Centre for excellence
for Children's Care and Protection

CHILDREN'S SERVICES REFORM RESEARCH:

MAPPING INTEGRATION AND OUTCOMES ACROSS SCOTLAND: A STATISTICAL ANALYSIS

TECHNICAL REPORT

August 2023

Authors: Dr Joanna Soraghan, Professor Adrian Bowman (University of Glasgow), Micky Anderson, Dr Alex McTier

Table of contents

CELCIS, the Centre for Excellence for Children's Care and Protection, is a leading improvement and innovation centre in Scotland. We improve children's lives by supporting people and organisations to drive long lasting change in the services they need, and the practices used by people responsible for their care.

How to read this report

This report is a supplementary document to the main 'Mapping Integration & Outcomes Across Scotland: A Statistical Analysis' report (Anderson et al., 2023) and should be read alongside that document. It is primarily intended for statisticians and researchers who wish to gain a better understanding of the methodology that was applied for this strand of the Children's Services Reform Research study.

It is important to emphasise that it is not necessary to understand all aspects of the methods applied in this research to be able to understand what the key learning has been from this piece of work. Our key findings are provided within the main report alongside a more accessible overview of the methodology. A shorter summary report is also available.

In addition to this Technical Report, an example dataset has been provided for one indicator along with a copy of the code that was utilised to conduct the analysis. We hope that together these will provide full clarity and transparency to people who may be interested in the specifics of the statistical modelling approach we have taken for this work. All data analysis was undertaken within the R Statistical Software environment (version 4.2.2)(R Core Team 2022).

We have sought to provide details on specific aspects of the analyses that were undertaken, beginning with an explanation of multilevel models and why these were selected for use in this study.

Note: For clarity, figures presented within this Technical Report have been labelled with letters (such as 'Figure A'). Any reference to Tables and Figures that are labelled with numbers (such as 'Figure 1') refer to those in the main 'Mapping Integration & Outcomes Across Scotland: A Statistical Analysis' report.

Readers may also find it helpful to refer to the language section of the 'Mapping Integration & Outcomes Across Scotland: A Statistical Analysis' report.

Modelling the data

Multilevel models

When looking at performance for a given indicator across all of the 32 local authority areas of Scotland, it would be expected that indicator measurements from within a particular local authority area would be more similar than measurements from across different authorities. As such, these measurements are related and would not be classed as 'independent' – a property that is required by many conventional statistical approaches. This correlation (or relationship) between the measurements of a particular local authority needs to be considered when selecting a modelling approach. Multilevel models (also known as mixed effects models) are the ideal statistical tool for this scenario. The technique is commonly used with longitudinal data (that is, data captured on the same subjects at different points in time). This type of model allows us to estimate the overall trends present while accounting for variation across the different subjects (with 'subjects' in the context of this work meaning the 32 local authority areas).

Multilevel models were selected for use in this study for three key reasons:

1. They account for the fact that there will be a correlation between the measurements of a given local authority, a fact that makes simple regression models invalid with this type of data.
2. These models allow for the inclusion of local authority areas with partial missing data (that is, no data available at certain timepoints).
3. They are suitable for use with a lower number of time points than traditional time series approaches (as few as 3, as outlined in Curran et al. (2010)). This is particularly important as our time series are limited to approximately 5-10 annual time points.

An overview of multilevel models is given by Brown (2021).

Modelling distributions

Multilevel models are flexible in that they can model a variety of different types of response variable (or indicator), such as continuous, normally distributed data, binary responses (where the outcome is a percentage) or count data (non-negative whole numbers). This is done through specification of the appropriate distribution needed to model the data within the multilevel modelling framework. An introduction to statistical distributions is provided by Viti et al., 2015.

The indicators for our study fall into two categories. The first is count data (such as the number of child protection registrations). Indicators of this type were modelled using a Negative Binomial distribution, which was universally found to be more appropriate than a Poisson distribution due to high levels of variability within the data. For indicators of the count type, the population of the local authority area was used as an offset within the model to convert the counts to rates (for example, number per 10,000), thereby standardising for diverse sizes of local authorities.

The second type of indicators are percentage data (for example, percentage of looked after children with three or more placements in the last 12 months), for which the binomial distribution was specified. Due to a poor fit of the binomial model in two instances, these particular percentage indicators were instead modelled using a Gaussian (Normal) multilevel model, with the model weighted for the population size that the percentage was calculated from (that is, the denominator).

The data type of each indicator is shown in Table 2 of our main report (within the column labelled 'Measure'), while the specific distribution used to model each indicator can be seen in Table 4 of the main report.

Specification of fixed and random effects

Multilevel models are also known as 'mixed effects models' as they comprise both 'fixed effects' and 'random effects'. Generally speaking, fixed effects are variables that are believed to have an impact on our response variable (that is, our 'indicator') and it is hoped to estimate or quantify this effect through the modelling. Examples of fixed effects in the current context would be the level of structural integration present or the level of deprivation within the local authority areas. Random effects can be thought of as grouping variables, which indicate which measurements/data points have come from the same subjects and should be considered as being related. It is generally of less interest to quantify or estimate the effect of these grouping variables, but it is important to control for them to ensure accuracy of the results. Within the current context, this would be a variable indicating which local authority area the measurement pertains to.

The random effects within a model can take on the form of either 'random intercepts' or 'random slopes'. Random intercepts allow the outcome to be higher or lower for each local authority area through a simple vertical shift, whereas random slopes would allow for a different trend over time for each local authority area.

All multilevel models used in this study were fitted with random intercepts only, meaning that the models allowed each local authority area to have a different intercept (or positive/negative deviation from the average of all local authority areas), but did not attempt to estimate a unique trend or slope over time for each local authority area.

Similarly, when looking to model the 'fixed effects' for changes at the different integration levels (Full, Partial or No structural integration), changes to the intercept (that is, positive/negative deviations from the average) were estimated for each group while changes in the trends (or slopes) for each group were not.

The rationale for this approach was largely down to sample size. The move to a model with varying slopes for each level of integration plus the inclusion of random slopes for each local authority area would require the estimation of many more parameters. It was found that the number of datapoints available was not always sufficient to allow for the fitting of this more complex model and, as such, the decision was taken to investigate differences between the groups purely in terms of scale (or intercepts) rather than trend (or slope) in order to ensure consistency of the approach across all indicators. An additional benefit is that this allows for a more straightforward (scalar) interpretation of any differences discovered.

In a similar vein, the decision was taken to solely consider linear trends over time when modelling the data. While more complex non-linear trends are evident for some indicators in the national average plots seen in Table 4 of the main report, the relatively small number of timepoints available meant that it was not viable to investigate whether more complex patterns of change could be taking place (for example, polynomial or exponential trends).

Models to assess the impact of structural integration

Providing an overview of the specific models fitted to assess the impacts of structural integration requires us to outline two distinct approaches. Firstly, we describe the models used for indicators where data was available both before and after integration, and then describe the models used when only post-integration data was available.

For all models in Figures A-C, we have plotted pre- and post-integration trends to and from 2016 as this was the year in which most of the local authorities (22 of 32) formed their Health and Social Care Partnerships. Trends were calculated, however, using each local authority's specific year of integration. Information on the dates of formation for each of the Health and Social Care Partnerships, and several other characteristics of Scotland's 32 local authorities, can be found in Appendix 1 of the main report.

Indicators with data available before and after integration

To investigate the hypotheses provided within the Methodology section of the main report, the following models were fitted to each indicator for which data was available both prior to and after integration (n=20).

Model A0: No group effect

We acknowledge that changes may occur in our indicators over time for a variety of reasons. As such, it was necessary for us to account for this within our modelling approach. As a first step in determining the effects of integration, the model for trend analysis (Figure 3 of the main report) was extended to a model that allowed the trends to vary in the pre-integration and post-integration periods.

Figure A illustrates the estimated pre- and post-integration trends arising from this model, specifically for the indicator representing the number of children referred to the Children's Reporter on offence grounds per 10,000 children. The black lines indicate the overall trend that was estimated for each period through the multilevel model, while the underlying grey lines depict the raw data for each of the 32 local authorities.

As the line of best fit is calculated separately for the periods before and after integration, the lines will not necessarily meet at the point of integration (that is, at the grey dashed line at 2016).

For the indicator illustrated in Figure A, the trend was found to decrease pre-integration and remain relatively stable post-integration, however for a given indicator the trends were free to vary in any direction. As a contextual factor was included in the model indicating which data points were recorded after the onset of the COVID-19 pandemic, the estimated effect of the pandemic on this indicator can be seen in the downward trend between 2020 and 2021.

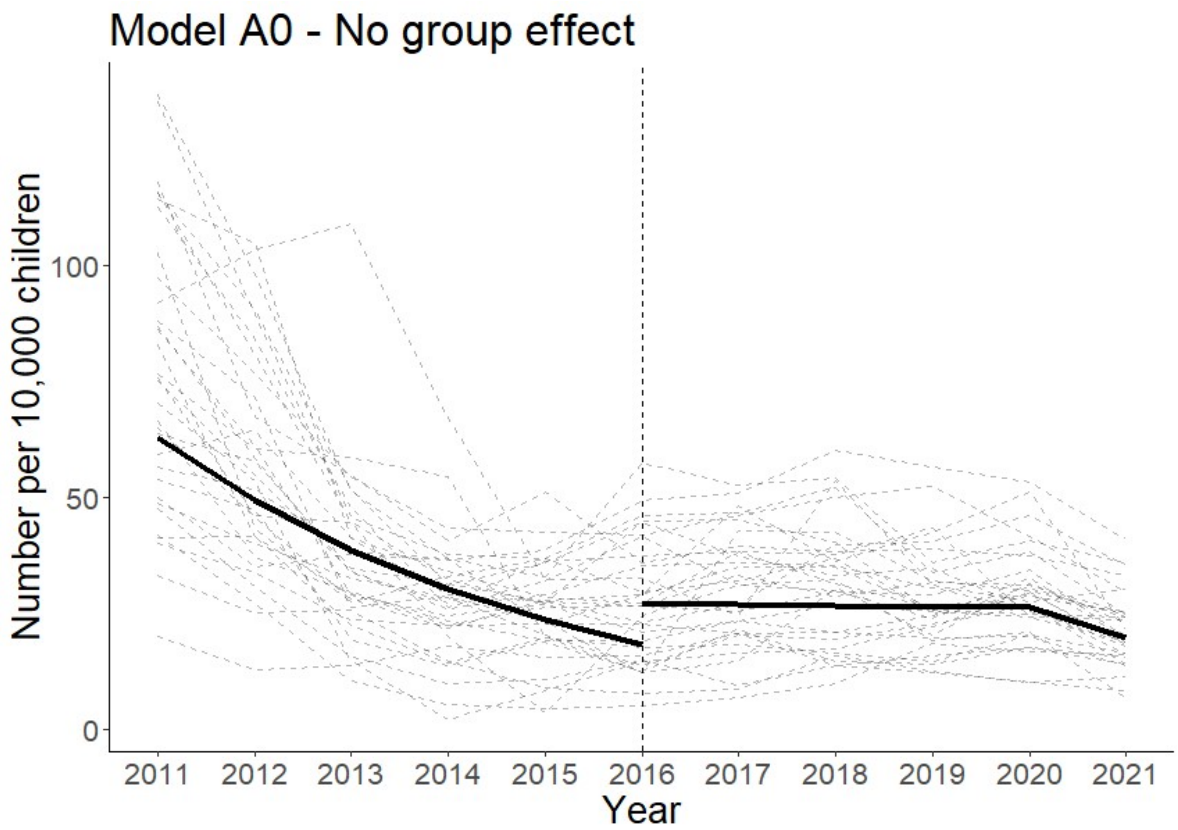

Figure A. Graphical depiction of model A0, where the trend is calculated separately in the periods before and after integration.

While Model A0 begins to provide a more detailed picture of changes in the indicator, it still does not take into account the level of structural integration present within each of the local authorities. As such, this model was not directly utilised to answer our research question, and has therefore not been included in the analysis flowchart seen in Figure 4 of the main report.

Model A1: Group effect constant over time

For any given indicator it is possible that there were pre-existing differences between the three categories of local authorities that went on to have Full, Partial or No structural integration of children's services. As such, Model A0 above was extended to a model allowing for differences between the local authorities within each of the three categories, as is depicted in Figure B. Within this new model, Model A1, differences could exist between the three categories of local authorities but were constant over time. That is, differences under this model were the same in the period before and after integration. This model therefore symbolises a situation where categories can have inherent differences but the level of integration present has not had an effect on the outcome (or 'indicator') of interest.

Note that under Model A1, the distance between any two categories is the same across both the pre- and post-integration periods. For this indicator representing the number of children referred to the Children's Reporter on offence grounds per 10,000 children, there was found to be very little difference between the three categories through this model, and as such a dashed line has been used for the Full structural integration trend to ensure that the three categories can be distinguished.

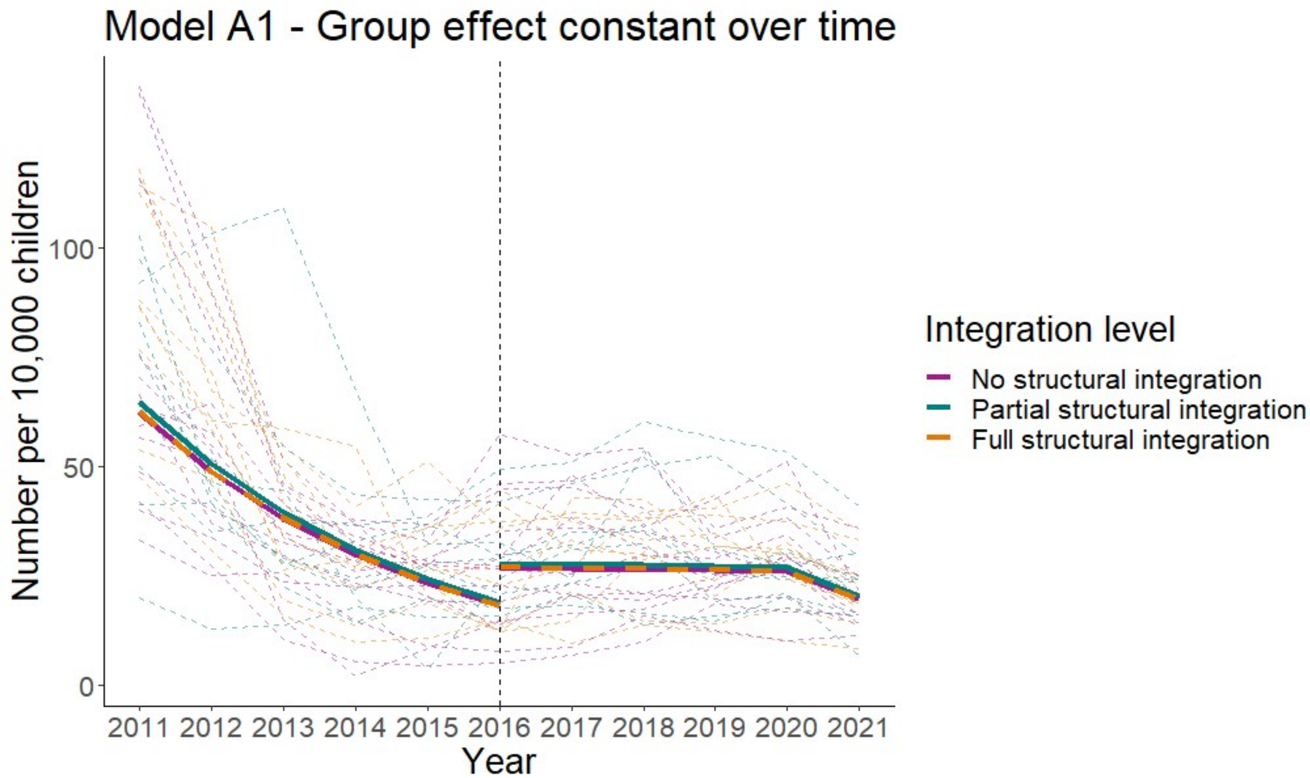

Figure B. Graphical depiction of model A1, which estimates the difference between local authorities in the three categories of integration over the time period as a whole.

Model A2: Group effect different pre-/post-integration

While Model A1 estimated the difference between the three categories over the time period as a whole, it did not allow for any changes in the differences between the three categories of local authorities after integration. In order to detect whether the level of structural integration present has had an impact on the indicator, Model A1 was then extended to allow for the differences between categories to vary before and after integration. This model is depicted in Figure C. In this particular example, it can be seen that there was very little difference between the No structural integration and Partial structural integration categories prior to integration, however these two categories have become slightly more distinct post-integration. It can also be seen that, prior to integration, local authority areas that went on to have Full structural integration had slightly higher values for this indicator (representing the number of children referred to the Children's Reporter on offence grounds per 10,000 children) on average than those with Partial or No structural integration. This pattern was then reversed in the period after integration, although all differences were small in scale.

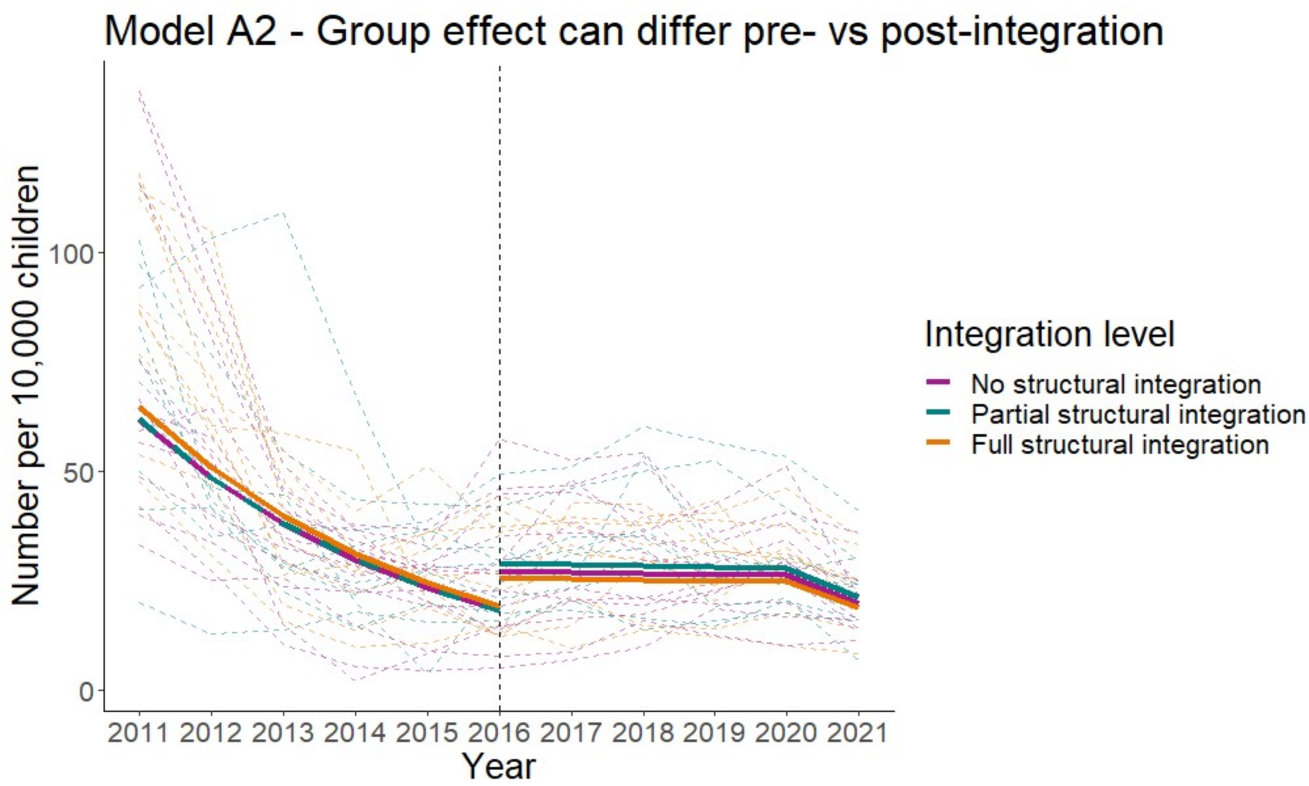

Figure C. Graphical depiction of model A2, which estimates the difference between local authorities in the three categories of integration in the pre-integration period and post-integration period separately.

What these two models tell us

These two models – Model A1 and Model A2 – are key in answering our research question. Any differences that exist between these two models – Model A1 and Model A2 – indicate the degree to which there has been a change in the relationship between the three categories after integration. That is, they show the degree to which the level of structural integration of children's services appears to be impacting upon the indicator.

As such, in order to determine if there had been a significant change across the three categories as a result of integration, a model comparison procedure was applied to compare Model A1 to Model A2. The specific procedure utilised to compare the models is called a 'parametric bootstrap model comparison' (see Model comparisons). This comparison procedure resulted in a *p*-value for each indicator, with a value of $p < 0.05$ indicating that the more complex model (Model A2) was a significantly better fit to the data[1]. Where that was the case, we determined that there was strong evidence for the differences between the three groups having changed from the pre- to post-integration periods. That is, the level of structural integration present was deemed to have had an impact on the indicator. In the example in Figures B and C, the difference between Model

[1] As we are assessing multiple outcomes, all *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method. Further details are provided later within Adjusting for multiple comparisons.

A1 and Model A2 was not found to be significant (as can be seen in Table 5 within the main report).

Where it was found that significant changes *had* taken place after integration, the specific differences between the three groups were explored in more detail through further exploration of Model A2.

Indicators with data only available after integration

Where data was not available prior to integration, the modelling approach was simplified slightly. We now required models to determine only whether there were significant differences between the three categories in the post-integration period, without any comparison to pre-integration differences.

Model B1: No group effect

As with Model A0, Model B1 does not account for the level of structural integration present within a local authority. The model purely calculates the optimal description of the trend across all local authorities in the post-integration period. The indicator represented in Figure D is 'the number of children and young people aged 12 to 20 proceeded against' (that is, who have had a criminal case brought against them) per 10,000 children and young people.

As with the previous example in Figure A, there was a clear indication of the COVID-19 pandemic impacting upon this indicator - illustrated by the more steeply downward trend between 2020 and 2021.

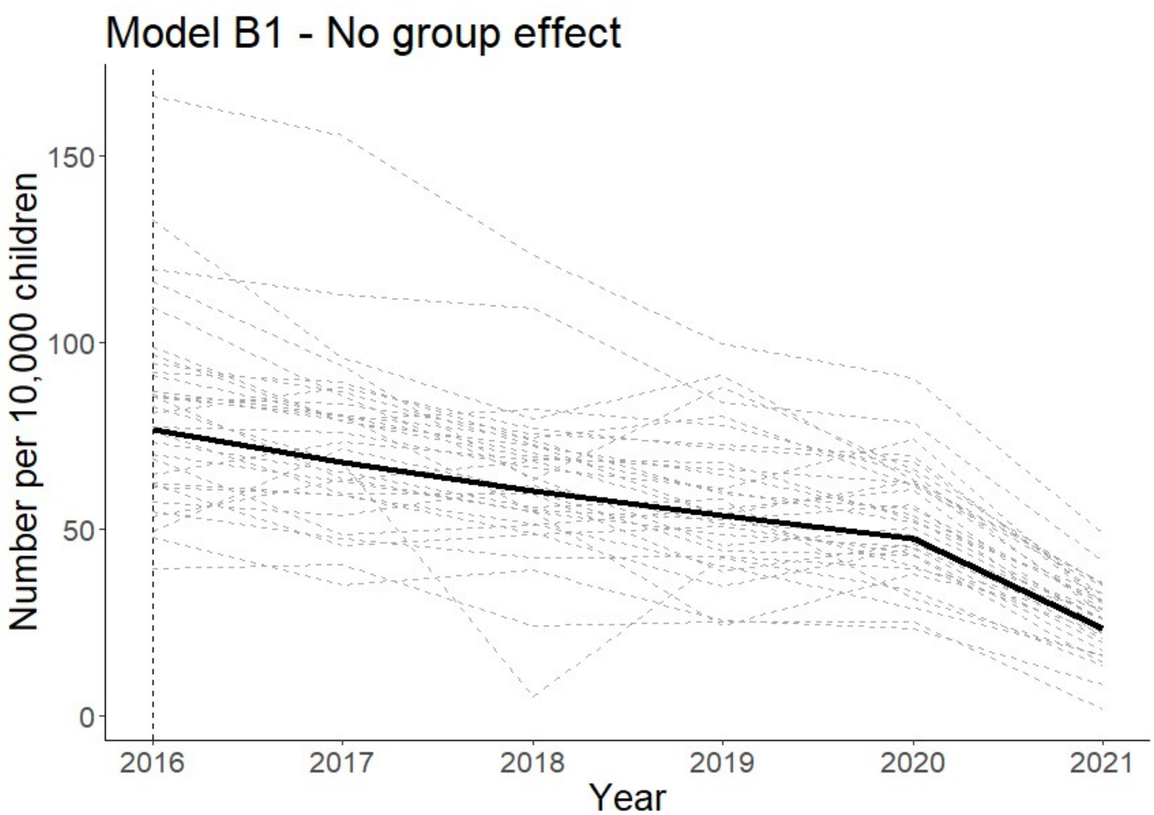

Figure D. Graphical depiction of model B1, which estimates the overall trend in the post-integration period.

Model B2: Group effect present

Model B1 was then extended to a model that accounts for the level of structural integration in each local authority. A depiction of Model B2 can be seen in Figure E and portrays the estimated trend for each of the three categories of local authorities (that is, those with Full, Partial and No structural integration) in the period after integration.

In order to determine if the differences between these three categories were significant, we again compared Model B1 to Model B2 by means of 'parametric bootstrap model comparisons'. Where Model B2 was a significantly better fit to the data than Model B1 (indicated by a *p*-value < 0.05 from the model comparison procedure[2]), we could then determine that there was strong evidence of a difference between the three categories of local authorities in terms of the given indicator.

As can be seen in Table 6 of the main report, the difference between the three categories of local authority areas was not found to be statistically significant for the particular indicator shown in Figure E (the number of children and young people aged 12 to 20 proceeded against per 10,000 children and young people).

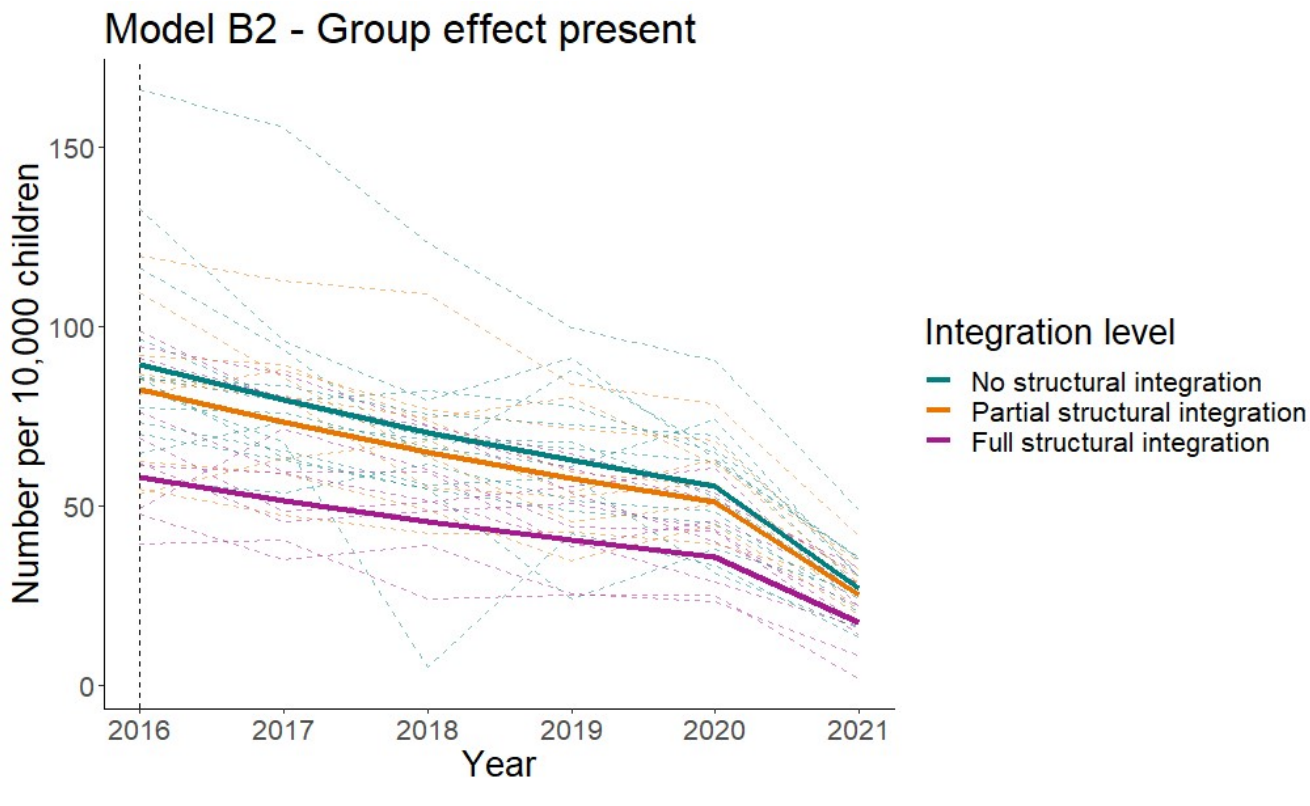

Figure E. Graphical depiction of model B2, which estimates the trend in the post-integration period for each distinct level of integration.

[2] As we are assessing multiple outcomes, all *p*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method. Further details are provided later within [Adjusting for multiple comparisons](#).

Again, where evidence was found that there *was* a difference between the three categories of integration, the specific differences could be explored through further examination of Model B2. In the absence of information prior to integration, however, there is difficulty in explicitly determining that any existing differences are a result of integration or that they were not already present prior to the Health and Social Care Partnerships being formed.

Model comparisons

There is ongoing discussion within the statistical literature on the most accurate way in which to determine *p*-values within a multilevel modelling framework. With traditional approaches to significance testing such as Wald tests, complications arise due to uncertainty in the determination of degrees of freedom that should be utilised for the distribution of the test statistic. As such, a variety of more precise alternatives have been adopted, an overview of which can be found in Luke (2017).

Parametric bootstrapping is widely considered to be a robust approach as it makes no assumptions about the distribution of the test statistic, and is suggested (Bates et al., 2015; Bolker, 2023) as a preferred approach by the authors of the prominent lme4 package for multilevel modelling within the R Statistical Software environment. This approach has been found to provide more accurate estimates than a conventional Likelihood Ratio Test for model comparisons with multilevel models, with a simulation study by Luke (2017) finding that parametric bootstrapping could produce acceptable error rates with all sample sizes. The parametric bootstrap approach to model comparison was additionally selected as it is appropriate for use with indicator variables that are not ‘normally’ distributed, as is the case for the indicators of interest in this study.

The R Statistical Software environment was utilised for all analysis within this strand of the research study. The lme4 (Bates et al., 2015) package was used to fit all models, while the pbkrtest (Halekoh & Højsgaard, 2014) package was utilised to conduct all model comparisons via a parametric bootstrapping approach, by means of the ‘PBmodcomp’ function.

Adjusting for multiple comparisons

As illustrated in the flowchart in Figure 4 of the main report, all *p*-values arising from the previously described model comparisons were adjusted for ‘multiple comparisons’. We have sought to explain here what we mean by multiple comparisons and provide more detail on the approach taken for this study.

When testing for significance at the 5% level (that is, using $p = 0.05$ as the threshold for significance), there is a 5% chance that we will incorrectly reject the null hypothesis when it is in fact true. That is to say, there is a 5% chance that we will determine to have found a real effect of our variable of interest (that is, integration) where in reality there is no effect present.

When we conduct more than one test of a hypothesis (or a set of closely related hypotheses), however, this can become problematic – it becomes more likely that some finding will appear ‘significant’ even when there is no underlying effect taking place. For example, in the case where we conduct 100 hypothesis tests where the null hypothesis is true and there is in fact no effect present, we would expect to incorrectly find evidence of an effect in 5 of these tests. As such, it is prudent to adjust for the number of tests conducted in order to have confidence in the resulting findings. This phenomena is referred to as the multiple comparisons problem, and an overview of the problem and approaches to dealing with multiple comparisons is given by Chen et al. (2017).

Given that we wanted to conduct a hypothesis test on each of our 25 indicator variables, we can consider these as 25 simultaneous hypotheses. Even where there has been no effect of the level of integration on the indicators, we would therefore expect to find incorrect significant results in approximately 1-2 (or 5%) of these tests if we did not account for the fact that multiple hypotheses have been tested. As such, an adjustment for multiple comparisons was applied to the *p*-values resulting from our model comparisons in order to prevent an increase in the Type I error rate (that is, an increase in the chance of finding an effect when there is no effect present). The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was employed to correct for multiple comparisons.

Results were then confirmed through application of the more stringent Bonferroni-Holm adjustment which led to the same interpretation, that is: the same three indicators were found to be statistically significant through this alternative approach to adjusting for multiple comparisons.

Further exploration of statistically significant results

There were three instances where we found via our model comparisons that the level of structural integration was in fact significantly associated with a given indicator. For each of these three indicators, the detail of this was explored further by exploration of the estimated marginal means of the three categories of local authority areas, as calculated by the *emmeans* package (Lenth 2023) in the R Statistical Software Environment. These means provide the average indicator value for each of the categories of local authority areas (that is, those with Full, Partial and No structural integration) in both the pre- and post-integration period, when all other variables in the model are held constant. Model variables that are continuous (that is, population density and deprivation) are set to their mean value, while for categorical variables (that is, whether the local authority and health board are coterminous, and whether or not the data point was measured during the COVID-19 pandemic or not) the estimated marginal means are calculated by averaging over the different categories (or levels) of that variable.

Dealing with data suppression

Data suppression is a method employed by many public bodies and other organisations who manage and publish data. It refers to the process of removing or masking certain data points within a dataset or table and is largely done to protect the anonymity and privacy of those that the data represents. This is generally done where small numbers are involved (with 'small' generally meaning numbers lower than 5 or 10).

While indicators with very high levels of data suppression were excluded from this study, several of the remaining indicators did contain a number of data points that had been suppressed. For the purposes of this research, the approach we took to deal with suppressed data was:

- **For indicators comprising count data (that is, those measured as 'Rate per…' in Table 2 of the main report)**, as suppressed values were known to be smaller than a particular number (usually five), twenty imputed datasets were created, within which each suppressed data point randomly took on one of the possible values that the suppression could be masking (specifically 1, 2, 3 or 4). Models were then fitted to each of these datasets, with the results of these models then being pooled.

- **For indicators comprising percentage data (that is, those measured as '%' in Table 2 of the main report)**, as both the numerator and denominator were suppressed where either of those values took on a value lower than five, the percentage could not be calculated and could plausibly take on any value from 1% (for example, one child out of 100) to 100% (for example, four children out of four). As such, these suppressed data points were treated as missing data – something that multilevel models are equipped to deal with.

Assessing correlation of contextual factors

Multicollinearity is the term used to describe high levels of correlation between two or more explanatory variables (or 'contextual factors') within a statistical model. Multicollinearity can be problematic as it makes it difficult to determine the individual effect of each of the correlated explanatory variables on the outcome variable of interest. To ensure that there were no problematic levels of correlation between any of our explanatory variables, the Variance Inflation Factor (VIF) was calculated for all explanatory variables in each model. All VIF values were found to be below the widely accepted threshold of 5 (with the vast majority being below 3), indicating that the levels of correlation between our explanatory variables were not providing cause for concern.

The correlation between deprivation and population density over the period 2011-2021 was 0.54, representing a moderate strength of relationship between the two variables. The fact that the correlation coefficient is positive indicates that, on the whole, areas that are more densely populated tend to have higher levels of deprivation than those that are less densely populated. The relationship between these two variables is displayed in Figure F. The measure of deprivation was based on the Scottish Index for Multiple Deprivation (SIMD) which is updated every four years (Scottish Government, 2020). As such, adjacent dots can be seen in the plot where the SIMD value has remained the same but there have been changes in the population density of an area.

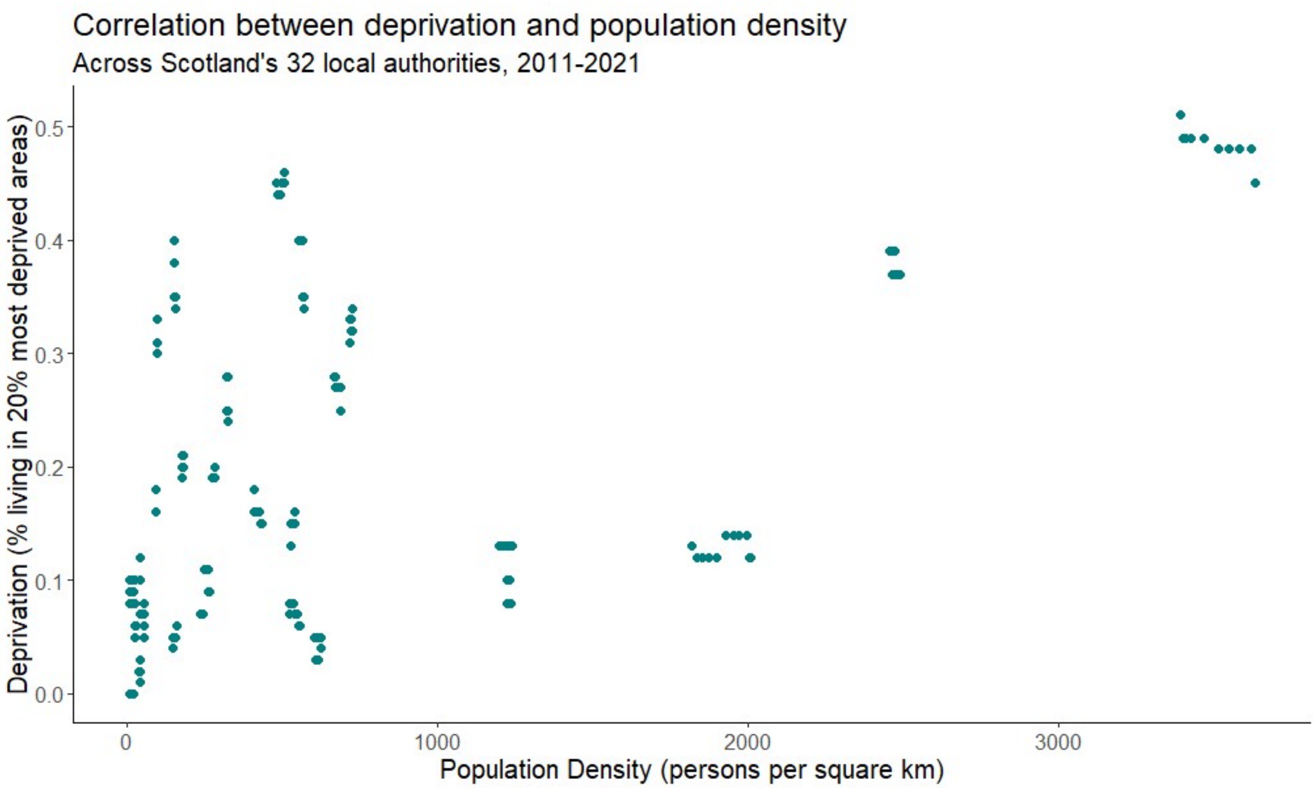

Figure F. Plot of population density against deprivation (taken to be the percentage of individuals within a local authority area who are living in the 20% most deprived areas in Scotland).

Summary

There are ongoing and varied changes taking place in children's outcomes, with a variety of factors contributing to these changes. Our report 'Mapping Integration & Outcomes Across Scotland: A Statistical Analysis' report (Anderson et al., 2023) details this more fully. A relatively complex approach was therefore required in order to analyse these changes appropriately and robustly, and in particular to identify any changes that could be attributed to the structural integration of children's services. This report has sought to provide additional detail that, alongside the main report, gives a comprehensive overview of the approach taken for this research study.

We hope that the detail provided in this Technical Report, combined with the data and code that have been published alongside it, will provide full clarity and transparency on the methodology implemented for this piece of research for the Children's Services Reform Research study.

References

Bates D, Mächler M, Bolker B, Walker S (2015). 'Fitting Linear Mixed-Effects Models Using lme4.' *Journal of Statistical Software*, 67(1), 1–48.

Benjamini, Y, and Hochberg, Y (1995). 'Controlling the false discovery rate: a practical and powerful approach to multiple testing'. *Journal of the Royal Statistical Society Series B*, 57, 289--300.

Bolker, B. (2023) GLMM FAQ. Available at: http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#inference-and-confidence-intervals

Chen SY, Feng Z, Yi X (2017). 'A general introduction to adjustment for multiple comparisons.' *J Thorac Dis*. 2017 Jun;9(6):1725-1729.

Halekoh U, Højsgaard S (2014). 'A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest.' *Journal of Statistical Software*, 59(9), 1–30.

Lenth R (2023). emmeans: Estimated Marginal Means, aka Least-Squares Means. R pack age version 1.8.4-1, Available at: https://CRAN.R-project.org/package=emmeans

Luke, S.G. 'Evaluating significance in linear mixed-effects models in R'. *Behav Res* 49, 1494–1502 (2017).

R Core Team (2022) *R: A Language and Environment for Statistical Computing.* [online] Vienna, Austria: R Foundation for Statistical Computing. [Software] Available at: https://www.r-project.org/

The Scottish Government (2020) *Scottish Index of Multiple Deprivation.* [Dataset] Available at: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/

Viti, A., Terzi, A., and Bertolaccini, L. (2015) 'A Practical Overview on Probability Distributions'. *Journal of Thoracic Disease* 7 (3), E7–E10.

About CELCIS

CELCIS, the Centre for Excellence for Children's Care and Protection, is a leading improvement and innovation centre in Scotland. We improve children's lives by supporting people and organisations to drive long-lasting change in the services they need, and the practices used by people responsible for their care.

For more information

Visit: www.celcis.org Email: celcis@strath.ac.uk **Tel**: 0141 444 8500