The Challenges of Assessing Sample Representativeness within Community-Based Evaluations of Parenting Programmes delivered in England and Ireland

Nicole Gridley[1], Grainne Hickey[2], Tracey Bywater[3], Sinead McGilloway[2] Sarah Blower[3], Karen Whittaker[4] & Vashti Berry[5]

[1]Carnegie School of Education, Leeds Beckett University, Leeds, LS6 3QW. ORCID ID: 0000-0002-1881-5502

[2] Centre for Mental Health and Community Research, Department of Psychology, Maynooth University, Maynooth, Ireland.

GH: 0000-0002-3192-7859

SM: 0000-0003-0988-3201

[3] Department of Health Sciences, University of York, York, YO10 5DD

TB: 0000-0001-7207-8753

SB: 0000-0002-9168-9995

[4]School of Nursing, University of Central Lancashire, Preston, PR1 2HE

KW: 0000-0002-3493-9396

[5]University of Exeter Medical School, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

VB: 0000-0001-6438-3731

Authors Contribution

All authors contributed to the design of the presented study. TB and SMG were PI's for the two respective trials from which the data were drawn. NG and GH collated the trial and national data for the respective jurisdictions. NG, GH and SB carried out the analysis advised by TB, SMG and VB. NG and GH led the writing of the manuscript with input from all authors.

Data Availability Statement

The data that support the findings of this study are available from the authors upon reasonable request.

Acknowledgements

The time of TB, SB and VB was supported by the National Institute for Health Research Yorkshire and Humber and South West Peninsula Applied Research Collaborations. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care. This work was also part-funded by the NIHR Public Health Research Program (PHR) as part of the study 'Enhancing Social-Emotional Health and Wellbeing in the Early Years: A Community-based Randomized Controlled Trial (and Economic) Evaluation of the Incredible Years Infant and Toddler (0-2) Parenting Programs (E-SEE trial, ref: 13/93/10), which is funded by NIHR from 2015 to 2020. This project was part funded by the University of York's Internal Impact Pump Priming funding. Lastly, the ENRICH research programme, from which some of the data used in this study were generated, was funded by the Health Research Board in Ireland under its 'Collaborative Applied Research Grants scheme in Population Health and Health Services Research 2012' which was awarded to Professor Sinéad McGilloway as the Principal Investigator (CARG/2012/17).

We would like to thank Shane Leavy for his initial input into the collation of the data for the ENRICH trial comparisons.

Conflict of Interest

The authors have no conflicts of interest to declare.

Word count:

Abstract 289/300

Main Article (minus abstract, references and tables) 4911 / 5000

Abstract

The findings from evaluations of parenting programmes can help inform policy and practice decisions, including how best to allocate scarce resources designed to support families. However, studies often fail to consider the extent to which the findings can be generalised to other settings or populations. One, yet unexplored solution, is to compare study findings and large-scale datasets including publicly available population data. The aims of this study were to assess the feasibility of assessing sample representativeness using publicly available data and to identify the challenges involved in considering the generalisability of study findings. Sociodemographic data from two community-based evaluations of parenting programmes conducted in England and Ireland between 2015-2018 were used in the study (N =395 parent-infant dyads). The results indicated some differences between the trial samples and the wider population. However, it is difficult to reach definitive conclusions about these findings due to the limitations associated with using the comparative datasets. Our study revealed three key challenges, including: (1) how best to define and conceptualise representativeness; (2) the availability of comparative datasets; and (3) the quality of the available data. Our study suggests that there is a need for up-to-date, good quality comparative datasets to allow for the assessment of representativeness. Further work is required to identify parameters for making claims about representativeness, specifically regarding the acceptable level of difference between the target population and the study sample. This is the first study to explore the feasibility of using publicly available population data in two jurisdictions, for the purpose of making judgements about the representativeness of the findings from parent programmes. It is hoped that our results will encourage further investigation around the reporting of trial external validity to enable effective decision making at policy and practice level.

Keywords: representativeness; parenting; trials; external validity; generalisability; England; Ireland

What is known about this topic?

- Findings from evaluations of interventions are often used to inform policy-based decisions.
- Sample representativeness guides decisions regarding how well findings will translate to other, or wider populations from which the sample has been drawn.
- The reporting of sample representativeness has seen a decline since reporting guidelines have been introduced.

What this paper adds?

- This article is the first to use population level publicly available data to assess sample representativeness for community-based parenting programme evaluations in two countries.
- Three key challenges or ‘lessons’ associated with accessing and using appropriate population level data are identified.
- Six key recommendations have been made to help researchers, policy makers and practitioners in overcoming these challenges into the future.

Introduction

Decision makers in the field of early intervention and prevention increasingly work with tight financial budgets and therefore have a growing need for evidence which can inform the development of best practice in early child and family service provision. Evaluations of parenting programmes, particularly evidence from randomised controlled trials (RCTs) and quasi-experimental studies, are crucial in determining 'what works' (Kennedy-Martin, Curtis, Faries, Robinson & Johnston, 2015). The results from such trials are normally collated and translated into guidelines or recommendations for evidence-based standards (Atkins et al., 2004; Travers et al., 2007). Arguably however, these can only be used to inform effective decision making if the evaluations are well designed, and the participants are representative of the populations for whom the programmes were originally designed (Travers et al., 2007).

A small body of research suggests that group-based early parenting programmes may be universally beneficial when implemented as a preventative measure (e.g. Feinberg et al., 2010; Lindsay & Totsika, 2017). However, systematic reviews have recently highlighted mixed effects of parenting interventions on child outcomes and parent-child relationships when implemented in the *earliest* years (Hurt et al., 2018). High-quality trials of parenting programmes implemented in the earliest years of life (birth to three years) with universal/targeted elements are important, in that their findings are likely to be used to guide public health policy and, therefore, careful interpretation of their context is needed. External validity encompasses a broad range of factors including population, setting, and intervention implementation. Guidelines are available to ensure transparency in the reporting of trials and to allow for the independent assessment of external validity (i.e. CONSORT; Schulz, Altman & Moher, 2010; TIDieR; Hoffmann et al., 2014). Despite this, it is suggested that there has been little improvement in the reporting of external validity (Candy, Vickerstaff, Jones & King, 2018). Stuart, Ackerman, and Westreich (2018) suggest that publicly available population data should be used to establish sample representativeness. However, this has received little attention in the literature, specifically in the field of early-years' service provision. The aim of this current study was to identify how one element of external validity, sample representativeness, can be assessed and measured as part of trial reporting.

External Validity

Well conducted trials/evaluations with high internal validity enable researchers to confidently make conclusions regarding intervention effectiveness. However, it is also vital for researchers, practitioners and policy makers to understand the real-world contexts in which a given intervention will work. External validity refers to how well the results of a trial/study represent the larger population from which the sample was drawn, or can be

generalised to different populations (Eldridge, Ashby, Bennett, Wakelin & Feder, 2008; Leviton, 2017; Murad, Katabi, Benkhadra & Montori, 2018). Subsequently, trials that report high levels of external validity can confidently state that the results would be similar in other contexts with a range of populations (Rothwell, 2006; Tomlinson, Ward & Marlow, 2015). Thus, evidence of both internal and external validity is necessary to ascertain whether an intervention is effective and can work (or will not work) when implemented in alternative community and practice situations.

Eldridge and colleagues (2008) recommend researchers to view external validity as the combination of four processes which describe the extent to which: 1) a trial is able to target all individuals who would be eligible for the intervention or, the representativeness of the sample to the population from which it was drawn ('Reach'); 2) the research settings are representative of the wider population of settings ('Adoption'); 3) the intervention is implemented as intended in real-world settings ('Implementation'); and, 4) the programme is sustained over time, even after trial completion ('Maintenance'). When evaluating psychosocial interventions, such as parenting programmes, the processes of implementation and maintenance are typically assessed by means of an accompanying process evaluation. Establishing reach and adoption can be considered within the impact evaluation, and measurement and monitoring of these processes should be considered prior to the study commencing (Rothwell, 2006) or during the after-the-fact analysis of the trial data (Stuart, Bradshaw & Leaf, 2015).

Several factors associated with the study design have been suggested to influence representativeness and generalisability. For example, methods of recruiting the research site and participants (i.e. via purposive or random selection), may positively or negatively impact the generalisability of findings to different settings (Olsen, Orr, Bell & Stuart, 2013), and the very nature of an RCT design may discourage participation, either because there is no guarantee of receiving the intervention, or there may be a perceived stigma associated with being offered an intervention (Stuart et al., 2015). Finally, differences between residents across specific geographical areas, and contrasting local health and social care systems, can also have implications for generalisability (Kennedy-Martin et al., 2015). Thus, there is a need for researchers to be thorough in their reporting of these facets of a study to facilitate effective decision-making with regards to potential roll-out of the intervention to other populations and contexts.

Reporting External Validity within Trials

Calls for greater transparency in this respect have led to the development of checklists for the general reporting of intervention complexity and delivery (e.g. TIDieR; Hoffmann et al., 2014) as well as standardised guidelines

for RCTs (CONSORT; Schulz et al., 2010) and non-RCTs (TREND; Des Jarlais, Lyles & Crepaz, 2004). However, despite recent revisions in reporting guidelines, critics argue that they do not go far enough in terms of encompassing all aspects of external validity that are required for making informed decisions at a practice and policy level (Kennedy-Martin et al., 2015). For example, there is no clear guidance or stipulated parameters regarding how assessments of comparability to the wider population, or different populations, should be made.

The results from several systematic reviews have indicated that despite changes to reporting guidelines, many studies consistently fail to report on the basic elements of external validity. Eldridge et al. (2008) reported that only 53% of 34 cluster-RCTs in primary care included a discussion of sample generalisability. This was found to be more likely when the authors had also reported on the characteristics of those who were approached to participate in the study as opposed to those who were recruited and took part. This is important as often those included in the final analysis are not always the same as those who are initially recruited due to attrition. More recently, Candy et al. (2018) indicated that there has been no significant improvement since 2002, in the reporting of external validity in psychotherapeutic and educational RCTs, including early parenting programmes, and specific characteristics pertinent to understanding representativeness are now being under-reported far more often than before guideline revisions in 2008. Subsequently, there are calls for researchers to enhance their level of reporting, specifically in relation to providing a direct comparison where possible between recruited sample characteristics and the target population.

It is important to define the 'target population' with whom the findings should be compared and generalised, in order to establish sample representativeness, Furthermore, there should be access to national data on that population (Stuart et al., 2015) and suggestions for achieving this have been made by several research teams. Drawing on previous work (Rothwell, 2006), Tomlinson et al. (2015) suggest that the sociodemographic characteristics of those enrolled in a study are the primary factors influencing generalisability, and that comparisons between those who were approached to take part but who did not consent to do so, versus those recruited into the study, should be the gold standard method of drawing conclusions regarding representativeness. Unfortunately, this solution presents a rather circular problem, as it is not always possible nor ethical to obtain such data from those individuals who do not wish to take part in the trial. Another suggestion comes from Millard et al. (2014) who proposed retrospective data collection using a community survey. The challenge here is that the time lapse between the original trial and the retrospective data collection period can be affected by shifting demographic patterns which could then render the assessment of external validity obsolete. Where previous evaluations of an intervention exist, comparisons between datasets can be made in the form of pooled studies.

However, when this is not possible, researchers may be faced with the prospect of using publicly available data collected from similar cohorts or national large-scale surveys. This alternative method of assessing sample representativeness has recently been recommended by Najafzadeh and Schneeweiss (2017) and Stuart et al. (2018), although the limitations of this approach are not clear.

Current Study

The aim of the current study - which was undertaken in two neighbouring jurisdictions - was to explore the utility of publicly available population data when trying to establish one of four elements of external validity, the representativeness of samples, when recruiting participants to community-based evaluations of universal parenting programmes. Specifically, we were interested in identifying the key challenges for researchers when using publicly available data to assess sample representativeness.

Methods

Data Sources

Sociodemographic data were drawn from two independent trials of community-based early parenting programmes conducted in two countries with different healthcare systems (Bywater et al., 2018; Hickey et al., 2016; 2019). These data were compared to publicly available Government and Local Authority data to represent a probability sample of national and local parents of young infants living within the settings targeted for the intervention.

The trials

1. The **E**valuatio**N** of w**R**aparound in **I**reland for **CH**ildren and families - **P**arent and **IN**fant (ENRICH-PIN) study (e.g. Hickey et al., 2016; 2019) was designed as a quasi-experimental evaluation of a newly developed, collaborative, multi-component, wraparound-inspired, community-based universal initiative implemented in Ireland (since 2014) by Public Health Nurses (PHNs) in partnership with community-based service providers. The PIN initiative includes a range of developmentally appropriate parent and infant support programmes which are delivered to mothers of young children in a group-based format, over a period of time, from birth to two years. The trial was conducted in two separate sites in the Eastern region of Ireland, both urban, which incorporate neighbourhoods characterised by significant levels of socioeconomic disadvantage. A total of 190 parent-infant dyads (96 intervention; 84 comparison) provided written informed consent to participate in the trial and completed the baseline assessments when the target child was under four months old. Details regarding inclusion/exclusion criteria, identification and recruitment methods are presented in Table 1. Ethical approval was granted by the host

institution's Social Research Ethics Committee and by the Health Service Executive (HSE) North East Area Research Ethics Committee.

(Table 1)

2. The **E**nhancing **S**ocial-**E**motional wellbeing in the **E**arly years trial (E-SEE) comprised an external pilot (Bywater et al., in press) and a definitive RCT evaluation (Bywater et al., 2018) of the Incredible Years Infant and Toddler programmes (Webster-Stratton, 2010) delivered in a proportionate universalism approach, i.e. testing E-SEE steps (not each IY programme independently). This study utilises the pilot data pertaining to the external pilot. Intervention delivery was conducted by Health Visitors and Family Support Workers in community settings as part of a universal proportionate model of need (Marmot, 2010). The pilot was conducted in two sites - one in the North and one in the South of England. Sites were initially selected due to their high levels of socioeconomic disadvantage, adequate number of live births per month, and matched or above average levels of postnatal depression in comparison to the UK norm. A total of 205 parent-infant dyads were formally recruited to the study. Apart from one father, all participating dyads included mothers and babies under two months old. Ethical approval was granted by the North West Wales NHS ethics committee (15/WA/0178) and the host institution's Research Governance Committee.

National and local level data

Comparative national and local level population data in both jurisdictions were drawn from several sources. For the ENRICH-PIN trial, population data were retrieved from the Central Statistics Office (CSO; 2014; 2016a; 2016b) census data. For the E-SEE pilot, population data were retrieved from the Office for National Statistics (2011; 2016a; 2016b; 2016c) and the Public Health England Public Health Profile database (PHE, 2019).

Data Identification

Initial steps in the process involved identifying data/variables that were common to both trials and that could be matched with national and local data sources. Based on the recommendation by Rothwell (2006) and Tomlinson and colleagues (2015), 10 sociodemographic variables typically used to assess representativeness were identified including, for example, parent age, ethnicity, marital status, lone parent household and income. The full list (plus definitions) is provided in Table 2.

(Table 2)

We also sought to locate national and local level equivalent data for the comparison of maternal depression rates and severity. This was measured in both studies using the Primary Health Questionnaire - 9 (PHQ-9; Kroenke, Spitzer & Williams, 2001). The PHQ-9 is a brief, 9-item self-report scale which provides an overall

indication of depression based on how the respondent has been feeling during the previous two weeks. Responses on each of the items are provided using a Likert scale ranging from 0 (not at all) to 3 (all of the time) and summed to generate a total score (minimum score = 1, maximum = 27) which provides an index of overall severity of depression. By convention, total scores ranging from 5 to 27 indicate mild to severe depression. For the E-SEE trial, the PHQ-9 was used to screen for different levels of intervention following the baseline assessment, and as an outcome measure to establish change over time. For the ENRICH-PIN trial, it was used only as an outcome measure.

Analysis

Due to the composition of the comparative population datasets, data that related only to parents of young children was difficult to identify and extricate. Subsequently, overall population estimates were used for seven sociodemographic variables where subset data was unavailable, including: ethnicity, religious orientation, marital status (England only), household size, educational qualifications (England only), income; and depression scores.

Where available, the frequencies, percentages, means and standard deviations for each of the included variables were extracted from national and local level data sets. Chi square tests with Yates correction were applied to categorical variables. Analysis could not be performed at the local authority level for the variable 'first child' within the England based E-SEE trial due to lack of publicly available data. We were also unable to perform analysis for one site in the E-SEE trial for both employment and education, due to the data violating statistical assumptions. Further, we were unable to perform comparisons between the religious orientation of the general population and the Irish ENRICH sample as this data was not collected at the trial level.

One sample *t*-tests were applied to variables with continuous data. Analysis could not be performed for income and depression rates for either trial due to difficulties accessing population level standard deviations. Similarly, incomplete data at the population level (national and site) also affected comparisons with E-SEE trial household size, and local authority level data comparisons for parent age. Consequently, data are presented for all valid cases.

Findings

How Representative are the Trial Samples of the Target Population based on Eligibility Criteria?

It is recommended that researchers define their 'target population' and look to the specific sociodemographic characteristics of the sample to draw comparisons to the wider population from which they are drawn (Tomlinson et al., 2015). The ENRICH-PIN trial was intended to be universal, targeting all mothers with new-borns living predominantly in socially disadvantaged areas. Our analysis revealed that the trial comprised

significantly more first-time mothers when compared to the general population, although this was not entirely unexpected given that these mothers may be more likely to take part in an intervention of this nature (Table 3). In contrast, the E-SEE pilot focused on recruiting all parents with new-born babies, but with a view to target those considered to be most in need according to depression scores on the PHQ-9 scores (falling within the 'mild' to 'severe' range). While this health indicator suggested that the E-SEE sample included a greater number of mothers with mild to severe depression when compared to the general population, we were unable to confirm this using statistical testing (Table 4).

(Tables 3 and 4)

How Representative are the Trial Samples to the General Population Along Other Characteristics?

Trial samples also need to represent the target population in terms of other variables which are predictive of treatment effects or the outcomes and which are likely to guide 'on-the-ground' implementation (Stuart et al., 2015). For this reason, we also explored other key characteristics.

Participants recruited to the ENRICH-PIN trial were statistically more likely to be unemployed, report fewer educational qualifications, larger household sizes, and were less likely to be a lone parent when compared to the general population. Findings were mirrored in the Site 1 data whereby trial participants were also more likely to be unemployed and to have low educational attainment compared to the local population. Differences for Site 2 versus the general population were only observed for household size and unemployment status, where again the trial sample reported a larger number of household residents and were more likely to be unemployed (Table 3).

Participants recruited to the E-SEE trial represented a significantly younger population and constituted a lower proportion of the national population who identified as White British. Significant differences to population level data also emerged on several other variables. Specifically, the E-SEE sample were less likely to report a religious orientation, and those who did, were also less likely to report identifying as Christian. The E-SEE combined sample, and those recruited in Site 1 were also less likely to report having no educational qualifications. On the other hand, the sample represented a significantly higher proportion of the national and local population (for Site 1 only) who reported being in a lone parent household and unemployed. The participants recruited to Site 1 for the E-SEE trial also represented a greater proportion of the local population who identified as single.

Discussion

The current study sought to establish the feasibility of using publicly available data to assess the representativeness of samples recruited to two large effectiveness trials in England and Ireland. We compared our samples to the wider population using 10 basic sociodemographic characteristics as recommended by Rothwell

(2006) and Tomlinson et al. (2015). Our relatively crude analyses suggest some interesting differences/variations, although given the difficulties associated with isolating subgroups within the population data, we are unable to draw any conclusions about what these results may mean for the trials themselves. The barriers that we encountered as part of this process are not unique (Stuart et al., 2015). Subsequently, we have identified and discuss three key challenges associated with this approach and make six recommendations for the conduct of future research in this area, as well as for the collection and publication of data sets against which analyses of representativeness can be conducted (Table 5).

1. Conceptualising the Target Population and Representativeness

Despite previous calls, little guidance currently exists regarding what level of 'difference' between the trial sample and the target population may be considered acceptable to establish whether a sample is representative (Rothwell, 2006). Moreover, there are no clear guidelines as to which characteristics should be used to define the target population (and their order of importance), and therefore, what kinds of comparisons should be made vis-à-vis local and nationally available data. Subsequently, we recommend that future work needs to consider the key characteristics that should be compared as a matter of course for different interventions or study designs. Further efforts should also focus on defining the statistical parameters for establishing whether the difference between the trial sample and the target population is acceptable (Recommendation 1).

2. Availability of Comparative Data

According to Stuart et al (2015), establishing the secondary data source(s) on which to make comparisons, is key to making claims and judgements about representativeness. We explored various data sources and several issues were identified, namely the availability of a) up-to-date data; b) matching time points; and, c) the inclusion of matching variables of interest.

Whilst there are a number of good quality cohort data sets available to researchers in England and Ireland (e.g. Growing up in Ireland [Nixon, Swords & Murray, 2013] and the British Cohort Study [Elliott and Shepherd, 2006]), many of these are now considered out of date and thus not appropriate for describing the representativeness of emerging trial samples. Given that local populations are often subject to shifting demographics, it was not deemed appropriate to use these cohort studies as comparative sources.

When drawing conclusions about representativeness, it is important to ensure that the data being compared are matched, specifically in relation to the time in an individual's life history when they were collected. This is of importance when comparing data for periods of infancy, childhood and adolescence as these life stages are characterised by critical periods of neurological and biological development (Bundy et al., 2018), and mismatched

ages may also reflect mismatched life stage and capability. Finding a recent comparative data set that included similar time points to both trials, proved problematic. For example, the first sweep of cohort data within the Millennium Cohort Study (Connelly and Platt, 2014) occurred when the children were 9-months old. Given that both of our trials recruited parents when their child was under four months old, it was not considered appropriate to use this data source as a means of comparison.

Similarly, it was difficult to source publicly available data sets that included the same variables reported in both trials; namely, maternal depression, income, and first-time mothers. There are two possible reasons for this. Firstly, such data may not be publicly available at a local level and therefore population estimates cannot be completed. This was most evident when trying to source data from Ireland. Secondly, where data are available, there is either very little consistency in the specific measure used for this purpose, or it is not clear how these data have been collected. For example, local level data from England available from the PHE website (2019) fail to indicate which measure of depression has been used to collect prevalence rates. Given that there is evidence to suggest that some measures of depression are more sensitive and reliable than others, there is potential for the choice of measure to impact on reported prevalence rates and, in turn, representativeness (Blower et al., 2019).

Based on these challenges, we have identified two recommendations that we suggest, should be followed up in the future by trialists and researchers.

Recommendation 2: There should be greater access to, and availability of, similar trial data to enable such comparisons by means of, for example, the archiving or pooling of anonymised datasets. However, we are aware that there are currently several cohort trials ongoing across the UK which should, in due course, help in this regard (UK Data Service; https://ukdataservice.ac.uk/).

Recommendation 3: Greater efforts are needed to identify an appropriate battery of measures that can be applied across studies and within practice-based settings (Blower et al., 2019). Trialists/researchers might consider using measures (within reasonable limits and depending on their overall suitability) which have been used in larger national studies to enable appropriate cross-comparisons. Such efforts could encourage trials to gather data for which there are known national comparative data, rather than collecting data and then trying to find suitable comparative data in which to compare. However, this data may also have limitations.

3. The Quality of Comparative Data

Stuart et al. (2015) highlighted the need for population level datasets to be of sufficiently high-quality to enable researchers to characterise potential target populations. Exploration of the secondary data sources for the purposes of this study revealed several issues associated with the quality of publicly available population data, and

specifically how the data had been compiled and presented. This had implications for identifying our own subgroups, but also for how the trial data had to be grouped to draw comparisons.

One of the main challenges is that most of the sociodemographic data available in the public domain represents information pertaining to both men and women and not necessarily just parents with children. In addition, it was not always possible to filter this information by age or other sociodemographic characteristics that defined our population of interest. Subsequently, it was difficult to identify a subset of mothers within most of the population data variables, the quality of the existing data was poor, and we were unable to make valid like-with-like comparisons.

There was also a need to transform some of the trial data to improve comparability with the national and local population datasets. For example, the variables of ethnicity, religion, lone parent household, and median weekly income all needed some form of manipulation and transformation at the trial level so that they were identical in composition to the population data. Information relating to how the population data had been grouped or compiled, was often limited or difficult to source, thereby leading to a significant investment of time and effort with little reward. Collectively, our work suggests that researchers should not underestimate the task of assessing representativeness even when considering basic sociodemographic variables. Based on these observations, we make a further three recommendations that trialists, researchers and curators of public databases should strive to address:

Recommendation 4: Provide explicit information relating to the composition and categorisation of population level data.

Recommendation 5: Modify the quality of publicly available data systems so that they may be used for multiple purposes. This should include the ability to filter the data to better identify target populations, and the inclusion of some key variables which might be useful for researchers in this area (e.g. maternal depression); and

Recommendation 6: Invest in the development of a national and accessible health visiting/midwife/Public Health Nurse service dataset of new families which, unlike most study/cohort datasets, are not limited by selection bias. While it could be argued that routine data could be accessed i.e. NHS digital, the cost and process required to access this data is not justifiable for the purposes of exploring representativeness.

Strengths and Limitations

This is the first study to assess one element of external validity, sample representativeness, using national and local level population data in conjunction with data from two trials in two different jurisdictions. Recent research indicates that external validity is being under-reported (Candy et al., 2018) and current recommendations

encourage the use of publicly available data to establish sample representativeness (Stuart et al., 2018). The findings from the present study are therefore timely and provide important evidence to suggest that the current infrastructure to support such assessments needs improvement before researchers can confidently use these datasets. Our findings should stimulate further discussion with regards to how representativeness should be defined and measured. Moreover, the findings should provide the impetus for policy makers and researchers to work collaboratively to establish what improvements are needed to enhance the quality and availability of population level data.

A key limitation of this study is that the estimates regarding representativeness are affected by the quality of comparative data and should, therefore, be interpreted with caution. This is mostly due to the way these data are collected and presented. Specifically, we were unable to filter the national, publicly available datasets to identify our 'target population' i.e. mothers of young infants, across all variables. Subsequently, we cannot be confident about the extent to which the trial samples are comparable to the wider population of mothers of young infants locally and nationally who did not take part in the respective studies. This highlights the importance of comprehensive and high-quality data gathering systems as a driver of innovation within health care delivery, including – but not limited to – child and family service settings (Fixsen, Blasé, Naoom & Wallace, 2009). Thus, greater attention is needed at the policy level to ensure the systematic collection of high-quality data on service users. This would ensure more effective identification of needs and service requirements, whilst also providing a stronger evidence base for service innovations.

Implications and Conclusions

Representativeness is only one of four elements that need to be considered when assessing external validity, but it is important in allowing us to make a judgement as to whether an intervention is likely to be effective in the wider population from which it is drawn, or from populations elsewhere. This, coupled with other aspects of external validity (e.g. aspects of implementation), is an important consideration for intervention roll-out and/or making decisions to proceed with a larger, more definitive trial, as was the case with E-SEE.

Our study has shown that using publicly available population data to ascertain sample representativeness can be problematic. While we believe there is too much error built into our findings to confidently reach conclusions about the representativeness of our samples, we have been able to tease out key learning points for the challenges we faced relevant to researchers, practitioners and policy makers. These are summarised in Table 5 but include a need for: (1) defining the parameters for establishing whether the differences between the trial sample and the target population are acceptable; (2) increased or improved availability of up-to-date comparative data, both at

local and national levels; (3) consistency in the application of measures used to collected data; (4) clearer guidance on how categories within national datasets have been composed; and (5 and 6) increased or improved level, and quality, of publicly available data. Whilst all of these should be considered carefully in the context of ethical guidelines (e.g. the need to ensure confidentiality and anonymity), they are crucial in improving access to, and the availability of, appropriate data, to allow researchers and policy makers to make informed decisions regarding population health interventions.

References

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., Liberati, A., O'Connell, D., Oxman, A. D., Phillips, B.,Schünemann, H., Edejer, T. T., Vist, G. E., Williams, J. W., and The GRADE Working Group (2004). Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches The GRADE Working Group. *BMC Health Services Research, 4,* 38-45. doi:10.1186/1472-6963-4-38

Blower, S., Gridley, N., Dunn, A., Bywater, T., Hindson, Z. and Bryant, M. (2019). Psychometric Properties of Parent Outcome Measures Used in RCTs of Antenatal and Early Years Parent Programs: A Systematic Review. *Clinical Child and Family Psychology Review*, *22*, 3, 367–387. doi.org/10.1007/s10567-019-00276-2

Bundy, D. A. P., de Silva, N., Horton, S., Patton, G. C., Schultz, L., Jamison, D. T., et al. (2018). Investment in child and adolescent health and development: Key messages from *Disease Control Priorities*, 3rd edition. *The Lancet, 391*(10121), 687-699.

Bywater, T., Berry, V., Blower, S. L., Cohen, J., Gridley, N., Kiernan, K., ...and Pickett, K. (2018). Enhancing Social-Emotional Health and Wellbeing in the Early Years (E-SEE): a study protocol of a community-based randomised controlled trial with process and economic evaluations of the incredible years infant and toddler parenting programmes, delivered in a proportionate universal model. *BMJ open*, *8*(12), e026906.

Candy, B., Vickerstaff, V., Jones, L., and King, M. (2018). Description of complex interventions: analysis of changes in reporting in randomised trials since 2002. *Trials*, *19*(1), 110.

Central Statistics Office (2014). *Census data.* Retrieved from https://www.cso.ie/en/census/

Central Statistics Office (2016a). *Census data.* Retrieved from https://www.cso.ie/en/census/

Central Statistics Office (2016b). *Small area population statistics.* Retrieved from https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/

Connelly, R., and Platt, L. (2014). Cohort profile: UK millennium Cohort study (MCS). *International Journal of Epidemiology*, *43*(6), 1719-1725.

Des Jarlais, D. C., Lyles, C., and Crepaz, N. (2004). Improving the reporting quality of nonrandomised evaluations of behavioural and public health interventions. The TREND Statement. *American Journal of Public Health, 94.* 361-366.

Eldridge, S., Ashby, D., Bennett, C., Wakelin, M., and Feder, G. (2008). Internal and external validity of cluster randomised trials: systematic review of recent trials. *British Medical Journal,* doi:10.1136/bmj.39517.495764.25

Elliott, J., and Shepherd, P. (2006). Cohort profile: 1970 British birth cohort (BCS70). *International Journal of Epidemiology*, *35*(4), 836-843.

Feinberg, M. E., Jones, D. E., Kan, M. L., & Goslin, M. C. (2010). Effects of family foundations on parents and children: 3.5 years after baseline. *Journal of Family Psychology, 24,* 532-542.

Fixsen, D. L., Blasé, K. A., Naoom, S. F., and Wallace, F. (2009). Core implementation components. *Research on Social Work in Practice, 19,* 531-540. doi: 10.1177/1049731509335549

Hickey, G., McGilloway, S., Furlong, M., Leckey, Y., Bywater, T., and Donnelly, M. (2016). Understanding the implementation and effectiveness of a group-based early parenting intervention: a process evaluation protocol. *BMC Health Services Research, 16.* Doi: 10.1186/s12913-016-1737-3

Hickey, G., McGilloway, S., Leckey, Y., Furlong, M., Leavy, S., Stokes, A., O'Connor, S., Bywater, T. and Donnelly, M (2019). Mothers' wellbeing, parenting attitudes and home environment: Cumulative risk and parity in early motherhood. *Child: Care, Health and Development* 45(4):523-530. Doi: 10.1111/cch.12677.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... and Lamb, S. E. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal*, *348*, g1687.

Hurt, L., Paranjothy, S., Lucas, P.J., Watson, D., Mann, M., Griffiths, L.J., Ginja, S., …. & Lingam, R. (2018). Interventions that enhance health services for parents and infants to improve child development and social and emotional well-being in high income countries: A systematic review. *British Medical Journal Open*, 8, e014899. doi: 10.1136/bmjopen-2016-014899

Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S., and Johnston, J. (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials, 16,* 495-509. doi: 10.1186/s13063-015-1023-4

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine, 16.* 606-613.

Leviton, L. C. (2017). Generalizing about public health interventions: a mixed-methods approach to external validity. *Annual Review of Public Health, 38.* 371-391.

Lindsay, G., & Totsika, V. (2017). The effectiveness of universal parenting programmemes: The CANparent trial. *BMC Psychology, 5*, 35.

Marmot, M., Allen J., Goldblatt, P., Boyce, T., McNeish, D., and Grady, M. (2010). Fair society, healthy lives. *The Marmot Review, 14.*

Millard, J. D., Muhangi, L., Sewankambo, M., Ndibazza, J., Elliott, A. M., and Webb, E. L. (2014). Assessing the external validity of a randomized controlled trial of anthelminthics in mothers and their children in Entebbe, Uganda. *Trials*, *15*(1), 310.

Murad, M. H., Katabi, A., Benkhadra, R., and Montori, V. M. (2018). External validity, generalisability, applicability and directness: a brief primer. *BMJ evidence-based medicine*, *23*(1), 17.

Najafzadeh, M., and Schneeweiss, S. (2017). From trial to target populations—calibrating real-world data. *New England Journal of Medicine*, *376*(13), 1203-5.

Nixon, E., Swords, L., and Murray, A. (2013). *Parenting and Infant Development*. Stationery Office.

Office for National Statistics (2011). *2011 Census.* Retrieved from https://www.ons.gov.uk/census/2011census

Office for National Statistics (2016a). *Annual survey of hours and earnings.* Retrieved from https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2017provisionaland2016revisedresults

Office for National Statistics (2016b). *Births by usual area of residence in the UK.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/birthsbyareaofusualresidenceofmotheruk

Office for National Statistics (2016c). *Families and household in the UK.* Retrieved from https://www.ons.gov.uk/releases/familiesandhouseholdsintheuk2016

Olsen, R. B., Orr, L. L., Bell, S. H., and Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management, 32,* 107-121. doi:10.1002/pam.21660

Public Health England. Public Health Profiles. [13th December 2019] https://fingertips.phe.org.uk

Raynor, P. (2008). Born in Bradford, a cohort study of babies born in Bradford, and their parents: protocol for the recruitment phase. *BMC Public Health*, *8*(1), 1-13.

Rothwell, P. M. (2006). Factors that can affect the external validity of randomised controlled trials. *PLOS Clinical Trials, 1.* doi: 10.1371/journal.pctr.0010009

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised controlled trials. *British Medical Journal,* 340. doi: 10.1136/bmj.c332

Stuart, E. A., Ackerman, B., and Westreich, D. (2018). Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on Social Work Practice*, *28*(5), 532-537.

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*(3), 475-485.

Tomlinson, M., Ward, C. L., and Marlow, M. (2015). Improving the efficiency of evidence-based interventions: the strengths and limitations of randomised controlled trials. *South African Crime Quarterly, 51.* 43-53. doi: 10.4314/sacq.v51i1.5

Travers, J., Marsh, S., Williams, M., Weatherall, M., Caldwell, B., Shirtcliffe, P., Aldington, S., and Beasley, R. (2007). External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax, 62.* 219-223. doi: 10.1136/thx.2006.066837

Webster-Stratton, C. (2010). *The Incredible Years parenting programme's user manual.* Unpublished manuscript.