labelsize = 4,
             pointsize = 3,
             col.var = turbo_colours[16],
             repel = TRUE,
             ggtheme = theme_minimal()) +
  labs(title = "Categorical features by first two dimensions, no imputation")

# Total contribution to dimension 1 and 2
fviz_contrib(res_mca_patients_noimp, choice = "var", axes = 1:2, top = 15) +
  labs(title = "Contribution of categories to first two dimensions, no imputation",
       x = "Category",
       y = "Contribution (%)")

# Plot categories
fviz_mca_var(res_mca_patients_noimp,
             col.var = "contrib",
             labelsize = 4,
             pointsize = 3,
             gradient.cols = c(turbo_colours[2], turbo_colours[5],
                               turbo_colours[8], turbo_colours[10]),
             repel = TRUE,
             ggtheme = theme_minimal()) +
  labs(title = "Categories by first two dimensions, no imputation",
       colour = "Contribution")

mca_noimp_sex_plot <- fviz_mca_ind(res_mca_patients_noimp,
             label = "none", # hide individual labels
             habillage = "Sex", # color by groups
             palette = c(turbo_colours[14], turbo_colours[17]),
             addEllipses = TRUE, ellipse.type = "confidence",
             title = "Patients by sex, no imputation",
             ggtheme = theme_minimal())
mca_noimp_age_plot <- fviz_mca_ind(res_mca_patients_noimp,
             label = "none",
             habillage = "Age band",
             palette = c(turbo_colours[2], turbo_colours[5], turbo_colours[7],
                         turbo_colours[11], turbo_colours[14], turbo_colours[17]),
             addEllipses = TRUE, ellipse.type = "confidence",
             title = "Patients by age band, no imputation",
             ggtheme = theme_minimal())
mca_noimp_commune_plot <- fviz_mca_ind(res_mca_patients_noimp,
             label = "none",
```

```
            habillage = "Commune",
            palette = c(turbo_colours[2], turbo_colours[3], turbo_colours[5],
                        turbo_colours[7], turbo_colours[9], turbo_colours[11],
                        turbo_colours[12], turbo_colours[14], turbo_colours[15],
                        turbo_colours[17], turbo_colours[18]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by commune of residence, no imputation",
            ggtheme = theme_minimal())
mca_noimp_ses_plot <- fviz_mca_ind(res_mca_patients_noimp,
            label = "none",
            habillage = "Health insurance",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11],
                        turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by proxy SES, no imputation",
            ggtheme = theme_minimal())
mca_noimp_rural_plot <- fviz_mca_ind(res_mca_patients_noimp,
            label = "none",
            habillage = "Rurality",
            palette = c(turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by rurality, no imputation",
            ggtheme = theme_minimal())
mca_noimp_ethnic_plot <- fviz_mca_ind(res_mca_patients_noimp,
            label = "none",
            habillage = "Ethnicity",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11],
                        turbo_colours[14]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by ethnicity, no imputation",
            ggtheme = theme_minimal())
mca_noimp_disab_plot <- fviz_mca_ind(res_mca_patients_noimp,
            label = "none",
            habillage = "Disability",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by disability status, no imputation",
            ggtheme = theme_minimal())
mca_noimp_foster_plot <- fviz_mca_ind(res_mca_patients_noimp,
            label = "none",
            habillage = "Foster care",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by foster care status, no imputation",
            ggtheme = theme_minimal())
mca_noimp_age_plot
mca_noimp_ethnic_plot
mca_noimp_commune_plot
mca_noimp_ses_plot
mca_noimp_sex_plot
mca_noimp_rural_plot
mca_noimp_disab_plot
mca_noimp_foster_plot
```

```r
### MCA with imputation

patients_mca_imp <- patients_mca_small %>%
   select(sex_desc,
          age_group,
          commune_name,
          ses_status,
          ethnicity,
          rurality,
          disability_imp,
          foster_care_imp) %>%
  rename("Sex" = sex_desc,
         "Age band" = age_group,
         "Commune" = commune_name,
         "Health insurance" = ses_status,
         "Rurality" = rurality,
         "Ethnicity" = ethnicity,
         "Disability imputed" = disability_imp,
         "Foster care imputed" = foster_care_imp)

res_mca_patients_imp <- FactoMineR::MCA(patients_mca_imp,
                                        ncp = 8, # Needs to match number of features
                                        graph = FALSE)

# Plot features
fviz_mca_var(res_mca_patients_imp,
             choice = "mca.cor",
             labelsize = 4,
             pointsize = 3,
             col.var = turbo_colours[16],
             repel = TRUE,
             ggtheme = theme_minimal()) +
  labs(title = "Categorical features by first two dimensions, with imputation")

# Total contribution to dimension 1 and 2
fviz_contrib(res_mca_patients_imp, choice = "var", axes = 1:2, top = 15) +
  labs(title = "Contribution of categories to first two dimensions, with imputation",
       x = "Category",
       y = "Contribution (%)")

# Plot categories
fviz_mca_var(res_mca_patients_imp,
             col.var = "contrib",
             labelsize = 4,
             pointsize = 3,
             gradient.cols = c(turbo_colours[2], turbo_colours[5],
                               turbo_colours[8], turbo_colours[10]),
             repel = TRUE,
             ggtheme = theme_minimal()) +
  labs(title = "Categories by first two dimensions, with imputation",
       colour = "Contribution")

mca_imp_sex_plot <- fviz_mca_ind(res_mca_patients_imp,
```

```
            label = "none", # hide individual labels
            habillage = "Sex", # color by groups
            palette = c(turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by sex, with imputation",
            ggtheme = theme_minimal())
mca_imp_age_plot <- fviz_mca_ind(res_mca_patients_imp,
            label = "none",
            habillage = "Age band",
            palette = c(turbo_colours[2], turbo_colours[5], turbo_colours[7],
                        turbo_colours[11], turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by age band, with imputation",
            ggtheme = theme_minimal())
mca_imp_commune_plot <- fviz_mca_ind(res_mca_patients_imp,
            label = "none",
            habillage = "Commune",
            palette = c(turbo_colours[2], turbo_colours[3], turbo_colours[5],
                        turbo_colours[7], turbo_colours[9], turbo_colours[11],
                        turbo_colours[12], turbo_colours[14], turbo_colours[15],
                        turbo_colours[17], turbo_colours[18]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by commune of residence, with imputation",
            ggtheme = theme_minimal())
mca_imp_ses_plot <- fviz_mca_ind(res_mca_patients_imp,
            label = "none",
            habillage = "Health insurance",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11],
                        turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by proxy SES, with imputation",
            ggtheme = theme_minimal())
mca_imp_rural_plot <- fviz_mca_ind(res_mca_patients_imp,
            label = "none",
            habillage = "Rurality",
            palette = c(turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by rurality, with imputation",
            ggtheme = theme_minimal())
mca_imp_ethnic_plot <- fviz_mca_ind(res_mca_patients_imp,
            label = "none",
            habillage = "Ethnicity",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11],
                        turbo_colours[14]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by ethnicity, with imputation",
            ggtheme = theme_minimal())
mca_imp_disab_plot <- fviz_mca_ind(res_mca_patients_imp,
            label = "none",
            habillage = "Disability imputed",
            palette = c(turbo_colours[3], turbo_colours[7]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by disability status, with imputation",
```

```
              ggtheme = theme_minimal())
mca_imp_foster_plot <- fviz_mca_ind(res_mca_patients_imp,
              label = "none",
              habillage = "Foster care imputed",
              palette = c(turbo_colours[3], turbo_colours[7]),
              addEllipses = TRUE, ellipse.type = "confidence",
              title = "Patients by foster care status, with imputation",
              ggtheme = theme_minimal())
mca_imp_age_plot
mca_imp_ethnic_plot
mca_imp_commune_plot
mca_imp_ses_plot
mca_imp_sex_plot
mca_imp_rural_plot
mca_imp_disab_plot
mca_imp_foster_plot

### MCA 3 features

patients_mca_aut <- patients_mca_small %>%
  filter(autism == "Yes autism") %>%
   select(age_group,
          commune_name,
          ethnicity) %>%
  rename("Age band" = age_group,
         "Commune" = commune_name,
         "Ethnicity" = ethnicity)

res_mca_patients_aut <- FactoMineR::MCA(patients_mca_aut,
                                        ncp = 3, # Needs to match number of features
                                        graph = FALSE)

# Plot features
fviz_mca_var(res_mca_patients_aut,
              choice = "mca.cor",
              labelsize = 4,
              pointsize = 3,
              col.var = turbo_colours[16],
              repel = TRUE,
              ggtheme = theme_minimal()) +
  labs(title = "Categorical features by first two dimensions, 3 features")

# Total contribution to dimension 1 and 2
fviz_contrib(res_mca_patients_aut, choice = "var", axes = 1:2, top = 15) +
  labs(title = "Contribution of categories to first two dimensions, 3 features",
       x = "Category",
       y = "Contribution (%)")

# Plot categories
fviz_mca_var(res_mca_patients_aut,
              col.var = "contrib",
              labelsize = 4,
              pointsize = 3,
```

```
            gradient.cols = c(turbo_colours[2], turbo_colours[5],
                              turbo_colours[8], turbo_colours[10]),
            repel = TRUE,
            ggtheme = theme_minimal()) +
  labs(title = "Categories by first two dimensions, 3 features",
       colour = "Contribution")

mca_aut_age_plot <- fviz_mca_ind(res_mca_patients_aut,
            label = "none",
            habillage = "Age band",
            palette = c(turbo_colours[2], turbo_colours[5], turbo_colours[7],
                        turbo_colours[11], turbo_colours[14], turbo_colours[17]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by age band, 3 features",
            ggtheme = theme_minimal())
mca_aut_commune_plot <- fviz_mca_ind(res_mca_patients_aut,
            label = "none",
            habillage = "Commune",
            palette = c(turbo_colours[2], turbo_colours[3], turbo_colours[5],
                        turbo_colours[7], turbo_colours[9], turbo_colours[11],
                        turbo_colours[12], turbo_colours[14], turbo_colours[15],
                        turbo_colours[17], turbo_colours[18]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by commune of residence, 3 features",
            ggtheme = theme_minimal())
mca_aut_ethnic_plot <- fviz_mca_ind(res_mca_patients_aut,
            label = "none",
            habillage = "Ethnicity",
            palette = c(turbo_colours[3], turbo_colours[7], turbo_colours[11],
                        turbo_colours[14]),
            addEllipses = TRUE, ellipse.type = "confidence",
            title = "Patients by ethnicity, 3 features",
            ggtheme = theme_minimal())
mca_aut_age_plot
mca_aut_ethnic_plot
mca_aut_commune_plot
```

```
### School data for linkage

school_ssas <- chile_bayes %>%
  select(-student_commune_name) %>%
  filter(commune_code %in% araucsur_communes$commune_code) %>%
  mutate(dob = ymd(dob),
         ses_status = ifelse(school_fee == "Free", 1,
                      ifelse(school_fee == "$1,000-$10,000" |
                               school_fee == "$10,001-$50,000" |
                               school_fee == "$50,001-$100,000", 2,
                      ifelse(school_fee == "$100,001+", 3, NA)))
         ) %>%
  rename(student_commune_name = commune_name) %>%
  select(dob,
         sex_desc,
         student_commune_name,
         autism,
```

```
         ses_status,
         age_june30,
         age_cat_name)

school_ssas_aut <- school_ssas %>%
  filter(autism == 1) %>%
  # We only want to find additional autism cases in the clinical records
  select(-age_june30, -autism) %>%
  rowid_to_column("id")

# Add a fake row so that pairing works later
school <- rbind(select(school_ssas_aut, -age_cat_name),
                c(nrow(school_ssas_aut)+1, "2023-06-01", NA, NA, NA))

### School data chi squared tests

# Check if students in SSAS are different to other students
chile_bayes_ssas_grouped <- chile_bayes %>%
  filter(ssas == 1) %>%
  group_by(sex_desc) %>%
  summarise(variable = "Sex",
            count = n(),
            ssas_perc = paste0(sprintf("%.2f",
                                       round(count / chile_bayes %>%
                                               filter(ssas == 1) %>%
                                               nrow() * 100, 2)), "%")) %>%
  rename(values = sex_desc) %>%
  rbind(chile_bayes %>%
          filter(ssas == 1) %>%
          group_by(age_cat_name) %>%
          summarise(variable = "Age band",
                    count = n(),
                    ssas_perc = paste0(sprintf("%.2f",
                                               round(count / chile_bayes %>%
                                                       filter(ssas == 1) %>%
                                                       nrow() * 100, 2)), "%")) %>%
          rename(values = age_cat_name),
        chile_bayes %>%
          filter(ssas == 1) %>%
          group_by(school_fee) %>%
          summarise(variable = "School fee",
                    count = n(),
                    ssas_perc = paste0(sprintf("%.2f",
                                               round(count / chile_bayes %>%
                                                       filter(ssas == 1) %>%
                                                       nrow() * 100, 2)), "%")) %>%
          rename(values = school_fee),
        chile_bayes %>%
          filter(ssas == 1) %>%
          group_by(ethnic_2_group) %>%
          summarise(variable = "Ethnicity",
                    count = n(),
                    ssas_perc = paste0(sprintf("%.2f",
```

```
                                              round(count / chile_bayes %>%
                                                      filter(ssas == 1) %>%
                                                      nrow() * 100, 2)), "%")) %>%
          rename(values = ethnic_2_group),
        chile_bayes %>%
          filter(ssas == 1) %>%
          group_by(school_rurality) %>%
          summarise(variable = "Rurality",
                    count = n(),
                    ssas_perc = paste0(sprintf("%.2f",
                                              round(count / chile_bayes %>%
                                                      filter(ssas == 1) %>%
                                                      nrow() * 100, 2)), "%")) %>%
          rename(values = school_rurality),
        chile_bayes %>%
          filter(ssas == 1) %>%
          group_by(special_needs_status) %>%
          summarise(variable = "Accesses SEED",
                    count = n(),
                    ssas_perc = paste0(sprintf("%.2f",
                                              round(count / chile_bayes %>%
                                                      filter(ssas == 1) %>%
                                                      nrow() * 100, 2)), "%")) %>%
          rename(values = special_needs_status) %>%
          mutate(values = ifelse(values == 0, "No", "Yes"))) %>%
  select(variable, values, ssas_perc)

chile_bayes_nonssas_grouped <- chile_bayes %>%
  filter(ssas == 0) %>%
  group_by(sex_desc) %>%
  summarise(variable = "Sex",
            count = n(),
            nonssas_perc = paste0(sprintf("%.2f",
                                          round(count / chile_bayes %>%
                                                  filter(ssas == 0) %>%
                                                  nrow() * 100, 2)), "%")) %>%
  rename(values = sex_desc) %>%
  rbind(chile_bayes %>%
          filter(ssas == 0) %>%
          group_by(age_cat_name) %>%
          summarise(variable = "Age band",
                    count = n(),
                    nonssas_perc = paste0(sprintf("%.2f",
                                                  round(count / chile_bayes %>%
                                                          filter(ssas == 0) %>%
                                                          nrow() * 100, 2)), "%")) %>%
          rename(values = age_cat_name),
        chile_bayes %>%
          filter(ssas == 0) %>%
          group_by(school_fee) %>%
          summarise(variable = "School fee",
                    count = n(),
                    nonssas_perc = paste0(sprintf("%.2f",
```

```
                                                round(count / chile_bayes %>%
                                                        filter(ssas == 0) %>%
                                                        nrow() * 100, 2)), "%")) %>%
          rename(values = school_fee),
        chile_bayes %>%
          filter(ssas == 0) %>%
          group_by(ethnic_2_group) %>%
          summarise(variable = "Ethnicity",
                    count = n(),
                    nonssas_perc = paste0(sprintf("%.2f",
                                                round(count / chile_bayes %>%
                                                        filter(ssas == 0) %>%
                                                        nrow() * 100, 2)), "%")) %>%
          rename(values = ethnic_2_group),
        chile_bayes %>%
          filter(ssas == 0) %>%
          group_by(school_rurality) %>%
          summarise(variable = "Rurality",
                    count = n(),
                    nonssas_perc = paste0(sprintf("%.2f",
                                                round(count / chile_bayes %>%
                                                        filter(ssas == 0) %>%
                                                        nrow() * 100, 2)), "%")) %>%
          rename(values = school_rurality),
        chile_bayes %>%
          filter(ssas == 0) %>%
          group_by(special_needs_status) %>%
          summarise(variable = "Accesses SEED",
                    count = n(),
                    nonssas_perc = paste0(sprintf("%.2f",
                                                round(count / chile_bayes %>%
                                                        filter(ssas == 0) %>%
                                                        nrow() * 100, 2)), "%")) %>%
          rename(values = special_needs_status) %>%
          mutate(values = ifelse(values == 0, "No", "Yes"))) %>%
  select(variable, values, nonssas_perc)

chisq_pvals <- c(ifelse(chisq.test(chile_bayes$sex_desc, chile_bayes$ssas)$p.value < 0.0001,
                        "<<0.001", signif(chisq.test(chile_bayes$sex_desc,
                                                     chile_bayes$ssas)$p.value, 3)), "",
                 ifelse(chisq.test(chile_bayes$age_cat_name,
                                   chile_bayes$ssas)$p.value < 0.0001,
                        "<<0.001", signif(chisq.test(chile_bayes$age_cat_name,
                                                     chile_bayes$ssas)$p.value, 3)), "", "", "",
                 ifelse(chisq.test(chile_bayes$school_fee,
                                   chile_bayes$ssas)$p.value < 0.0001,
                        "<<0.001", signif(chisq.test(chile_bayes$school_fee,
                                                     chile_bayes$ssas)$p.value, 3)), "", "", "", "", "",
                 ifelse(chisq.test(chile_bayes$ethnic_2_group,
                                   chile_bayes$ssas)$p.value < 0.0001,
                        "<<0.001", signif(chisq.test(chile_bayes$ethnic_2_group,
                                                     chile_bayes$ssas)$p.value, 3)), "", "",
                 ifelse(chisq.test(chile_bayes$school_rurality,
```

```
                                    chile_bayes$ssas)$p.value < 0.0001,
                          "<<0.001", signif(chisq.test(chile_bayes$school_rurality,
                                                       chile_bayes$ssas)$p.value, 3)), "",
                  ifelse(chisq.test(chile_bayes$special_needs_status,
                                    chile_bayes$ssas)$p.value < 0.0001,
                          "<<0.001", signif(chisq.test(chile_bayes$special_needs_status,
                                                       chile_bayes$ssas)$p.value, 3)), "")

chile_bayes_grouped <- cbind(chile_bayes_ssas_grouped,
                             chile_bayes_nonssas_grouped[, 3],
                             chisq_pvals) %>%
  mutate(variables_short = c("Sex", "",
                             "Age band", "", "", "",
                             "School fee", "", "", "", "", "", "",
                             "Ethnicity", "", "",
                             "Rurality", "",
                             "Accesses SEED", ""
                             )) %>%
  select(variables_short, values, ssas_perc, nonssas_perc, chisq_pvals)

kbl(chile_bayes_grouped,
    col.names = c("Feature", "Available values", "% for SSAS", "% for non-SSAS", "p-value"),
    align = "llrrr",
    booktabs = TRUE,
    format = "latex",
    linesep = c("", "\\addlinespace", # Sex
                "", "", "", "\\addlinespace", # Age
                "", "", "", "", "", "", "", "\\addlinespace", # SES
                "", "", "\\addlinespace", # Ethnicity
                "", "\\addlinespace", # Rurality,
                "", "\\addlinespace"), # SEED
    caption = "Percentage of features' values for each category for the students in the school data resi
  kable_styling(latex_options = c("repeat_header"))

# SSAS school data table

school_sass_aut.table <- school_ssas_aut %>%
  group_by(sex_desc) %>%
  summarise(variable = "Sex",
            count = n(),
            perc = sprintf("%.2f", round(count/nrow(school_ssas_aut)*100, 2))) %>%
  rename(values = sex_desc) %>%
  rbind(school_ssas_aut %>%
          group_by(age_cat_name) %>%
          summarise(variable = "Age band",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(school_ssas_aut)*100, 2))) %>%
          rename(values = age_cat_name),
        school_ssas_aut %>%
          group_by(student_commune_name) %>%
          summarise(variable = "Commune",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(school_ssas_aut)*100, 2))) %>%
```

```
          rename(values = student_commune_name),
        school_ssas_aut %>%
          group_by(ses_status) %>%
          summarise(variable = "Proxy SES",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(school_ssas_aut)*100, 2))) %>%
          rename(values = ses_status)) %>%
  mutate(count_suppress = ifelse(count < 20, "<20", count),
         count_perc = ifelse(count < 20, count_suppress, paste0(count_suppress,
                                                                " (", perc, "%)")),
         variables_short = c("Sex", "",
                             "Age band", "", "", "",
                             "Commune", "", "", "", "", "", "", "", "", "", "",
                             "", "", "", "", "", "", "", "", "", "",
                             "Proxy SES", "", "", "")) %>%
  select(variables_short, values, count_perc)

kbl(school_sass_aut.table,
    col.names = c("Feature", "Available values", "Count (%)"),
    align = "llr",
    booktabs = TRUE,
    format = "latex",
    linesep = c("", "\\addlinespace", # Sex
                "", "", "", "\\addlinespace", # Age
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "", "\\addlinespace", # Commune
                "", "", "", "\\addlinespace"), # SES
    caption = "Count and percentage of features' values in the SSAS school dataset which comprises stud
  kable_styling(latex_options = c("repeat_header"))

### Patient data for linkage

patients_age <- clinical %>%
  filter(commune_code %in% araucsur_communes$commune_code,
         age_june30 >= 6 & age_june30 <= 18) %>%
  group_by(id, gender, dob, commune_name, region_name) %>%
  summarise(ses_status = get.mode.na(ses_status),
            autism = get.max.na(autism),
            aut_rank = get.min.na(aut_rank),
            age_june30 = get.min.na(age_june30)) %>%
  ungroup() %>%
  rename("student_commune_name" = "commune_name",
         "student_region_name" = "region_name",
         "sex_desc" = "gender") %>%
  rowid_to_column("row_id") %>%
  select(row_id,
         id,
         dob,
         age_june30,
         sex_desc,
         student_commune_name,
```

```
        ses_status)

patients <- patients_age %>%
  select(-age_june30) %>%
  mutate(ses_status = as.character(ses_status))

patients_distinct <- patients %>%
  group_by(id) %>%
  summarise(commune_count = n(), communes = list(student_commune_name))

### Patient data table

ssas_patients.table <- patients_age %>%
  group_by(sex_desc) %>%
  summarise(variable = "Sex",
            count = n(),
            perc = sprintf("%.2f", round(count/nrow(patients_age)*100, 2))) %>%
  rename(values = sex_desc) %>%
  rbind(patients_age %>%
          mutate(age_cat_name = factor(ifelse(age_june30 <= 8, "6-8",
                                       ifelse(age_june30 <= 11, "9-11",
                                       ifelse(age_june30 <= 14, "12-14", "15-18"))),
                                       levels = c("6-8", "9-11", "12-14", "15-18"))) %>%
          group_by(age_cat_name) %>%
          summarise(variable = "Age band",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_age)*100, 2))) %>%
          rename(values = age_cat_name),
        patients_age %>%
          group_by(student_commune_name) %>%
          summarise(variable = "Commune",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_age)*100, 2))) %>%
          rename(values = student_commune_name),
        patients_age %>%
          group_by(ses_status) %>%
          summarise(variable = "Proxy SES",
                    count = n(),
                    perc = sprintf("%.2f", round(count/nrow(patients_age)*100, 2))) %>%
          rename(values = ses_status)) %>%
  mutate(count_suppress = ifelse(count < 20, "<20", count),
         count_perc = ifelse(count < 20, count_suppress, paste0(count_suppress,
                                                                " (", perc, "%)")),
         variables_short = c("Sex", "",
                             "Age band", "", "", "",
                             "Commune", "", "", "", "", "", "", "", "", "", "",
                             "", "", "", "", "", "", "", "", "", "",
                             "Proxy SES", "", "")) %>%
  select(variables_short, values, count_perc)

kbl(ssas_patients.table,
    col.names = c("Feature", "Available values", "Count (%)"),
    align = "llr",
```

```
    booktabs = TRUE,
    format = "latex",
    linesep = c("", "\\addlinespace", # Sex
                "", "", "", "\\addlinespace", # Age
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "",
                "", "", "", "", "", "\\addlinespace", # Commune
                "", "", "\\addlinespace"), # SES
    caption = "Count and percentage of features' values in the patient dataset. Proxy SES is health insu
  kable_styling(latex_options = c("repeat_header"))

### Manual linkage

patients_grouped_ses <- patients %>%
  group_by(sex_desc,
           dob,
           student_commune_name,
           ses_status) %>%
  summarise(count = n(),
            ids = list(id)) %>%
  ungroup()

school_grouped_ses <- school %>%
  group_by(sex_desc,
           dob,
           student_commune_name,
           ses_status) %>%
  summarise(count = n(),
            ids = list(rowid)) %>%
  ungroup()

merged_ses <- merge(school, patients, by = c("sex_desc",
                                             "dob",
                                             "student_commune_name",
                                             "ses_status"), all = FALSE)

patients_grouped <- patients %>%
  group_by(sex_desc,
           dob,
           student_commune_name) %>%
  summarise(count = n(),
            ids = list(id))

school_grouped <- school %>%
  group_by(sex_desc,
           dob,
           student_commune_name) %>%
  summarise(count = n(),
            ses = list(ses_status))

merged <- merge(school, patients, by = c("sex_desc",
                                         "dob",
```

```
                                        "student_commune_name"), all = FALSE)

###  Probabilistic Weights

pairs_blocked <- compare.linkage(school,
                      select(patients, -row_id),
                      blockfld = c("sex_desc", "dob"),
                       # Block on sex and dob because we really want them to be
                      # the same and helps computation time
                      exclude = c(1) # Exclude the id column in both datasets
                      )

pairs_weighted <- epiWeights(pairs_blocked, e = c(0.01, # Default for DOB
                           0.01, # Default for sex
                           0.01, # Default for commune because we want a good match
                           0.1   # Have more error for SES because it is loosely defined
                           ))

pairs_classified <- emClassify(pairs_weighted, threshold.upper = 1, threshold.lower = 0.8)

pairs_blocked.df <- pairs_blocked$pairs %>%
  mutate(weight = pairs_classified$Wdata,
         pred = pairs_classified$prediction) %>%
  rename(".x" = id1, ".y" = id2) %>%
  select(-is_match)

# Turn into a 'pairs' type object so it can be fed to select_n_to_m()
pairs_blocked.pairs <- pair_blocking(school, patients, on = c("sex_desc", "dob")) %>%
         mutate(student_commune_name = (school$student_commune_name[.x] ==
                                          patients$student_commune_name[.y])) %>%
  left_join(pairs_blocked.df, by = c(".x", ".y")) %>%
  select(c(-student_commune_name.x)) %>%
  rename("student_commune_name" = "student_commune_name.y")

matches <- select_n_to_m(pairs_blocked.pairs, threshold = 0.5, score = "weight",
                         n = 1, m = 1, var = "match") %>%
  filter(match == TRUE) %>%
  rename("id" = ".x",
         "row_id" = ".y") %>%
  mutate(id = as.character(id))

# Match school records

# Now add the matched clinical records to the school records
school_matched <- school %>%
  mutate(id = as.character(id)) %>%
  filter(!is.na(sex_desc)) %>%
  left_join(matches, by = "id") %>%
  rename(id.school = id,
         dob.school = dob.x,
         sex_desc.school = sex_desc.x,
         student_commune_name.school = student_commune_name.x,
         ses_status.school = ses_status.x,
```

```
         dob.matched = dob.y,
         sex_desc.matched = sex_desc.y,
         student_commune_name.matched = student_commune_name.y,
         ses_status.matched = ses_status.y) %>%
  select(c(-pred, -match)) %>%
  left_join(patients, by = "row_id") %>%
  rename(id.patient = row_id,
         patient_id = id,
         dob.patient = dob,
         sex_desc.patient = sex_desc,
         student_commune_name.patient = student_commune_name,
         ses_status.patient = ses_status) %>%
  select(id.school, id.patient, patient_id,
         dob.school, dob.patient, dob.matched,
         sex_desc.school, sex_desc.patient, sex_desc.matched,
         student_commune_name.school, student_commune_name.patient,
         student_commune_name.matched,
         ses_status.school, ses_status.patient, ses_status.matched,
         weight) %>%
  arrange(desc(weight))

school_matched_small <- school_matched %>%
  mutate(matched = ifelse(is.na(patient_id), 0, 1),
         sex.school = ifelse(sex_desc.school == "Male", 1,
                      ifelse(sex_desc.school == "Female", 2, NA))) %>%
  merge(region_service_commune_lookup %>% as.tibble() %>% select(-geometry),
        by.x = "student_commune_name.school", by.y = "commune_name") %>%
  select(id.school,
         dob.school,
         sex_desc.school, sex.school,
         student_commune_name.school, commune_code,
         ses_status.school,
         matched)
# Fake record has now been excluded

### Match patients

# Now add the matched clinical records to the school records
patients_matched <- patients %>%
  left_join(matches, by = "row_id") %>%
  rename(id.patient = row_id,
         patient_id = id.x,
         dob.patient = dob.x,
         sex_desc.patient = sex_desc.x,
         student_commune_name.patient = student_commune_name.x,
         id = id.y,
         ses_status.patient = ses_status.x,
         dob.matched = dob.y,
         sex_desc.matched = sex_desc.y,
         student_commune_name.matched = student_commune_name.y,
         ses_status.matched = ses_status.y) %>%
  select(c(-pred, -match)) %>%
  left_join(school, by = "id") %>%
```

```
  rename(id.school = id,
         dob.school = dob,
         sex_desc.school = sex_desc,
         student_commune_name.school = student_commune_name,
         ses_status.school = ses_status) %>%
  select(id.school, id.patient, patient_id,
         dob.school, dob.patient, dob.matched,
         sex_desc.school, sex_desc.patient, sex_desc.matched,
         student_commune_name.school, student_commune_name.patient,
         student_commune_name.matched,
         ses_status.school, ses_status.patient, ses_status.matched,
         weight) %>%
  arrange(desc(weight))

patients_matched_small <- patients_matched %>%
  mutate(matched = ifelse(is.na(id.school), 0, 1),
         sex.patient = ifelse(sex_desc.patient == "Male", 1,
                       ifelse(sex_desc.patient == "Female", 2, NA))) %>%
  merge(region_service_commune_lookup %>% as.tibble() %>% select(-geometry),
        by.x = "student_commune_name.patient", by.y = "commune_name") %>%
  select(patient_id, id.patient,
         dob.patient,
         sex_desc.patient, sex.patient,
         student_commune_name.patient, commune_code,
         ses_status.patient,
         matched)

### Unique patients

# Check if any of the patients that lived in multiple communes matched to
# multiple school records
patients_matched_unique <- patients_matched_small %>%
  group_by(matched, patient_id) %>%
  summarise(count = n())

patients_dup <- patients_matched_unique %>%
  filter(matched == 1, count > 1) %>%
  select(patient_id)
# No patient is inadvertently matched to two school records

### Komogorov school

school_yes <- school_matched_small %>% filter(matched == 1)
school_no <- school_matched_small %>% filter(matched == 0)

# Kolmogorov tests for our matched results
ks.school.sex <- ks.test(na.omit(school_yes$sex.school),
                         na.omit(school_no$sex.school),
                         alternative = "two.sided", simulate.p.value = TRUE)

ks.school.ses_status <- ks.test(as.numeric(na.omit(school_yes$ses_status.school)),
                                as.numeric(na.omit(school_no$ses_status.school)),
                                alternative = "two.sided", simulate.p.value = TRUE)
```

```r
ks.school.commune_code<- ks.test(as.numeric(na.omit(school_yes$commune_code)),
                                 as.numeric(na.omit(school_no$commune_code)),
                                 alternative = "two.sided", simulate.p.value = TRUE)

# Kolmogorov tests with permutation distributions
set.seed(123)
nPerm <- 2000
ks_perm.school.pvals <- data.frame(sex = numeric(nPerm),
                                   commune_code = numeric(nPerm),
                                   ses_status = numeric(nPerm))

school_matched_small_perm <- school_matched_small

for (i in 1:nPerm) {
  # permute matched status
  school_matched_small_perm$matched <- school_matched_small$matched[sample(nrow(school_matched_small))]
  school_perm_yes <- school_matched_small_perm %>% filter(matched == 1)
  school_perm_no <- school_matched_small_perm %>% filter(matched == 0)

  ks_perm.school.sex <- ks.test(na.omit(school_perm_yes$sex.school),
                                na.omit(school_perm_no$sex.school),
                                alternative = "two.sided")
  ks_perm.school.commune_code <- ks.test(as.numeric(na.omit(school_perm_yes$commune_code)),
                                         as.numeric(na.omit(school_perm_no$commune_code)),
                                         alternative = "two.sided")
  ks_perm.school.ses_status <- ks.test(as.numeric(na.omit(school_perm_yes$ses_status.school)),
                                       as.numeric(na.omit(school_perm_no$ses_status.school)),
                                       alternative = "two.sided")

  ks_perm.school.pvals$sex[i] <- ks_perm.school.sex$p.value
  ks_perm.school.pvals$commune_code[i] <- ks_perm.school.commune_code$p.value
  ks_perm.school.pvals$ses_status[i] <- ks_perm.school.ses_status$p.value
}

# Kolmogorov patients

patients_yes <- patients_matched_small %>% filter(matched == 1)
patients_no <- patients_matched_small %>% filter(matched == 0)

# Kolmogorov tests for our matched results
ks.patients.sex <- ks.test(na.omit(patients_yes$sex.patient),
                           na.omit(patients_no$sex.patient),
                           alternative = "two.sided", simulate.p.value = TRUE)

ks.patients.ses_status <- ks.test(as.numeric(na.omit(patients_yes$ses_status.patient)),
                                  as.numeric(na.omit(patients_no$ses_status.patient)),
                                  alternative = "two.sided", simulate.p.value = TRUE)

ks.patients.commune_code<- ks.test(as.numeric(na.omit(patients_yes$commune_code)),
                                   as.numeric(na.omit(patients_no$commune_code)),
                                   alternative = "two.sided", simulate.p.value = TRUE)

# Kolmogorov tests with permutation distributions
```

```
set.seed(123)
nPerm <- 2000
ks_perm.patients.pvals <- data.frame(sex = numeric(nPerm),
                                     commune_code = numeric(nPerm),
                                     ses_status = numeric(nPerm))

patients_matched_small_perm <- patients_matched_small

for (i in 1:nPerm) {
  # permute matched status
  patients_matched_small_perm$matched <- patients_matched_small$matched[sample(nrow(patients_matched_sma
  patients_perm_yes <- patients_matched_small_perm %>% filter(matched == 1)
  patients_perm_no <- patients_matched_small_perm %>% filter(matched == 0)

  ks_perm.patients.sex <- ks.test(na.omit(patients_perm_yes$sex.patient),
                                  na.omit(patients_perm_no$sex.patient),
                                  alternative = "two.sided")
  ks_perm.patients.commune_code <- ks.test(as.numeric(na.omit(patients_perm_yes$commune_code)),
                                           as.numeric(na.omit(patients_perm_no$commune_code)),
                                           alternative = "two.sided")
  ks_perm.patients.ses_status <- ks.test(as.numeric(na.omit(patients_perm_yes$ses_status.patient)),
                                         as.numeric(na.omit(patients_perm_no$ses_status.patient)),
                                         alternative = "two.sided")

  ks_perm.patients.pvals$sex[i] <- ks_perm.patients.sex$p.value
  ks_perm.patients.pvals$commune_code[i] <- ks_perm.patients.commune_code$p.value
  ks_perm.patients.pvals$ses_status[i] <- ks_perm.patients.ses_status$p.value
}

# Kolmogorov proportions

# Results for school by sex
school_match_yes.sex <- school_yes %>%
  group_by(sex.school) %>%
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 1)
school_match_no.sex <- school_no %>%
  group_by(sex.school) %>%
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 0)
school_match.sex <- rbind(school_match_yes.sex, school_match_no.sex) %>%
  mutate(sex_desc = ifelse(sex.school == 1, "Male",
                           ifelse(sex.school == 2, "Female", NA))) %>%
  arrange(sex_desc, matched)

# Results for patients by sex
patients_match_yes.sex <- patients_yes %>%
  group_by(sex.patient) %>%
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 1)
patients_match_no.sex <- patients_no %>%
  group_by(sex.patient) %>%
  summarise(count = n()) %>%
```

```
  mutate(freq = count/sum(count), matched = 0)
patients_match.sex <- rbind(patients_match_yes.sex, patients_match_no.sex) %>%
  mutate(sex_desc = ifelse(sex.patient == 1, "Male",
                           ifelse(sex.patient == 2, "Female", NA))) %>%
  arrange(sex_desc, matched)

# Results for school by commune
school_match_yes.student_commune_name <- school_yes %>%
  group_by(student_commune_name.school) %>%
  summarise(count = n()) %>% mutate(freq = count/sum(count)) %>%
  mutate(matched = 1)
school_match_no.student_commune_name <- school_no %>%
  group_by(student_commune_name.school) %>%
  summarise(count = n()) %>% mutate(freq = count/sum(count)) %>%
  mutate(matched = 0)
school_match.student_commune_name <- rbind(school_match_yes.student_commune_name,
                                           school_match_no.student_commune_name) %>%
  arrange(student_commune_name.school, matched)

# Results for patients by commune
patients_match_yes.student_commune_name <- patients_yes %>%
  group_by(student_commune_name.patient) %>%
  summarise(count = n()) %>% mutate(freq = count/sum(count)) %>%
  mutate(matched = 1)
patients_match_no.student_commune_name <- patients_no %>%
  group_by(student_commune_name.patient) %>%
  summarise(count = n()) %>% mutate(freq = count/sum(count)) %>%
  mutate(matched = 0)

patients_match.student_commune_name <- rbind(patients_match_yes.student_commune_name,
                                             patients_match_no.student_commune_name) %>%
  arrange(student_commune_name.patient, matched)

# Results for school by SES
school_match_yes.ses_status <- school_yes %>%
  group_by(ses_status.school) %>%
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 1)
school_match_no.ses_status <- school_no %>%
  group_by(ses_status.school) %>%
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 0)
school_match.ses_status <- rbind(school_match_yes.ses_status,
                                 school_match_no.ses_status) %>%
  arrange(ses_status.school, matched)

# Results for patients by SES
patients_match_yes.ses_status <- patients_yes %>%
  group_by(ses_status.patient) %>%
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 1)
patients_match_no.ses_status <- patients_no %>%
  group_by(ses_status.patient) %>%
```

```
  summarise(count = n()) %>%
  mutate(freq = count/sum(count), matched = 0)
patients_match.ses_status <- rbind(patients_match_yes.ses_status,
                                   patients_match_no.ses_status) %>%
  arrange(ses_status.patient, matched)
```

```
# Aggregate patients to single row per patient by choosing the commune
# that has a match where available.
patients_matched_small_dedup <- patients_matched_small %>%
  group_by(patient_id, dob.patient, sex_desc.patient) %>%
  summarise(commune_list = list(student_commune_name.patient),
            student_commune_name.patient = student_commune_name.patient[which.max(matched)],
            ses_list = list(ses_status.patient),
            ses_status.patient = ses_status.patient[which.max(matched)],
            #autism.patient = max(autism.patient),
            matched = max(matched))



n_aut_araucS <- nrow(school_matched) +
  length(unique(patients_matched$patient_id)) - nrow(matches)
# OR
n_aut_araucS <- school_matched_small %>% filter(matched == 0) %>% nrow() +
  dim(patients_matched_small_dedup %>% filter(matched == 0) %>% select(patient_id)
      %>% unique())[1] + nrow(matches)

# Get count of autism and no autism in school data
school_ssas_grouped <- school_ssas %>%
  group_by(age_june30, sex_desc, autism) %>%
  summarise(count = n()) %>%
  pivot_wider(names_from = "autism", values_from = "count") %>%
  rename(no_autism_school = "0", yes_autism_school = "1")

# Remove matched patients because they are already counted as yes autism in
# the school data. Then get count of autism patients
patients_nomatch_grouped <- patients_matched_small_dedup %>%
  filter(matched == 0) %>%
  mutate(age_june30 = trunc(time_length(interval(ymd(dob.patient),
                                                 ymd("2021-06-30")), unit = "year"))) %>%
  rename(sex_desc = sex_desc.patient) %>%
  group_by(age_june30, sex_desc) %>%
  summarise(yes_autism_patients = n())

# Combine the counts. Move the number of autism patients from the no autism in
# school data to yes autism in school data
school_patients_grouped <- merge(school_ssas_grouped, patients_nomatch_grouped,
                                 by = c("age_june30", "sex_desc")) %>%
  mutate(no_autism_linked = no_autism_school - yes_autism_patients,
         yes_autism_linked = yes_autism_school + yes_autism_patients,
         age_cat_name = factor(ifelse(age_june30 <= 8, "6-8",
                                ifelse(age_june30 <= 11, "9-11",
                                ifelse(age_june30 <= 14, "12-14", "15-18"))),
                                levels = c("6-8", "9-11", "12-14", "15-18")),
         sex = ifelse(sex_desc == "Female", 2, 1))
```

```r
### Find updated prevalece

n_stdpop <- sum(chile_stdpop$std_pop)

aut_newprev_araucS <- school_patients_grouped %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
           # If there are no cases of autism in the group, input 0
         sample_pop_size = no_autism_linked + yes_autism_linked,
           # Total sample population is autism cases + not cases
         sample_prevalence = yes_autism_linked / sample_pop_size) %>%
           # Prevalence of autism in the group
  left_join(chile_stdpop, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
           # Prevalence of autism in the group, standardised to standard population
         w = std_pop / (sample_pop_size * n_stdpop),
           # Weight of the group using standard population
         w2 = pop_prop / sample_pop_size,
           #sum_std_pop = sum(std_pop)
           ) %>%
  ungroup()

aut_newprev_araucS_adj <- aut_newprev_araucS %>%
  summarise(sum_sample_pop_size = sum(sample_pop_size),
            crude_rate = sum(yes_autism_linked) / sum(sample_pop_size),
            crude_count = sum(yes_autism_linked),
            adjusted_rate = sum(yes_autism_linked / sample_pop_size * pop_prop),
            adjusted_count = round(adjusted_rate * sum_sample_pop_size, 0),
            var = sum(pop_prop^2 * yes_autism_linked / sample_pop_size^2),
            w_M = max(w),
            crude_ci_lower = crude_rate - (1.96 * sqrt(crude_rate *
              (1 - crude_rate) / sum_sample_pop_size)),
            crude_ci_upper = crude_rate + (1.96 * sqrt(crude_rate *
              (1 - crude_rate) / sum_sample_pop_size)),
            adjusted_ci_lower = ifelse(var == 0, 0, var / (2*adjusted_rate) *
              qchisq(p = 0.05/2, df = 2*adjusted_rate^2 / var)),
            adjusted_ci_upper = (var + w_M^2) / (2*(adjusted_rate + w_M)) *
              qchisq(p = 1-0.05/2, df = 2*(adjusted_rate+w_M)^2 / (var+w_M^2))) %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Adjusted prevalence by sex
aut_newprev_araucS_f <- school_patients_grouped %>%
  filter(sex == 2) %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
         sample_pop_size = no_autism_linked + yes_autism_linked,
         sample_prevalence = yes_autism_linked / sample_pop_size) %>%
  left_join(chile_stdpop_f, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
         w = std_pop / (sample_pop_size * n_stdpop),
         w2 = pop_prop / sample_pop_size,
           #sum_std_pop = sum(std_pop)
```

```
          ) %>%
  ungroup()

aut_newprev_araucS_adj_f <- get_adjusted_prev(rename(aut_newprev_araucS_f,
                                                    "n_disease" = "yes_autism_linked"),
                                              grouping_vars = c()) %>%
  mutate(sex_desc = "Female") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

aut_newprev_araucS_m <- school_patients_grouped %>%
  filter(sex == 1) %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
         sample_pop_size = no_autism_linked + yes_autism_linked,
         sample_prevalence = yes_autism_linked / sample_pop_size) %>%
  left_join(chile_stdpop_m, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
         w = std_pop / (sample_pop_size * n_stdpop),
         w2 = pop_prop / sample_pop_size,
           #sum_std_pop = sum(std_pop)
           ) %>%
  ungroup()

aut_newprev_araucS_adj_m <- get_adjusted_prev(rename(aut_newprev_araucS_m,
                                                    "n_disease" = "yes_autism_linked"),
                                              grouping_vars = c()) %>%
  mutate(sex_desc = "Male") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

aut_newprev_araucS_sex_adj <- rbind(aut_newprev_araucS_adj_m, aut_newprev_araucS_adj_f)

# Adjusted prevalence by age band
chile_stdpop_6 <- chile_stdpop %>%
  filter(age >= 6 & age <= 8) %>%
  mutate(pop_prop = std_pop / sum(std_pop))
chile_stdpop_9 <- chile_stdpop %>%
  filter(age >= 9 & age <= 11) %>%
  mutate(pop_prop = std_pop / sum(std_pop))
chile_stdpop_12 <- chile_stdpop %>%
  filter(age >= 12 & age <= 14) %>%
  mutate(pop_prop = std_pop / sum(std_pop))
chile_stdpop_15 <- chile_stdpop %>%
  filter(age >= 15 & age <= 18) %>%
  mutate(pop_prop = std_pop / sum(std_pop))

# Age 6-8
aut_newprev_araucS_6 <- school_patients_grouped %>%
  filter(age_cat_name == "6-8") %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
         sample_pop_size = no_autism_linked + yes_autism_linked,
```

```
        sample_prevalence = yes_autism_linked / sample_pop_size) %>%
  left_join(chile_stdpop_6, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
         w = std_pop / (sample_pop_size * n_stdpop),
         w2 = pop_prop / sample_pop_size) %>%
  ungroup()

aut_newprev_araucS_adj_6 <- get_adjusted_prev(rename(aut_newprev_araucS_6,
                                                     "n_disease" = "yes_autism_linked"),
                                              grouping_vars = c()) %>%
  mutate(sex_desc = "Male",
         age_cat_name = "6-8") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Age 9-11
aut_newprev_araucS_9 <- school_patients_grouped %>%
  filter(age_cat_name == "9-11") %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
         sample_pop_size = no_autism_linked + yes_autism_linked,
         sample_prevalence = yes_autism_linked / sample_pop_size) %>%
  left_join(chile_stdpop_9, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
         w = std_pop / (sample_pop_size * n_stdpop),
         w2 = pop_prop / sample_pop_size) %>%
  ungroup()

aut_newprev_araucS_adj_9 <- get_adjusted_prev(rename(aut_newprev_araucS_9,
                                                     "n_disease" = "yes_autism_linked"),
                                              grouping_vars = c()) %>%
  mutate(sex_desc = "Male",
         age_cat_name = "9-11") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Age 12-14
aut_newprev_araucS_12 <- school_patients_grouped %>%
  filter(age_cat_name == "12-14") %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
         sample_pop_size = no_autism_linked + yes_autism_linked,
         sample_prevalence = yes_autism_linked / sample_pop_size) %>%
  left_join(chile_stdpop_12, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
         w = std_pop / (sample_pop_size * n_stdpop),
         w2 = pop_prop / sample_pop_size) %>%
  ungroup()

aut_newprev_araucS_adj_12 <- get_adjusted_prev(rename(aut_newprev_araucS_12,
                                                      "n_disease" = "yes_autism_linked"),
                                               grouping_vars = c()) %>%
  mutate(sex_desc = "Male",
```

```
         age_cat_name = "12-14") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

# Age 15-18
aut_newprev_araucS_15 <- school_patients_grouped %>%
  filter(age_cat_name == "15-18") %>%
  rename(age = age_june30) %>%
  mutate(yes_autism_linked = ifelse(is.na(yes_autism_linked), 0, yes_autism_linked),
         sample_pop_size = no_autism_linked + yes_autism_linked,
         sample_prevalence = yes_autism_linked / sample_pop_size) %>%
  left_join(chile_stdpop_15, by = c("age", "sex")) %>%
  mutate(disease_prev_std = yes_autism_linked / sample_pop_size * pop_prop,
         w = std_pop / (sample_pop_size * n_stdpop),
         w2 = pop_prop / sample_pop_size) %>%
  ungroup()

aut_newprev_araucS_adj_15 <- get_adjusted_prev(rename(aut_newprev_araucS_15,
                                                      "n_disease" = "yes_autism_linked"),
                                               grouping_vars = c()) %>%
  mutate(sex_desc = "Male",
         age_cat_name = "15-18") %>%
  mutate(crude_ci_lower = ifelse(crude_ci_lower < 0, 0, crude_ci_lower),
         adjusted_ci_lower = ifelse(adjusted_ci_lower < 0, 0, adjusted_ci_lower))

aut_newprev_araucS_agecat_adj <- rbind(aut_newprev_araucS_adj_6,
                                       aut_newprev_araucS_adj_9,
                                       aut_newprev_araucS_adj_12,
                                       aut_newprev_araucS_adj_15)

aut_newprev_araucS_agecat_adj.table <- aut_newprev_araucS_agecat_adj %>%
  select(age_cat_name, crude_prev_ci, adjusted_prev_ci)

kbl(aut_newprev_araucS_agecat_adj.table,
  col.names = c("Age band", "Crude prevalence (95% CI)", "Adjusted prevalence (95% CI)"),
  align = "lrr",
  booktabs = TRUE,
  format = "pandoc",
  linesep = c(""),
  caption = "Age- and sex-adjusted updated autism prevalence from data linkage in SSAS by age band with

### Adjusted new prevalence and table

araucS_lower_prev <- aut_prev_health_adj %>%
  filter(health_service_name == "Araucanía Sur") %>%
  select(adjusted_rate) %>% as.numeric()
araucS_upper_prev <- aut_newprev_araucS_adj %>%
  select(adjusted_rate) %>% as.numeric()

prev_delta <- araucS_upper_prev - araucS_lower_prev
prev_delta_ci_width <- max(aut_prev_health_adj %>%
                           filter(health_service_name == "Araucanía Sur") %>%
                           select(adjusted_ci_upper) %>% as.numeric() -
```

```
                              aut_prev_health_adj %>%
                              filter(health_service_name == "Araucanía Sur") %>%
                              select(adjusted_ci_lower) %>% as.numeric(),
                            aut_newprev_araucS_adj %>%
                              select(adjusted_ci_upper) %>% as.numeric() -
                              aut_newprev_araucS_adj %>%
                              select(adjusted_ci_lower) %>% as.numeric())

araucS_lower <- aut_prev_health_adj %>%
  filter(health_service_name == "Araucanía Sur") %>%
  select(adjusted_count) %>% as.numeric()
araucS_upper <- aut_newprev_araucS_adj %>%
  select(adjusted_count) %>% as.numeric()
# araucS_mean <- (araucS_upper + araucS_lower) / 2
# araucS_sd <- (araucS_upper - araucS_lower) / (2*1.96)

aut_prev_health_adj$new_rate <- numeric(nrow(aut_prev_health_adj))
aut_prev_health_adj$new_ci_lower <- numeric(nrow(aut_prev_health_adj))
aut_prev_health_adj$new_ci_upper <- numeric(nrow(aut_prev_health_adj))
aut_prev_health_adj$new_count <- numeric(nrow(aut_prev_health_adj))

for(i in 1:nrow(aut_prev_health_adj)) {
  if(aut_prev_health_adj$health_service_name[i] == "Araucanía Sur") {
    aut_prev_health_adj$new_rate[i] <- araucS_upper_prev
    aut_prev_health_adj$new_ci_lower[i] <- aut_newprev_araucS_adj$adjusted_ci_lower
    aut_prev_health_adj$new_ci_upper[i] <- aut_newprev_araucS_adj$adjusted_ci_upper
    aut_prev_health_adj$new_count[i] <- araucS_upper
  }
  else {
    adjusted_ci_width <- aut_prev_health_adj$adjusted_ci_upper[i] - aut_prev_health_adj$adjusted_ci_lowe
    aut_prev_health_adj$new_rate[i] <- aut_prev_health_adj$adjusted_rate[i] + prev_delta
    aut_prev_health_adj$new_ci_lower[i] <- aut_prev_health_adj$new_rate[i] - 0.5 * max(adjusted_ci_widtl
    aut_prev_health_adj$new_ci_upper[i] <- aut_prev_health_adj$new_rate[i] + 0.5 * max(adjusted_ci_widtl
    aut_prev_health_adj$new_count[i] <- round(aut_prev_health_adj$new_rate[i] * aut_prev_health_adj$sum_
  }
}

aut_newprev_health_adj.table <- aut_prev_health_adj %>%
  mutate(new_prev_ci = paste0(sprintf("%.2f", round(new_rate * 100, 2)),
                              " (",
                              sprintf("%.2f", round(new_ci_lower * 100, 2)),
                              ", ",
                              sprintf("%.2f", round(new_ci_upper * 100, 2)),
                              ")")) %>%
  select(health_service_name, adjusted_prev_ci, new_prev_ci)
kbl(aut_newprev_health_adj.table,
    col.names = c("Health service", "Adjusted prevalence (95% CI)", "Adjusted updated prevalence (Maxima
    align = "lrr",
    booktabs = TRUE,
    format = "pandoc",
    linesep = c(""), # No extra white space
    caption = "Adjusted prevalence and adjusted updated prevalence of autism by health service in Chile
  add_header_above(c(" " = 1, "Adjusted prevalence" = 2))
```

```
### Bayesian projection

nObs <- nrow(chile_bayes_aut)
nIter <- 2000
nBurn <- 2000
nFeat <- length(unique(aut_prev_health_adj$health_service_code))
FeatNames <- sort(unique(aut_prev_health_adj$health_service_code))

### Prior 1

# Same prior for all health services - specifically the national prevalence
# in school data
aut_theta_mu_prior <- aut_prev_adj$adjusted_rate
aut_theta_sigma_prior <- (aut_prev_adj$adjusted_ci_upper -
                            aut_prev_adj$adjusted_ci_lower) / (2*1.96)

rand_model_constant <- "model {
  for(i in 1:nFeat) { # For each category in the feature grouping
    theta[i] ~ dbeta(theta_a, theta_b)
    disease_sample[i] ~ dbin(theta[i], nObs[i])
    disease_pred[i] ~ dbin(theta[i], nObs[i])
  }
}"

pars <- c("theta_a", "theta_b", "theta", "disease_sample", "disease_pred")

health_low_post <- do_jags_rand_model(x = aut_prev_health_adj,
                              feat = "health_service_name",
                              model = rand_model_constant,
                              theta_mu = aut_theta_mu_prior,
                              theta_sigma = aut_theta_sigma_prior,
                              pars = pars,
                              convergence_checks = FALSE) %>%
  rename("health_service_name" = "Feat_names")

# Plot posterior densities
health_const_post_ci <- health_low_post %>%
  group_by("health_service_name") %>%
  summarise(post_lower = quantile(predicted_prev, 0.025),
            post_upper = quantile(predicted_prev, 0.975))

ggplot() +
  geom_density(data = health_low_post, aes(x = predicted_prev*100)) +
  geom_vline(data = health_const_post_ci, aes(xintercept = post_lower*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = health_const_post_ci, aes(xintercept = post_upper*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_lower*100),
             color = "red", linetype = "dashed") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_upper*100),
             color = "red", linetype = "dashed") +
  facet_wrap(~health_service_name) +
  labs(title = "Posterior predictive distributions for common lower bound",
```

```
       x = "Prevalence %",
       y = "Density")

### Prior 2

# Now have health service specific priors so need to provide them as vectors
# and index each as needed
rand_model_low <- "model {
  for(i in 1:nFeat) { # For each category in the feature grouping
    theta[i] ~ dbeta(theta_a[i], theta_b[i])
    disease_sample[i] ~ dbin(theta[i], nObs[i])
    disease_pred[i] ~ dbin(theta[i], nObs[i])
  }
}"

pars <- c("theta_a", "theta_b", "theta", "disease_sample", "disease_pred")

# Define beta prior
theta_mu <- aut_prev_health_adj$adjusted_rate
theta_sigma <- (aut_prev_health_adj$adjusted_ci_upper -
                  aut_prev_health_adj$adjusted_ci_lower) / (2*1.96)
theta_a <- theta_mu * (theta_mu * (1-theta_mu) / theta_sigma^2 - 1)
theta_b <- (1 - theta_mu) * (theta_mu * (1-theta_mu) / theta_sigma^2 - 1)

# Initial values for model chains
rand_ini <- list(list(theta = rep(0.001, nFeat)),
                 list(theta = rep(0.01, nFeat)))
# Run JAGS model
rand_data <- list(theta_a = theta_a,
                  theta_b = theta_b,
                  nObs = aut_prev_health_adj$sum_sample_pop_size,
                  aut_sample = aut_prev_health_adj$adjusted_count,
                  nFeat = nFeat)
rand_jag <- jags.model(textConnection(rand_model_low),
                       data = rand_data,
                       inits = rand_ini,
                       n.chains = 2,
                       quiet = TRUE)
update(rand_jag, n.iter = nBurn)
rand_sam <- coda.samples(model = rand_jag,
                         variable.names = pars,
                         n.iter = nIter)

convergence_check <- FALSE
if(convergence_check) {
  print(mcmc_trace(rand_sam, paste0("theta[", 1:nFeat, "]")))
  print(mcmc_trace(rand_sam, paste0("disease_pred[", 1:nFeat, "]")))
  rand_summ <- summary(subset_draws(as_draws(rand_sam), pars),
                           ~quantile(.x, probs=c(0.025, 0.5, 0.975)),
                           ~mcse_quantile(.x, probs=c(0.025, 0.5, 0.975)),
                           "rhat") %>%
        arrange(desc(rhat))
  print(rand_summ)
```

```
}

aut_prev_health_post_new <- as_tibble(as_draws_matrix(rand_sam),
                                      rownames = "Iteration") %>%
    select(c("Iteration", contains("theta["))) %>%
    pivot_longer(cols = contains("theta["),
                 names_to = "health_service_code",
                 values_to = "predicted_prev") %>%
    mutate(Feat_names = factor(health_service_code,
                               levels = c(paste0("theta[",1:nFeat,"]")),
                               labels = FeatNames)) %>%
    select(Iteration, Feat_names, predicted_prev) %>%
  rename(health_service_code = Feat_names) %>%
  merge(health_service_lookup, by = "health_service_code") %>%
  rename(health_service_name_long = health_service_name,
         health_service_name = health_service_name_short)

# Plot posterior densities
aut_prev_health_post_new_ci <- aut_prev_health_post_new %>%
  group_by(health_service_name) %>%
  summarise(post_lower = quantile(predicted_prev, 0.025),
            post_upper = quantile(predicted_prev, 0.975))

ggplot() +
  geom_density(data = aut_prev_health_post_new, aes(x = predicted_prev*100)) +
  geom_vline(data = aut_prev_health_post_new_ci, aes(xintercept = post_lower*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_post_new_ci, aes(xintercept = post_upper*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_lower*100),
             color = "red", linetype = "dashed") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_upper*100),
             color = "red", linetype = "dashed") +
  facet_wrap(~health_service_name) +
  labs(title = "Posterior predictive distribution for health service specific lower bound priors",
       x = "Prevalence %",
       y = "Density")

### Prior 3

# Now have health service specific priors so need to provide them as
# vectors and index each as needed
rand_model_high <- "model {
  for(i in 1:nFeat) { # For each category in the feature grouping
    theta[i] ~ dbeta(theta_a[i], theta_b[i])
    disease_sample[i] ~ dbin(theta[i], nObs[i])
    disease_pred[i] ~ dbin(theta[i], nObs[i])
  }
}"

pars <- c("theta_a", "theta_b", "theta", "disease_sample", "disease_pred")

# Define beta prior
```

```
theta_mu <- aut_prev_health_adj$new_rate
theta_sigma <- (aut_prev_health_adj$new_ci_upper - aut_prev_health_adj$new_ci_lower) / (2*1.96)
theta_a <- theta_mu * (theta_mu * (1-theta_mu) / theta_sigma^2 - 1)
theta_b <- (1 - theta_mu) * (theta_mu * (1-theta_mu) / theta_sigma^2 - 1)

# Initial values for model chains
rand_ini <- list(list(theta = rep(0.001, nFeat)),
                 list(theta = rep(0.01, nFeat)))
# Run JAGS model
rand_data <- list(theta_a = theta_a,
                  theta_b = theta_b,
                  nObs = aut_prev_health_adj$sum_sample_pop_size,
                  aut_sample = aut_prev_health_adj$adjusted_count,
                  nFeat = nFeat)
rand_jag <- jags.model(textConnection(rand_model_high),
                       data = rand_data,
                       inits = rand_ini,
                       n.chains = 2,
                       quiet = TRUE)
update(rand_jag, n.iter = nBurn)
rand_sam <- coda.samples(model = rand_jag,
                         variable.names = pars,
                         n.iter = nIter)

convergence_check <- FALSE
if(convergence_check) {
  print(mcmc_trace(rand_sam, paste0("theta[", 1:nFeat, "]")))
  print(mcmc_trace(rand_sam, paste0("disease_pred[", 1:nFeat, "]")))
  rand_summ <- summary(subset_draws(as_draws(rand_sam), pars),
                           ~quantile(.x, probs=c(0.025, 0.5, 0.975)),
                           ~mcse_quantile(.x, probs=c(0.025, 0.5, 0.975)),
                           "rhat") %>%
        arrange(desc(rhat))
  print(rand_summ)
}

aut_prev_health_post_new <- as_tibble(as_draws_matrix(rand_sam),
                                      rownames = "Iteration") %>%
    select(c("Iteration", contains("theta["))) %>%
    pivot_longer(cols = contains("theta["),
                 names_to = "health_service_code",
                 values_to = "predicted_prev") %>%
    mutate(Feat_names = factor(health_service_code,
                               levels = c(paste0("theta[",1:nFeat,"]")),
                               labels = FeatNames)) %>%
    select(Iteration, Feat_names, predicted_prev) %>%
  rename(health_service_code = Feat_names) %>%
  merge(health_service_lookup, by = "health_service_code") %>%
  rename(health_service_name_long = health_service_name,
         health_service_name = health_service_name_short)

# Plot posterior densities
aut_prev_health_post_new_ci <- aut_prev_health_post_new %>%
```

```
  group_by(health_service_name) %>%
  summarise(post_lower = quantile(predicted_prev, 0.025),
            post_upper = quantile(predicted_prev, 0.975))

ggplot() +
  geom_density(data = aut_prev_health_post_new, aes(x = predicted_prev*100)) +
  geom_vline(data = aut_prev_health_post_new_ci, aes(xintercept = post_lower*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_post_new_ci, aes(xintercept = post_upper*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_lower*100),
             color = "red", linetype = "dashed") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_upper*100),
             color = "red", linetype = "dashed") +
  facet_wrap(~health_service_name) +
  labs(title = "Posterior predictive distribution for health service specific upper bound priors",
       x = "Prevalence %",
       y = "Density")

### Prior 4

# Now have health service specific uniform priors so need to provide them as
# vectors and index each as needed
rand_model_uniform <- "model {
  for(i in 1:nFeat) { # For each category in the feature grouping
    theta[i] ~ dunif(theta_lower[i], theta_upper[i])     #dbeta(theta_a[i], theta_b[i])
    disease_sample[i] ~ dbin(theta[i], nObs[i])
    disease_pred[i] ~ dbin(theta[i], nObs[i])
    #theta_a[i] <- theta_mu * (theta_mu * (1-theta_mu) / pow(theta_sigma, 2) - 1)
    #theta_b[i] <- (1 - theta_mu) * (theta_mu * (1-theta_mu) / pow(theta_sigma, 2) - 1)
  }
  #theta_mu ~ dunif(0, 1)
  #theta_sigma ~ dexp(1)
}"

pars <- c("theta", "disease_sample", "disease_pred")

# Define uniform prior
theta_lower <- aut_prev_health_adj$adjusted_rate
theta_upper <- aut_prev_health_adj$new_rate

# Initial values for model chains
rand_ini <- list(list(theta = rep(0.001, nFeat)),
                 list(theta = rep(0.01, nFeat)))
# Run JAGS model
rand_data <- list(theta_lower = theta_lower,
                  theta_upper = theta_upper,
                  nObs = aut_prev_health_adj$sum_sample_pop_size,
                  aut_sample = aut_prev_health_adj$adjusted_count,
                  nFeat = nFeat)
rand_jag <- jags.model(textConnection(rand_model_uniform),
                       data = rand_data,
                       inits = rand_ini,
```

```
                        n.chains = 2,
                        quiet = TRUE)
update(rand_jag, n.iter = nBurn)
rand_sam <- coda.samples(model = rand_jag,
                         variable.names = pars,
                         n.iter = nIter)

convergence_check <- FALSE
if(convergence_check) {
  print(mcmc_trace(rand_sam, paste0("theta[", 1:nFeat, "]")))
  print(mcmc_trace(rand_sam, paste0("disease_pred[", 1:nFeat, "]")))
  rand_summ <- summary(subset_draws(as_draws(rand_sam), pars),
                           ~quantile(.x, probs=c(0.025, 0.5, 0.975)),
                           ~mcse_quantile(.x, probs=c(0.025, 0.5, 0.975)),
                           "rhat") %>%
        arrange(desc(rhat))
  print(rand_summ)
}

aut_prev_health_post_new <- as_tibble(as_draws_matrix(rand_sam),
                                      rownames = "Iteration") %>%
    select(c("Iteration", contains("theta["))) %>%
    pivot_longer(cols = contains("theta["),
                 names_to = "health_service_code",
                 values_to = "predicted_prev") %>%
    mutate(Feat_names = factor(health_service_code,
                               levels = c(paste0("theta[",1:nFeat,"]")),
                               labels = FeatNames)) %>%
    select(Iteration, Feat_names, predicted_prev) %>%
  rename(health_service_code = Feat_names) %>%
  merge(health_service_lookup, by = "health_service_code") %>%
  rename(health_service_name_long = health_service_name,
         health_service_name = health_service_name_short)


# Plot posterior densities
aut_prev_health_post_new_ci <- aut_prev_health_post_new %>%
  group_by(health_service_name) %>%
  summarise(post_lower = quantile(predicted_prev, 0.025),
            post_upper = quantile(predicted_prev, 0.975))

ggplot() +
  geom_density(data = aut_prev_health_post_new, aes(x = predicted_prev*100)) +
  geom_vline(data = aut_prev_health_post_new_ci, aes(xintercept = post_lower*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_post_new_ci, aes(xintercept = post_upper*100),
             color = "blue", linetype = "dotted") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_lower*100),
             color = "red", linetype = "dashed") +
  geom_vline(data = aut_prev_health_adj, aes(xintercept = adjusted_ci_upper*100),
             color = "red", linetype = "dashed") +
  facet_wrap(~health_service_name) +
  labs(title = "Posterior predicitve distribution for health service specific uniform priors",
```

```
       x = "Prevalence %",
       y = "Density")
```

```
### Age band table

# Autism
aut_prev.agecat <- get_grouped_prev_plot(x = chile_bayes_aut,
                                         grouping_vars = c("age_cat_name", "autism"),
                                         disease = "autism") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

aut_prev.agecat.table <- aut_prev.agecat %>%
  select(age_cat_name, n_disease, prev_ci)

kbl(aut_prev.agecat.table,
    col.names = c("Age band", "Autism cases", "Prevalence % (95% CI)"),
    align = "rrr",
    booktabs = TRUE,
    format = "latex",
    caption = "Count and prevalence of autism cases by age band in Chile school data with normal confid
  kable_styling(latex_options = "HOLD_position")

# ADHD
adhd_prev.agecat <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                         grouping_vars = c("age_cat_name", "adhd"),
                                         disease = "adhd") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

adhd_prev.agecat.table <- adhd_prev.agecat %>%
  select(age_cat_name, n_disease, prev_ci)

kbl(adhd_prev.agecat.table,
    col.names = c("Age band", "ADHD cases", "Prevalence % (95% CI)"),
    align = "rrr",
    booktabs = TRUE,
    format = "latex",
    caption = "Count and prevalence of ADHD cases by age band in Chile school data with normal confiden
  kable_styling(latex_options = "HOLD_position")

### Sex table

# Autism
aut_prev.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                         grouping_vars = c("age_cat_name", "sex_desc", "autism"),
                                         disease = "autism") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

aut_prev.agecat_sex.table <- aut_prev.agecat_sex %>% filter(sex_desc == "Female") %>%
  select(age_cat_name, n_disease, prev_ci) %>%
  cbind(aut_prev.agecat_sex %>% filter(sex_desc == "Male") %>% select(n_disease, prev_ci))

kbl(aut_prev.agecat_sex.table,
    col.names = c("Age band", "Autism cases", "Prevalence % (95% CI)", "Autism cases", "Prevalence % (95
    align = "rrrrr",
```

```
    booktabs = TRUE,
    format = "latex",
    caption = "Count and prevalence of autism cases by age band in Chile school data for females and mal
  add_header_above(header = c(" " = 1, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = "HOLD_position")

# ADHD
adhd_prev.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                              grouping_vars = c("age_cat_name", "sex_desc", "adhd"),
                                              disease = "adhd") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

adhd_prev.agecat_sex.table <- aut_prev.agecat_sex %>% filter(sex_desc == "Female") %>%
  select(age_cat_name, n_disease, prev_ci) %>%
  cbind(adhd_prev.agecat_sex %>% filter(sex_desc == "Male") %>% select(n_disease, prev_ci))

kbl(adhd_prev.agecat_sex.table,
    col.names = c("Age band", "ADHD cases", "Prevalence % (95% CI)", "ADHD cases", "Prevalence % (95% CI
    align = "rrrrr",
    booktabs = TRUE,
    format = "latex",
    caption = "Count and prevalence of ADHD cases by age band in Chile school data for females and males
  add_header_above(c(" " = 1, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = "HOLD_position")

### Health service table

# Autism
aut_prev_health.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                                    grouping_vars = c("health_service_name", "age_cat_na
                                                    disease = "autism") %>%
  arrange(age_cat_name)

aut_prev_health.agecat_sex.table <- aut_prev_health.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(health_service_name, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(aut_prev_health.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(aut_prev_health.agecat_sex.table,
  col.names = c("Health service", "Age band", "Autism cases", "Prevalence % (95% CI)", "Autism cases",
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  longtable = TRUE,
  linesep = c("", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "", "",
```

```
              "", "", "", "\\addlinespace"), # Extra white space after every 29th line
  caption = "Count and prevalence of autism cases by health service and age band in Chile school data fo
  add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = c("repeat_header"), font_size = 8) %>%
  kable_styling(latex_options = "HOLD_position")

# ADHD
adhd_prev_health.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                          grouping_vars = c("health_service_name", "age_cat_name", "sex_
                                          disease = "adhd") %>%
  arrange(age_cat_name)

adhd_prev_health.agecat_sex.table <- adhd_prev_health.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(health_service_name, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(adhd_prev_health.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(adhd_prev_health.agecat_sex.table,
  col.names = c("Health service", "Age band", "ADHD cases", "Prevalence % (95% CI)", "ADHD cases", "Prev
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  longtable = TRUE,
  linesep = c("", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "", "",
              "", "", "", "\\addlinespace"), # Extra white space after every 29th line
  caption = "Count and prevalence of ADHD cases by health service and age band in Chile school data for
  add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = c("repeat_header"), font_size = 8) %>%
  kable_styling(latex_options = "HOLD_position")

### Health service plot

# Autism
aut_prev_health.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                          grouping_vars = c("health_service_name" ,"age_cat_name", "sex_
                                          disease = "autism") %>%
  arrange(age_cat_name)

ggplot(data = aut_prev_health.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~health_service_name) +
  labs(title = "Autism prevalence by health service",
       x = "Sex",
```

```
       y = "Crude prevalence (95% CI)",
       fill = "Age category")

# ADHD
adhd_prev_health.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                          grouping_vars = c("health_service_name", "age_cat_name", "sex
                                          disease = "adhd") %>%
  arrange(age_cat_name)

ggplot(data = adhd_prev_health.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~health_service_name) +
  labs(title = "ADHD prevalence by health service",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age band")

### SES table

# Autism
aut_prev_econA.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                                   grouping_vars = c("school_fee", "age_cat_name", "sex
                                                   disease = "autism") %>%
  arrange(age_cat_name)

aut_prev_econA.agecat_sex.table <- aut_prev_econA.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(school_fee, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(aut_prev_econA.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(aut_prev_econA.agecat_sex.table,
  col.names = c("School fee", "Age band", "Autism cases", "Prevalence % (95% CI)", "Autism cases", "Prev
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  #longtable = TRUE,
  linesep = c("", "", "", "", "", "", "\\addlinespace"), # Extra white space after every 7th line
  caption = "Count and prevalence of Autism cases by school fee and age band in Chile school data for fe
  add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = "HOLD_position")

# ADHD
adhd_prev_econA.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                                   grouping_vars = c("school_fee", "age_cat_name", "sex
                                                   disease = "adhd") %>%
  arrange(age_cat_name)
```

```
adhd_prev_econA.agecat_sex.table <- adhd_prev_econA.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(school_fee, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(adhd_prev_econA.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(adhd_prev_econA.agecat_sex.table,
  col.names = c("School fee", "Age band", "ADHD cases", "Prevalence % (95% CI)", "ADHD cases", "Prevale
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  #longtable = TRUE,
  linesep = c("", "", "", "", "", "", "\\addlinespace"), # Extra white space after every 7th line
  caption = "Count and prevalence of ADHD cases by school fee and age band in Chile school data for fema
add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = "HOLD_position")

### SES plot

# Autism
aut_prev_econA.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                                   grouping_vars = c("school_fee", "age_cat_name", "sex
                                                   disease = "autism") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = aut_prev_econA.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~school_fee) +
  labs(title = "Autism prevalence by school fee",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age category")

# ADHD
adhd_prev_econA.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                                   grouping_vars = c("school_fee", "age_cat_name", "sex
                                                   disease = "adhd") %>%
  arrange(age_cat_name)

ggplot(data = adhd_prev_econA.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~school_fee) +
  labs(title = "ADHD prevalence by school fee",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age category")
```

```r
### Ethnicity table

# Autism
aut_prev_ethnic.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut_ethnic,
                                                    grouping_vars = c("ethnic_2_group", "age_cat_name",
                                                    disease = "autism") %>%
  arrange(age_cat_name)

aut_prev_ethnic.agecat_sex.table <- aut_prev_ethnic.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(ethnic_2_group, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(aut_prev_ethnic.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(aut_prev_ethnic.agecat_sex.table,
  col.names = c("Ethnicity", "Age band", "Autism cases", "Prevalence % (95% CI)", "Autism cases", "Preva
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  #longtable = TRUE,
  linesep = c("", "", "\\addlinespace"), # Extra white space after every 3rd line
  caption = "Count and prevalence of Autism cases by ethnicity and age band in Chile school data for fe
  add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(font_size = 8) %>%
  kable_styling(latex_options = "HOLD_position")

# ADHD
adhd_prev_ethnic.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd_ethnic,
                                                    grouping_vars = c("ethnic_2_group", "age_cat_name",
                                                    disease = "adhd") %>%
  arrange(age_cat_name)

adhd_prev_ethnic.agecat_sex.table <- adhd_prev_ethnic.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(ethnic_2_group, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(adhd_prev_ethnic.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(adhd_prev_ethnic.agecat_sex.table,
  col.names = c("Ethnicity", "Age band", "ADHD cases", "Prevalence % (95% CI)", "ADHD cases", "Prevalen
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  #longtable = TRUE,
  linesep = c("", "", "\\addlinespace"), # Extra white space after every 3rd line
  caption = "Count and prevalence of ADHD cases by ethnicity and age band in Chile school data for fema
add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
```

```
  kable_styling(font_size = 8) %>%
  kable_styling(latex_options = "HOLD_position")

### Ethnicity plot

# Autism
aut_prev_ethnic.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut_ethnic,
                                                    grouping_vars = c("ethnic_2_group", "age_cat_name",
                                                    disease = "autism") %>%
  arrange(age_cat_name)

ggplot(data = aut_prev_ethnic.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~ethnic_2_group) +
  labs(title = "Autism prevalence by ethnicity",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age category")

# ADHD
adhd_prev_ethnic.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd_ethnic,
                                                    grouping_vars = c("ethnic_2_group", "age_cat_name",
                                                    disease = "adhd") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = adhd_prev_ethnic.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~ethnic_2_group) +
  labs(title = "ADHD prevalence by ethnicity",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age category")

### Rurality table

# Autism
aut_prev_rural.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                                    grouping_vars = c("school_rurality", "age_cat_name"
                                                    disease = "autism") %>%
  arrange(age_cat_name)

aut_prev_rural.agecat_sex.table <- aut_prev_rural.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(school_rurality, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(aut_prev_rural.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))
```

```
kbl(aut_prev_rural.agecat_sex.table,
  col.names = c("School rurality", "Age band", "Autism cases", "Prevalence % (95% CI)", "Autism cases",
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  #longtable = TRUE,
  linesep = c("", "\\addlinespace"), # Extra white space after every 2nd line
  caption = "Count and prevalence of Autism cases by school's rurality and age band in Chile school data
  add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = "HOLD_position")

# ADHD
adhd_prev_rural.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                                    grouping_vars = c("school_rurality", "age_cat_name"
                                                    disease = "adhd") %>%
  arrange(age_cat_name)

adhd_prev_rural.agecat_sex.table <- adhd_prev_rural.agecat_sex %>%
  filter(sex_desc == "Female") %>%
  mutate(n_disease_f = ifelse(n_disease < 20, "<20", n_disease)) %>%
  select(school_rurality, age_cat_name, n_disease_f, prev_ci) %>%
  cbind(adhd_prev_rural.agecat_sex %>%
          filter(sex_desc == "Male") %>%
          mutate(n_disease_m = ifelse(n_disease < 20, "<20", n_disease)) %>%
          select(n_disease_m, prev_ci))

kbl(adhd_prev_rural.agecat_sex.table,
  col.names = c("School rurality", "Age band", "ADHD cases", "Prevalence % (95% CI)", "ADHD cases", "Pr
  align = "lrrrrr",
  booktabs = TRUE,
  format = "latex",
  #longtable = TRUE,
  linesep = c("", "\\addlinespace"), # Extra white space after every 2nd line
  caption = "Count and prevalence of ADHD cases by school's rurality and age band in Chile school data
add_header_above(c(" " = 2, "Female" = 2, "Male" = 2)) %>%
  kable_styling(latex_options = "HOLD_position")

### Rurality plot

# Autism
aut_prev_rural.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_aut,
                                                    grouping_vars = c("school_rurality", "age_cat_name"
                                                    disease = "autism") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = aut_prev_rural.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~school_rurality) +
  labs(title = "Autism prevalence by school's rurality",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
```

```
       fill = "Age category")

# ADHD
adhd_prev_rural.agecat_sex <- get_grouped_prev_plot(x = chile_bayes_adhd,
                                                    grouping_vars = c("school_rurality", "age_cat_name"
                                                    disease = "adhd") %>%
  mutate(ci_lower = ifelse(ci_lower < 0, 0, ci_lower))

ggplot(data = adhd_prev_rural.agecat_sex) +
  geom_col(aes(x = sex_desc, y = sample_prevalence*100, group = age_cat_name, fill = age_cat_name), posi
  geom_errorbar(aes(x = sex_desc, ymin = ci_lower*100, ymax = ci_upper*100, group = age_cat_name), width
  scale_fill_manual(values = c(plasma_colours[3], plasma_colours[5], plasma_colours[7], plasma_colours[9
  facet_wrap(~school_rurality) +
  labs(title = "ADHD prevalence by school's rurality",
       x = "Sex",
       y = "Crude prevalence % (95% CI)",
       fill = "Age category")

### Kolmogorov plots

# School by sex
kolmog_school_sex.bar <- ggplot(school_match.sex) +
  geom_col(aes(x = factor(matched, labels = c("Unmatched", "Matched")),
               y = freq,
               fill = factor(matched, labels = c("Unmatched", "Matched")))) +
  theme(legend.position = "none") +
  facet_wrap(~sex_desc) +
  labs(title = "Matched status of SSAS school records by sex",
       x = "Matched status",
       y = "Feature frequency",
       fill = "Matched status")

kolmog_school_sex.density <- ggplot(ks_perm.school.pvals, aes(x = sex, y = after_stat(density))) +
  geom_density() +
  geom_vline(xintercept = ks.school.sex$p.value, color = "red") +
  labs(title = "Kolmogorov-Smirnov permutation test on matched status of SSAS school records by sex, wi
       x = "Sex",
       y = "Density")

grid.arrange(kolmog_school_sex.bar, kolmog_school_sex.density, nrow = 2)

# Patients by sex
kolmog_patient_sex.bar <- ggplot(patients_match.sex) +
  geom_col(aes(x = factor(matched, labels = c("Unmatched", "Matched")),
               y = freq,
               fill = factor(matched, labels = c("Unmatched", "Matched")))) +
  theme(legend.position = "none") +
  facet_wrap(~sex_desc) +
  labs(title = "Matched status of SSAS patient records by sex",
       x = "Matched status",
       y = "Feature frequency",
       fill = "Matched status")
```

```r
kolmog_patient_sex.density <- ggplot(ks_perm.patients.pvals, aes(x = sex, y = after_stat(density))) +
  geom_density() +
  geom_vline(xintercept = ks.patients.sex$p.value, color = "red") +
  labs(title = "Kolmogorov-Smirnov permutation test on matched status of patient records by sex, with ob
       x = "Sex",
       y = "Density")

grid.arrange(kolmog_patient_sex.bar, kolmog_patient_sex.density, nrow = 2)

# School by commune
kolmog_school_commune.bar <- ggplot(school_match.student_commune_name) +
  geom_col(aes(x = factor(matched, labels = c("Unmatched", "Matched")),
               y = freq,
               fill = factor(matched, labels = c("Unmatched", "Matched")))) +
  theme(legend.position = "none") +
  facet_wrap(~student_commune_name.school, scales = "fixed") +
  labs(title = "Matched status of SSAS school records by commune",
       x = "Matched status",
       y = "Feature frequency",
       fill = "Matched status")
# most of the difference in matched commune frequency is for Temuco which is the biggest commune.

kolmog_school_commune.density <- ggplot(ks_perm.school.pvals, aes(x = commune_code, y = after_stat(densi
  geom_density() +
  geom_vline(xintercept = ks.school.commune_code$p.value, color = "red") +
  labs(title = "Kolmogorov-Smirnov permutation test on matched status of SSAS school records by commune
       x = "Commune",
       y = "Density")

grid.arrange(kolmog_school_commune.bar, kolmog_school_commune.density, nrow = 2)

# Patients by commune
kolmog_patient_commune.bar <- ggplot(patients_match.student_commune_name) +
  geom_col(aes(x = factor(matched, labels = c("Unmatched", "Matched")),
               y = freq,
               fill = factor(matched, labels = c("Unmatched", "Matched")))) +
  theme(legend.position = "none") +
  facet_wrap(~student_commune_name.patient, scales = "fixed") +
  labs(title = "Matched status of SSAS patient records by commune",
       x = "Matched status",
       y = "Feature frequency",
       fill = "Matched status")

kolmog_patient_commune.density <- ggplot(ks_perm.patients.pvals, aes(x = commune_code, y = after_stat(de
  geom_density() +
  geom_vline(xintercept = ks.patients.commune_code$p.value, color = "red") +
  labs(title = "Kolmogorov-Smirnov permutation test on matched status of patient records by commune, wi
       x = "Commune",
       y = "Density")

grid.arrange(kolmog_patient_commune.bar, kolmog_patient_commune.density, nrow = 2)

# School by SES
```

```r
kolmog_school_ses.bar <- ggplot(school_match.ses_status) +
  geom_col(aes(x = factor(matched, labels = c("Unmatched", "Matched")),
               y = freq,
               fill = factor(matched, labels = c("Unmatched", "Matched")))) +
  theme(legend.position = "none") +
  facet_wrap(~ses_status.school) +
  labs(title = "Matched status of SSAS school records by SES",
       x = "Matched status",
       y = "Feature frequency",
       fill = "Matched status")

kolmog_school_ses.density <- ggplot(ks_perm.school.pvals, aes(x = ses_status, y = after_stat(density)))
  geom_density() +
  geom_vline(xintercept = ks.school.ses_status$p.value, color = "red") +
  labs(title = "Kolmogorov-Smirnov permutation test on matched status of SSAS school records by SES, wi
       x = "Proxy SES",
       y = "Density")

grid.arrange(kolmog_school_ses.bar, kolmog_school_ses.density, nrow = 2)

# Patients by SES
kolmog_patient_ses.bar <- ggplot(patients_match.ses_status) +
  geom_col(aes(x = factor(matched, labels = c("Unmatched", "Matched")),
               y = freq,
               fill = factor(matched, labels = c("Unmatched", "Matched")))) +
  theme(legend.position = "none") +
  facet_wrap(~ses_status.patient) +
  labs(title = "Matched status of SSAS patient records by SES",
       x = "Matched status",
       y = "Feature frequency",
       fill = "Matched status")

kolmog_patient_ses.density <- ggplot(ks_perm.patients.pvals, aes(x = ses_status, y = after_stat(density)
  geom_density() +
  geom_vline(xintercept = ks.patients.ses_status$p.value, color = "red") +
  labs(title = "Kolmogorov-Smirnov permutation test on matched status of patient records by SES, with ob
       x = "Proxy SES",
       y = "Density")

grid.arrange(kolmog_patient_ses.bar, kolmog_patient_ses.density, nrow = 2)
```

11 Appendix B | Research Protocol

See overleaf.

Investigating the autism diagnostic pathway through unsupervised machine learning and clinical record data linkage | Research protocol

1 Introduction

Autism Spectrum Disorder (ASD) is a collection of neurodevelopmental conditions characterised by social difficulties and restricted or repetitive behaviours (1). Autism is typically diagnosed in childhood through assessment of behaviour, cognitive function and developmental milestone attainment, using standardised definitions of the condition such as the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition, and through interviews such as the Autism Diagnostic Observation Schedule and the Autism Diagnostic Interview-Revised that can be completed by caregivers, individuals or clinicians (1, 2, 3).

The estimated prevalence of autism is 1-2% however prevalence differs with sex, ethnicity, socio-economic status and geography (1, 4). Meta-analysis suggests that in childhood and adolescence, three males are diagnosed for every female and females are more at risk of being diagnosed later or not at all (5). Among English school students, prevalence is highest in students aged 15-19, with age-specific prevalence differences across sex, region, first language spoken, access to free school meals and ethnicity, with prevalence highest among Black students at 2.11% and lowest among Roma/Irish Traveller students at 0.53% which likely reflects both true prevalence differences and differing access to diagnosis (4). Many people with autism have co-diagnoses of developmental and psychiatric conditions such as behavioural problems, ADHD, anxiety and depression which can mask the symptoms of autism and delay diagnosis by several years (6). There are clear links between some socio-demographic features and diagnosis age, but a gap in understanding how they interact to impact age (4).

It is important to understand the relationships between demographics and autism diagnosis age because early diagnosis provides individuals with earlier access to support and management of co-morbidities. Early interventions, particularly behavioural management such as Early Start Denver Model, are associated with positive outcomes including improved cognitive development, adaptive behaviours and IQ, and downgraded diagnoses (7-10).

This project will address the gap in understanding the relative contributions of socio-demographic features on autism diagnosis age using unsupervised machine learning (ML). Unsupervised ML is a powerful tool for discovering latent patient clusters in electronic health record data (11) and this project will use the ML techniques of Confirmatory Factor Analysis (CFA) and clustering to quantify association of socio-demographic features with diagnosis age groups in a large clinical record dataset. Unsupervised ML is

an appropriate choice because it does not impose any preconceptions on the data and therefore allows unexpected insights to emerge which may not be possible with supervised ML (11). Unsupervised ML is frequently used to enhance autism diagnosis techniques (12) but there is little evidence of its use in understanding patterns of diagnosis, as will be conducted here. This project will explore whether individuals diagnosed later have distinct demographic profiles to those diagnosed earlier which is important to understand because autism is a spectrum disorder with heterogeneous presentation (1). This could allow additional diagnostic resources to be directed to those more likely to have later diagnoses and hence earlier treatment could be provided.

This project will also be the first of its kind to validate school records of autism from the English national school census in the National Pupil Database against clinical records which will be valuable to understand the validity of inferences drawn from schools data.

2 Aims

This project aims to better understand the features that influence autism diagnosis age. To that end, its objectives are:

1. To quantify the association between socio-demographic features and age at autism diagnosis by developing ML techniques to classify clinical records into diagnosis age groups.
2. To create a novel validation of school autism diagnosis records using clinical diagnosis records.

3 Methods

3.1 Data collection

This project will use data from the Cambridgeshire and Peterborough NHS Foundation Trust (CPFT) Research Database (the clinical data). It contains anonymised clinical records of approximately 260,000 patients that have received treatment from CPFT since 2005 and are primarily resident in Cambridgeshire county which has population of approximately 1 million (13). This project will use data for approximately 3,000 patients diagnosed with autism at age 19 or younger.

The project will also use data from the Department for Education's (DfE) National Pupil Database (NPD) which contains anonymised records of demographics, schooling attendance and attainment of approximately 7.5 million pupils enrolled in state-funded education in England each year (the schools data) (4, 13-14). This project will use data for children enrolled in Cambridgeshire County since 2013.

Data from both datasets will be collected under data-sharing agreements between CPFT and the University of Cambridge and will be provided by CPFT. An ethics approval request for this project has been submitted to the Cambridge Psychology Research Ethics Committee and it will also be assessed under the CPFT research database governance process.

3.2 Data management

All data will be stored in the School of Clinical Medicine's Secure Data Hosting Service (SDHS). It will be deleted at conclusion of the project.

The statistical software package R will be used to clean, analyse and visualise data.

R analysis scripts and version control will be managed in GitHub and will be available at https://github.com/delatee/Autism-diagnosis-age-ML. Raw data and potentially sensitive information such as individual identifiers and data for groups with fewer than ten members will not be uploaded to GitHub.

Any additional data management requirements of CPFT, the SDHS, and ethics approval will be complied with.

3.3 Data cleaning and assessment

The clinical and schools data will be cleaned using R's {tidyverse} package to ensure comparability across features and records. This will include:

- Removing duplicated records.
- Standardising units of numerical features, e.g. age measured in years rather than years and months.
- Standardising text entries for categorical features, e.g. mapping free-text input for ethnicity to the NPD's ethnic groups (4).

The quality and completeness of the data will be assessed to understand the strength of conclusions it can provide. This will include:

- Assessing missingness.
- Identifying statistical distributions likely to have generated features.
- Identifying outliers and removing if appropriate.

3.4 Features of interest

To achieve the first objective of identifying socio-demographic features associated with autism diagnosis age, this project will select all socio-demographic features that are available in the clinical data and have known or suspected association with autism diagnosis, as shown in Table 1.

Feature name	Included values	Data type
Autism status (ICD-10 codes) (15)	• F84.0 (Autistic disorder) • F84.1 (Atypical autism) • F84.5 (Asperger syndrome) • F84.8 (Other pervasive developmental disorders) • F84.9 (Pervasive developmental disorder, unspecified)	Categorical
Age at diagnosis (years)	0-19	Numerical
Diagnosis route (NB: this feature will be exploratory)	• Education-based assessment • NHS clinical assessment • Private clinical assessment	Categorical
Co-diagnoses (4, 16)	• Intellectual disability • Attention-deficit hyperactivity disorder (ADHD) • Bipolar disorder • Anxiety • Depression & mood disorders • Psychosis • Learning difficulties • Sensory impairment including blindness and deafness	Categorical
Sex	• Female • Male	Categorical
Residential Output Area (OA) or Lower Layer Super Output Area (LSOA)	OAs or LSOAs in Cambridgeshire	Categorical
Deprivation quintile of OA or LSOA of residence	1-5	Categorical
NPD ethnicity categories (4)	• Any other racial/ethnic group • Asian • Black • Chinese • Mixed • Unclassified • White • Roma/Irish Traveller	Categorical
Ethnic density, i.e. the density of NPD ethnicity categories within residential OA or LSOA (17)	0-100% of LSOA comprises ethnic minorities	Numerical

Table 1: Features in the CPFT clinical records dataset that are anticipated to be used in analysis, the values for which data will be included, and data types.

To achieve the second objective of validating diagnoses across records, this project will additionally use socio-demographic features available in the schools data, as shown in Table 2. These features have been selected because they are relevant to the autism diagnosis process and are valuable to validate across records.

Feature name	Included values	Data type
Special Educational Needs and Disability (SEND) status	• Specific learning difficulty • Moderate learning difficulty • Severe learning difficulty • Profound and multiple learning difficulties • Speech, language, and communication needs • Hearing impairment • Visual impairment • Multisensory impairment • Physical disability • ASD • Other difficulty/disability • Social, emotional, and mental health • SEND status present without specialist assessment • Unclassified condition • No SEND status	Categorical
Age at diagnosis (years)	0-19	Numerical
Diagnosis route	• NHS clinical assessment • Private clinical assessment	Categorical
Sex	• Female • Male	Categorical
Residential Output Area (OA) or Lower Layer Super Output Area (LSOA)	OAs or LSOAs in Cambridgeshire	Categorical

Table 2: Features in the NPD school dataset that are anticipated to be used in linkage, the values for which data will be included, and data types.

If data assessment shows low data quality for some features, they may be excluded from further analysis.

3.5 Data preparation

In both datasets, autism diagnosis age will be mapped to the outcome feature, age group categories: 0-4, 5-9, 10-14, and 15-19 years. Categorical features will be transformed into dummy variables using the R package {tidymodels}. Geographic information will be transformed into additional features such rurality.

The clinical data will be split into training and test sets with respective ratios of 60:40, 70:30 and 80:20 and sensitivity analysis will be conducted to choose the optimal ratio. The below data preparation will be performed with the training set, the same preparations will then be applied to the testing set to ensure data in the testing set does not inform preparation of the training set.

If data assessment identified missingness, Multiple Imputation by Chained Equations (MICE) will be used to fill these gaps. Sensitivity analysis on MICE imputed values will be performed to validate their stability. MICE is an appropriate choice because it imputes plausible values from features' own distributions and is informed by each feature's relationship with all other features meaning covariance is maintained.

It is anticipated that some features, such as ethnicity, will have unbalanced classes and down-sampling and upweighting will be used to ensure majority classes do not dominate the models.

3.6 Model selection and fitting

Unsupervised machine learning will be conducted on the clinical data, using CFA then clustering. The outcome to be modelled will be diagnosis age group and all other features will be predictive. Although the data has labelled outcomes and supervised learning could be used, it is appropriate to use unsupervised learning because it allows hidden patterns and as-yet unknown relationships to be uncovered.

The factor constructs for CFA will be taken from analysis currently being conducted by Dr Varun Warrier, University of Cambridge Department of Psychiatry. CFA is an appropriate choice for this project because there is existing knowledge of causative factors which this technique will be able to quantify, and the predictive features are likely to be somewhat correlated. However, these features are unlikely to have common variance meaning principal component analysis (PCA) would be unsuitable.

Five repeats of *k*-fold cross-validation with stratification on age group will be used to avoid overfitting model parameters and ensure subsets are representative of age group distribution. *k* will be chosen based on accuracy, stability across seed values and computational runtime.

CFA with R's {lavaan} package will be used to reduce the dimensionality of the data by transforming it into principal factors that explain much of its variance. The number of factors to retain and the choice of threshold for variance they explain will be informed by

the data and scree plots of root mean squared errors will be used to choose the optimal number of components.

Clustering methods will be trialled to find a model that recovers the four known diagnosis age clusters, including *k*-means and hierarchical clustering with single, complete or average linkage measures, as appropriate to the data. These methods are appropriate because the number of clusters is known in advance. Density-based spatial clustering of applications with noise (DBSCAN) will also be trialled to validate whether four is the right number of clusters.

For each clustering method, Manhattan distances will be used as the data has high dimensionality, and grid search with comparison of accuracies will be used to tune hyperparameters.

The R packages {caret}, {pfa}, {stats}, {hclust} and {dbscan} will be used for model building, tuning and optimisation.

3.7 Assessing model fit

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) will be calculated for each model to quantify their predictive efficiency and choose the most parsimonious model. The optimal clustering method with optimal hyperparameters will be fitted to the testing data, and the accuracy, sensitivity, specificity, AIC, BIC, Comparative Fit Index and Tucker Lewis Index will be calculated to inform goodness of fit decisions.

Silhouette scores will be used to measure the similarity of cluster members. Confusion matrices will be created to understand how well individuals cluster into diagnosis age groups. Collinearity will be assessed using variance factor inflation; if high multicollinearity is found, implicated features may be removed and the models fitted again.

Further analysis may be conducted as appropriated based on findings from the above method.

3.8 Diagnosis validation

The clinical records will be linked to the school records using CPFT's Clinical Data Linkage Service (CDLS).

Confusion matrices will be constructed to validate autism diagnoses in the schools data against the clinical data.

4 Discussion

4.1 Strengths

Unsupervised ML can uncover relationships between features that could not be identified manually. CFA can identify clusters of individuals based on latent factors which could validate suspected causes of delayed autism diagnoses. This project's use of real-world data rather than simulation provides greater opportunity for novel findings as preconceptions are not embedded in the model and the large sample size increases statistical power and generalisability.

This project builds upon previous and ongoing research by Roman-Urrestarazu et al (4, 18) that used NPD data and will be strengthened by their learnings. CFA will validate current understanding of latent factors involved in diagnosis age and linkage to schools data will provide local validation of national diagnosis-age findings.

4.2 Limitations

The CFA can quantify the contribution of demographic factors to diagnosis age but cannot determine whether such relationships are causative; further work, such as randomised control trials, would be needed to establish causality.

The quality of the clinical data will largely determine the quality of this project's findings. While CPFT has done some clinical data pre-processing, it is not yet known how complete or standardised this data is. Many missing values would reduce the power of the model and many free-text fields would introduce additional error through parsing free-text into categorical features. Clinical coding does not map exactly to diagnoses meaning some autism diagnoses will be missed or reported erroneously in the clinical data, and therefore valid school diagnoses would be missed.

The schools data includes the 93% of students in state-funded education but not the 7% in independent schools or alternative arrangements meaning this dataset is not comprehensive and some valid clinical diagnoses will be missed (4). Additional clinical diagnoses will be missing from the school data as some population groups such as Roma/Irish Travellers are less likely to complete schooling (4).

4.3 Risks

Timely access to data poses a risk to this project; data requests and ethics approval have been sought. If timely access is not possible, a smaller set of schools data is currently available that could be used for unsupervised ML using much of the above method.

Unintended disclosure of personal information is possible if a collection of socio-demographic factors is unique to a person. This project will follow the UK Data Service's guidelines including not publishing data for groups with fewer than ten members (19).

5 Timeline

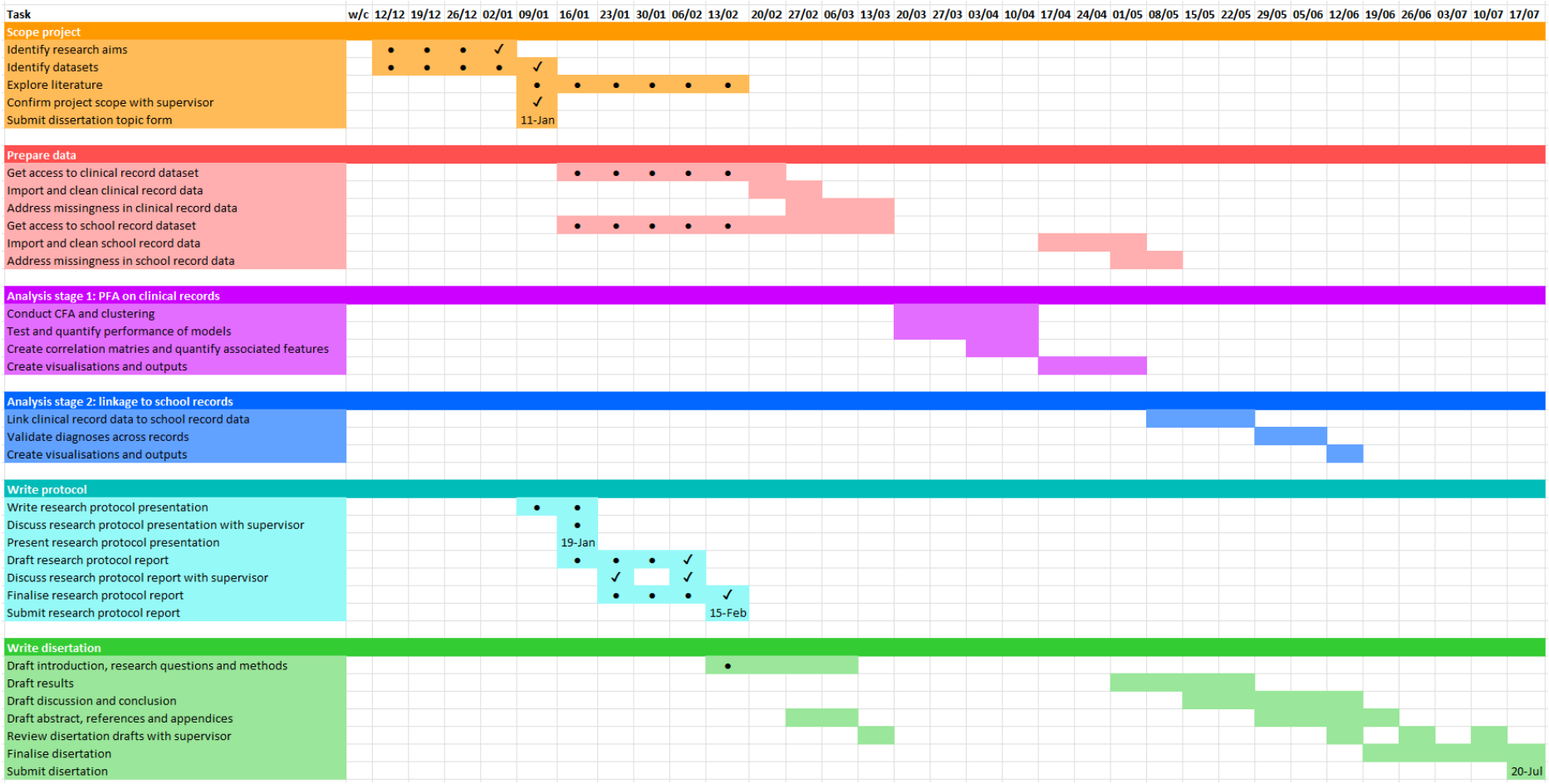


Figure 1: Timeline detailing activities and write up of the project. ● indicates an ongoing task, ✓ indicates completed task.

6 References

1. Lai MC, Lombardo MV, Baron-Cohen S. Autism. The Lancet. 2014 Mar 8;383(9920):896–910.

2. Hus Y, Segal O. Challenges Surrounding the Diagnosis of Autism in Children. Neuropsychiatr Dis Treat. 2021 Sep 3;Volume 17:3509–29.

3. Bildt A de, Sytema S, Ketelaars C, Kraijer D, Mulder E, Volkmar F, et al. Interrelationship Between Autism Diagnostic Observation Schedule-Generic (ADOS-G), Autism Diagnostic Interview-Revised (ADI-R), and the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) Classification in Children and Adolescents with Mental Retardation. J Autism Dev Disord. 2004 Apr;34(2):129–37.

4. Roman-Urrestarazu A, van Kessel R, Allison C, Matthews FE, Brayne C, Baron-Cohen S. Association of Race/Ethnicity and Social Disadvantage With Autism Prevalence in 7 Million School Children in England. JAMA Pediatr. 2021 Jun 7;175(6):e210054–e210054.

5. Loomes R, Hull L, Mandy WPL. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. J Am Acad Child Adolesc Psychiatry. 2017 Jun;56(6):466–74.

6. Leader G, Hogan A, Chen JL, Maher L, Naughton K, O'Rourke N, et al. Age of Autism Spectrum Disorder Diagnosis and Comorbidity in Children and Adolescents with Autism Spectrum Disorder. Dev Neurorehabilitation. 2020 May 14;25(1):29–37.

7. Abbas H, Garberson F, Liu-Mayo S, Glover E, Wall DP. Multi-modular AI Approach to Streamline Autism Diagnosis in Young Children. Sci Rep. 2020 Mar 19;10(1):5014.

8. Elder J, Kreider C, Brasher S, Ansell M. Clinical impact of early diagnosis of autism on the prognosis and parent-child relationships. Psychol Res Behav Manag. 2017 Aug 24;Volume 10:283–92.

9. Volkmar FR. Editorial: The Importance of Early Intervention. J Autism Dev Disord. 2014 Oct 18;44(12):2979–80.

10. Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, et al. Randomized, Controlled Trial of an Intervention for Toddlers With Autism: The Early Start Denver Model. Pediatrics. 2010 Jan 1;125(1):e17–23.

11. Wang Y, Zhao Y, Therneau TM, Atkinson EJ, Tafti AP, Zhang N, et al. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. J Biomed Inform. 2020 Feb;102:103364.

12. Küpper C, Stroth S, Wolff N, Hauck F, Kliewer N, Schad-Hansjosten T, et al. Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. Sci Rep. 2020 Mar 18;10(1):4805.

13. Research Database [Internet]. Cambridgeshire and Peterborough NHS Foundation Trust; Unknown. Available from: https://www.cpft.nhs.uk/research-database/

14. Jay MA, Mc Grath-Lone L, Gilbert R. Data Resource: the National Pupil Database (NPD). Int J Popul Data Sci [Internet]. 2019 Mar 20 [cited 2023 Feb 14];4(1). Available from: https://ijpds.org/article/view/1101

15. World Health Organisation. International Statistical Classification of Diseases and Related Health Problems (10th Revision) [Internet]. 2019. Available from: https://icd.who.int/browse10/2019/en#/F84

16. Rosenberg RE, Kaufmann WE, Law JK, Law PA. Parent Report of Community Psychiatric Comorbid Diagnoses in Autism Spectrum Disorders. Autism Res Treat. 2011 Aug 18;2011:1–10.

17. Das-Munshi J, Schofield P, Bhavsar V, Chang CK, Dewey ME, Morgan C, et al. Ethnic density and other neighbourhood associations for mortality in severe mental illness: a retrospective cohort study with multi-level analysis from an urbanised and ethnically diverse location in the UK. Lancet Psychiatry. 2019 May 13;6(6):506–17.

18. Roman-Urrestarazu A, Yang JC, van Kessel R, Warrier V, Dumas G, Jongsma H, et al. Autism incidence and spatial analysis in more than 7 million pupils in English schools: a retrospective, longitudinal, school registry study. Lancet Child Adolesc Health. 2022 Dec 1;6(12):857–68.

19. Griffiths E, Greci C, Kotrotsios Y, Parker S, Scott J, Welpton R, et al. Handbook on Statistical Disclosure Control for Outputs [Internet]. UK Data Service; 2019. Available from: https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf

Word count: 2,195