Psychological outcomes of 12-15 year olds with gender dysphoria receiving pubertal suppression: assessing reliable change and recovery

Short title: Psychological outcomes of puberty blockers

Professor Susan McPherson
School of Health and Social Care
University of Essex
Colchester CO4 3SQ, UK

David E.P. Freedman
Independent Researcher
London, UK

Abstract

In 2021 the results of the first UK study of GnHRA treatment in young adolescents was published. Based on means testing alone, the study found no deterioration in psychological wellbeing. The aim of the current study is to re-analyse the data to assess Reliable Change and Recovery rates in order to assess individual level change and provide a more accurate assessment of treatment risks. The data from the original uncontrolled prospective observational study were collected within a national specialist service for children with gender dysphoria. Participants were 44 12-15 year olds diagnosed with gender dysphoria meeting pre-specified eligibility criteria for GnHRA treatment. Puberty was suppressed using "triptorelin"; participants were followed-up for 36 months. A range of outcomes was assessed including response to pubertal suppression; bone health; adverse events; and psychological outcomes. Because the primary justification for use of GnHRA treatment in this age group is to relieve psychological distress caused by onset of puberty, our analysis focuses on general psychopathology measured by the Child Behaviour Checklist and Youth Self Report form. Contrary to conclusions of the original study, our results indicate that between 15% and 34% of participants reliably deteriorated depending on the subscale and time point. Relatively few participants fell into the borderline or clinical range at baseline (37%-58% depending on the scale and time point); recovery rates were low and fell to zero on both self-report and parent-report at 36 months. Rates of reliable deterioration as well as recovery rates were comparatively worse than typical child and adolescent mental health services where recovery rates of around 50% are found. These findings are concerning and indicate an urgent need to re-evaluate conclusions drawn from the original study and for more detailed analysis of outcomes to inform ongoing use of GnHRA treatment in young adolescents in the UK.

Introduction

The UK Gender Identity Development Service (GIDS) was established in 1989 providing therapeutic assessment and psychological intervention for children and young people experiencing gender dysphoria (GD). An early audit of the service by the current authors found that of the first 124 referrals to the service, the majority had complex social and/or psychological difficulties in addition to gender dysphoria [1]. At that time, GIDS was based on a psychotherapeutic service model based within an NHS specialist mental health Trust in London. When children reached adulthood (at least 16) and wanted to pursue medical interventions such as puberty blockers, cross-sex hormones or surgery, they would be referred to adult services in a hospital setting. In 2009 the service commissioning structure at GIDS changed from a London based service receiving extra-contractual referrals on a case-by-case basis to a nationally commissioned specialist service [2]. Referrals began to rapidly increase seeing a 2800% increase from 2009 to 2019 [2] while levels of complex social and psychological difficulties in the client group remained high [3].

Puberty blockers (GnRHa treatment) are a form of medical intervention, licenced for use to retard puberty in young people with precocious puberty, that have additionally been prescribed for people experiencing GD which had previously been restricted to those aged 16 years and over. The intervention had been trialled on younger children in the Netherlands since the late 1990s [4]. Pressure from patient groups to introduce the treatment for younger children in the UK grew [2]. From around 2009 a number of countries including USA and Australia began to permit its use in this younger age group [5] and in 2009 the British Society of Endocrinology and Diabetes published a statement advocating their use in a cautious and carefully managed way, monitored within a research context [2]. In 2011, GIDS applied for ethical approval to trial the treatment in children aged 12 to 15 within the context of a research study.

The proposed study received ethical approval from the UK National Research Ethics Service in 2011 and recruitment of participants took place from 2011 to 2014 [5]. The aims of the study were "to evaluate the benefits and risks for physical and mental health and wellbeing of mid-pubertal suppression in adolescents with GD; to add to the evidence base regarding the efficacy of GnRHa treatment for young people with GD; and to evaluate continuation and discontinuation of GD and the continued wish for gender reassignment within this group". The study assessed physical response to pubertal suppression; bone health; adverse events; and psychological outcomes. Psychological outcomes included general psychopathology, self-harm, quality of life, body image, GD, general functioning and patient experience. General psychopathology was assessed using the Child Behaviour Checklist (CBCL, parent report) and the Youth Self Report Form (YSR, self-report). These are validated measures which have been widely used internationally to assess children's mental health [6] and are arguably the most reliable indicator of general mental health among the outcomes assessed in the study.

The main argument for the introduction of puberty blockers in the UK for this age group had been on the grounds of their potential to relieve psychological distress.

> *"A key purpose of GnRHa treatment is to pause puberty, to avoid a deterioration in wellbeing and allow for further exploration of a young person's feelings about their gender identity and their wishes for the future, without the pressure or distress which may come from further unwanted bodily changes"* [7]

There were, however, unknown risks of treatment. The participant information leaflet [8] referred to a number of risks including the potential to limit bone strength, sexual development, height, fertility along with impacts on memory and concentration. Introducing a medical intervention for children with such serious risks was balanced against the potential for psychological benefit. Our focus in the

current paper is therefore on psychological outcomes of children receiving this intervention as measured by the general psychopathology indicators (CBCL and YSR).

The GIDS study was an uncontrolled pre-post design. Children meeting the eligibility criteria were referred by GIDS to University College London Hospital NHS Foundation Trust between 2011-2014, where study information was given and consent taken. Forty four children consented to take part in the study. Detailed study procedures are provided in the published paper [5]. Participants were assessed at baseline, 12 months, 24 months and 36 months. Analysis of psychological outcome data involved statistical tests to compare differences between mean scores on each subscale at each time point. This form of analysis falls within the statistical tradition of Null Hypothesis Significance Testing which involves comparing group means, in this case the same group at different time points.

"Significance testing almost invariably retards the search for knowledge by producing false conclusions about research literature" [9]. This is not an unusual or fringe perspective. Similar concerns have been outlined by key methodologists [10–12]. The central argument is that unlike natural sciences, when measuring social and psychological constructs, the null hypothesis is never true: there are always differences between people on these sorts of constructs. As such, evaluation of both adult and child psychological services in the UK use approaches which examine data at the level of individual change rather than group means [13,14]. These approaches typically assess two aspects of individual change: reliable change and clinically significant change, the latter sometimes referred to as recovery.

Reliable Change is individual change that is sufficiently unlikely to have arisen by measurement error alone. The approach provides summary information about what proportion of the sample improve, deteriorate or stay the same (no change). The formula takes into account the standard error of difference (before and after treatment) as well as the internal reliability of the measure [15]. The

analysis is "applicable, in one form or another, to the measurement of change on any continuous scale for *any* clinical problem"[15]. The approach can be used with clinical outcome data whether as part of a controlled or uncontrolled research study, or as part of routine outcome evaluation in a clinical setting. Clinically significant change or "recovery" refers to the proportion of patients who are within the clinical or borderline range at baseline, show reliable improvement and move into a non-clinical range [14].

The GIDS study had no control group or comparison group, justified on grounds that this was the only feasible design for this client group (see https://gids.nhs.uk/research/early-intervention-study/). The analysis therefore consisted of comparing the mean scores from the same group of individuals at different time points. The authors indicate that the analysis was designed "to minimise the likelihood of chance findings due to the large number of outcomes and small sample size" [5]. This means that the number of statistical tests was limited and some outcomes had to be analysed in broad groups with no sub-group analyses, such as by natal sex. Analysis of Reliable Change and Clinically Significant Change are not hampered by issues of multiple testing and is suitable for small or large samples.

The GIDS study found no statistical differences on the CBCL or YSR between time points, concluding that either puberty blockers "brought no measurable benefit nor harm to psychological function"; or that "treatment reduced [the] normative worsening of problems". The latter conclusion is based on evidence that psychological problems as measured by the YSR and CBCL tend to worsen during early adolescence. However, the evidence cited comes from non-clinical populations (Verhulst, 2003). Children who attend GIDS are a clinical population [1,3], with a similar range of social and psychological problems to those seen in CAMHS populations. Moreover, the children in the GIDS study were ostensibly receiving mental health support from the GIDS service in a therapeutic setting, making them a clinical population by definition. A reasonable benchmark is therefore the UK CAMHS

146 population rather than non-clinical populations. In this paper we argue that neither of the study's
147 conclusions is accurate, because the analysis did not take account of individual level change as is
148 standard practice when evaluating CAMHS services in the UK, nor did the authors benchmark their
149 interpretations against CAMHS psychological outcomes.
150
151 Data from the GIDS study were lodged at the UK Data Archive [8] and can be downloaded for
152 researcher use without any further additional ethical approval required. The aim of the current
153 study was therefore to re-analyse the data from the GIDS study to assess Reliable Change and
154 Recovery rates on the CBCL and YSR for the sample at 12, 24 and 36 months follow-up.
155
156 **Methods**
157
158 Study data were anonymised and lodged at the UK Data Archive in a format suitable for use with no
159 further ethical approval or consent required, as agreed with the Health Research Authority [8].
160 Please see the original publication for more detailed information on study methodology and further
161 details on the ethic approval which was obtained from the National Research Ethics Service (NRES:
162 reference 10/H0713/79) in February 2011 [5]. Subsequently, the team had discussions with the
163 Health Research Authority who provided permission for data to be deposited with the UK Data
164 Archive on the condition that sensitive data was removed to minimise disclosure risk of personal
165 information [5].
166
167 The study data were downloaded and converted to Excel. The data deposited were standardised for
168 age and sex and included individual level scores for CBCL Externalising, CBCL Internalising, CBCL Total
169 Problems, YSR Externalising, YSR Internalising and YSR Total Problems. These sub-scales are "higher-
170 order" subscales which combine other subscales into broad dimensions [6]. The Internalising domain
171 is considered a measure of general emotional problems including anxiety, depression, somatic

172 complaints and being withdrawn. The Externalising domain is considered an aggregate scale of
173 behavioural problems including attention problems and aggressive behaviour. The Total Problems
174 score is considered an aggregate of emotional and behavioural problems as well as sleep difficulties
175 [16]. The scales have published reliability data reporting good internal reliability with Chronbach
176 alphas ranging from 0.90 to 0.97 [6]. The scales come from a broad system of measures (ASEBA)
177 developed in the USA which have been widely used with published norms and reliability data and
178 clinical cut-offs for all versions. Established cut-points are <60 (normal range), 60-63 (borderline
179 range) and >63 (clinical range) [6].
180
181 The Reliable Change Index (RCI) for each sub-scale was calculated using the formula provided by
182 Evans, Margison and Barkham using Chronbach alpha as the reliability criterion [15]. Using the RCI,
183 we calculated the proportion of participants in each of the three categories: No change, Deteriorate.
184 Improve.
185
186 Recovery (clinically significant change) was calculated using the published clinical cut-points.
187 Following Gibbons, Harrison and Stallard [14], participants who scored in the borderline or clinical
188 range at baseline (greater than or equal to 60) were included in the analysis of clinically significant
189 change. Of those meeting this criterion, we calculated the proportion who had moved into the
190 normal range at each time point.
191 .
192
193 **Results**
194
195 The findings for Reliable Change and recovery are presented below by Internalising Problems (Table
196 1), Externalising Problems (Table 2) and Total Problems (Table 3). Data indicate that across all scales
197 with both self-report (YSR) and parent report (CBCL), the majority of participants experience no

reliable change in distress across all time points. Between 15% and 34% deteriorate and between 9% and 20% reliably improve.

There are relatively small numbers of participants in the clinical range at baseline: roughly 40-60% of the overall sample, depending on the scale. Rates of reliable recovery range from 0% to 45% depending on the scale and time point.

Table 1: CBCL and YSR Internalising Scales – reliable change and reliable recovery

CBCL Internalising				**YSR Internalising**		
Reliable change	12m	24m	36m	Reliable change	12m	24m
No change	68%	60%	64%	No change	56%	67%
Deteriorate	20%	20%	27%	Deteriorate	29%	20%
Improve	12%	20%	9%	Improve	15%	13%
N	41	20	11	N	41	15
Clinical/borderline range at baseline	Reliably recovered			Clinical/borderline range at baseline	Reliably recovered	
24/43 (56%)	2 (8%)	1 (4%)	0 (0%)	20/44 (45%)	3 (15%)	0 (0%)

211 Table 2: CBCL and YSR Externalising Scales – reliable change and reliable recovery

212

CBCL Externalising				**YSR Externalising**		
Reliable change	12m	24m	36m	Reliable change	12m	24m
No change	66%	70%	64%	No change	61%	67%
Deteriorate	15%	15%	18%	Deteriorate	22%	20%
Improve	20%	15%	18%	Improve	17%	13%
N	41	20	11	N	41	15
Clinical/borderline range at baseline	Reliably recovered			Clinical/borderline range at baseline	Reliably recovered	
16/43 (37%)	4 (25%)	1 (6%)	1 (6%)	12/44 (27%)	2 (17%)	0 (0%)

213

214

215

216

217

218

219

220

221

222

223

224 Table 3: CBCL and YSR Total Scales – reliable change and reliable recovery

225

CBCL Total				**YSR Total**		
Reliable change	12m	24m	36m	Reliable change	12m	24m
No change	49%	60%	55%	No change	37%	60%
Deteriorate	29%	20%	18%	Deteriorate	34%	27%
Improve	22%	20%	27%	Improve	29%	13%
N	41	20	11	N	41	15
Clinical/borderline range at baseline	Reliably recovered			Clinical/borderline range at baseline	Reliably recovered	
25/43 (58%)	4 (16%)	2 (8%)	0 (6%)	19/44 (43%)	6 (32%)	0 (0%)

226

227

228 **Discussion**

229

230 The GIDS Early Intervention Study authors claim that the study findings indicate there is no
231 deterioration in psychological wellbeing for these 12-15 year olds using puberty blockers [5]. This
232 conclusion was based on a statistical analysis comparing means at different time points. However,
233 this is not a suitable analysis to show the effect of the treatment on individuals in the study and does
234 not allow for comparison with other clinical populations experiencing psychological distress.

235

Our analysis indicates that, broadly speaking, for Internalising and Externalising Problems, 56%-68% experience no reliable change in distress across time points and although there is some variation, proportions do not appear markedly different between self-report and parent report. Between 15% and 29% deteriorate; and between 9% and 20% reliably improve. The Total Problems scale shows even higher proportions deteriorating (20%-34% depending on timepoint).

A rate of around 20% reliable improvement is not dissimilar to other CAMHS service evaluations. For example, using a similar global scale of psychopathology (RCADS), Gibbons, Harrison and Stallard [14] found between 20.7% and 24.4% of their sample reliably improved. However, the same study found that nearly all of the remaining participants showed No Change with only a very small proportion (0.7% - 5.7%) deteriorating. These proportions were similar for a range of other psychological outcome variables used[14]. Comparatively high levels of deterioration in the GIDS sample (ranging from 15-34%) is therefore very concerning. Note that the highest rate of deterioration (34%) is seen in the self-report scale at 12 months and only slightly reduces by 24 months to 27%.

Relatively few participants fell into the borderline or clinical range at baseline (37%-58% depending on the scale). This is lower than a typical CAMHS population; Gibbons, Harrison and Stallard found around 90% of their sample to be in the clinical or borderline range at baseline[14]. This suggests that the GIDS participants were less likely to be significantly distressed when they were referred to the clinic than typical CAMHS referrals and may indicate a relatively low referral threshold operating at the time period studied.

Given the relatively low proportion of the sample in the clinical or borderline range, the recovery rates should be treated cautiously. Nevertheless, a notable finding is that there is a tendency for parent report and self-report to show a discrepancy in recovery rates with self-report scores

indicating a higher rate of recovery after 12 months (around 25-45%). However, this discrepancy reduces over time so that at the later measurement points, self-report and parent report recovery rates converge at zero or near to zero. CAMHS services typically see recovery rates around 50% [14] and therefore the recovery rates seen in the GIDS sample are worryingly low, particularly given the rates fall to zero over a prolonged period (3 years).

Best interest decisions about treatments for children with significant risks to current and future physical health are normally taken on a balance of risks, that the treatment improves some other important aspect of wellbeing or health. The argument for use of puberty blockers in this age group was that they relieve psychological distress. Yet these data indicate that while reliable improvement rates are similar to CAMHS services, deterioration rates are considerably higher than in CAMHS services and reliable recovery rates considerably lower, reducing over time. This raises serious questions about the balance of risk.

In light of this analysis, the service should urgently conduct fuller analysis to ascertain whether there are any variables which might predict which 20% of children with GD are most likely to benefit psychologically and who are most likely to deteriorate. If there is a lack of robust data to conduct statistical modelling, this could at the very least consist of observation of patterns of individual change, taking into account demographic variables available including natal sex.

The GIDS study design was, in effect, a service evaluation conducted under the auspices of research ethics. Referrals for puberty blockers became routine practice in GIDS 2014 (Barnes, 2023), 7 years before the findings of the study were published. It would be expected therefore that the service would have continued to collect routine outcome data, as is expected across all UK CAMHS and IAPT CYP services and was standard practice in the CAMHS service in the Trust where the service was based. It should therefore be straightforward to also analyse a much larger data set of psychological

outcomes using the Reliable Change Index and clinically significant change approach to provide a more helpful picture of the impact of puberty blockers at the level of individual change. It should also be possible to analyse routinely collected outcome data for children who did not take up puberty blockers to generate comparable data on Reliable Change and recovery as a naturalistic control group. We recommend that the service conduct such an analysis urgently to inform the development of policy for new gender services being established in the UK.

Strengths and limitations

This is the first analysis of UK data on children aged 12-15 with GD taking puberty blockers demonstrating individual level change as opposed to testing differences between group averages. As such this analysis gives a much fuller picture of the benefits and risks of treatment. However, the analysis is limited by absence of UK norms for the measures used combined with the lack of a study control group to provide local norms; clinical cut points used were based on US normative data and it is possible that UK teenagers may present differently to US teenagers in respect of GD. The analysis is also limited by the limitations of the data set available for re-analysis which lacks differentiation by sex and item level data to look at more fine grained sub-scales.

References

1. Di Ceglie D, Freedman D, McPherson S, Richardson P. Children and adolescents referred to a specialist gender identity development service: Clinical features and demographic characteristics. Int J Transgenderism. 2002;6.

2. Barnes H. Time to Think: The Inside Story of the Collapse of the Tavistock's Gender Service for Children. London: Swift Press; 2023.

3. Holt V, Skagerberg E, Dunsford M. Young people with features of gender dysphoria: Demographics and associated difficulties. Clin Child Psychol Psychiatry. 2016;21: 108–118.

doi:10.1177/1359104514558431

4. Cohen-Kettenis PT, van Goozen SHM. Pubertal delay as an aid in diagnosis and treatment of a transsexual adolescent. Eur Child Adolesc Psychiatry. 1998;7: 246–248. doi:10.1007/s007870050073

5. Carmichael P, Butler G, Masic U, Cole TJ, De Stavola BL, Davidson S, et al. Short-term outcomes of pubertal suppression in a selected cohort of 12 to 15 year old young people with persistent gender dysphoria in the UK. Santana GL, editor. PLoS One. 2021;16: e0243894. doi:10.1371/journal.pone.0243894

6. Achenbach TM, Rescorla LA. Manual for the ASEBA School-Age Forms & Profiles. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families; 2001.

7. Gender Identity Development Service. The Early Intervention Study. Available: https://gids.nhs.uk/research/early-intervention-study

8. Carmichael P, Butler G, Masic U, Cole TJ, De Stavola BL, Davidson S, et al. Short-term outcomes of pubertal suppression in a selected cohort of 12 to 15 year old young people with persistent gender dysphoria in the UK 2019-2021 [Data Collection]. Colchester, Essex: UK Data Service; 2021. doi:10.5255/UKDA-SN-854413

9. Schmidt F, Hunter J. Are there benefits from NHST? Am Psychol. 2002;57: 65–66. doi:10.1037/0003-066X.57.1.65

10. Meehl PE. Theory-Testing in Psychology and Physics: A Methodological Paradox. Philos Sci. 1967;34: 103–115. doi:10.1086/288135

11. Clark-Carter D. Effect size: The missing piece of the jigsaw. Psychologist. 2003;16: 636–638.

12. Field A, Wright D. A bluffer's guide to effect sizes. PsyPAG Q. 2006;58: 9–23.

13. Wolpert M, Jacob J, Napoleone E, Whale A, Calderon A, Edbrooke-Childs J. Child- and Parent-Reported Outcomes and Experience from Child and Young People's Mental Health Services

2011–2015. Child Outcomes Research Consortium; 2016. Available: https://www.corc.uk.net/media/1544/0505207_corc-report_for-web.pdf

14. Gibbons N, Harrison E, Stallard P. Assessing recovery in treatment as usual provided by community child and adolescent mental health services. BJPsych Open. 2021;7: e87. doi:10.1192/bjo.2021.44

15. Evans C, Margison F, Barkham M. The contribution of reliable and clinically significant change methods to evidence-based mental health. Evid Based Ment Health. 1998;1: 70–72. doi:10.1136/ebmh.1.3.70

16. Guerrera S, Menghini D, Napoli E, Di Vara S, Valeri G, Vicari S. Assessment of Psychopathological Comorbidities in Children and Adolescents With Autism Spectrum Disorder Using the Child Behavior Checklist. Front Psychiatry. 2019;10. doi:10.3389/fpsyt.2019.00535