A randomized, controlled trial of the efficacy and harms of paroxetine and imipramine in the treatment of adolescent major depression: Restoring Study 329

Joanna Le Noury, School of Medical Sciences, Bangor University
Research psychologist

John M Nardo, Emory University,
Clinical assistant professor, retired

David Healy, School of Medical Sciences, Bangor University
Professor

Jon Jureidini, Paediatric Mental Health Training Unit, University of Adelaide
Clinical professor (corresponding author)

Melissa Raven, Discipline of Public Health, Flinders University
Adjunct lecturer

Catalin Tufanaru, Faculty of Health Sciences, University of Adelaide
Research associate

Elia Abi-Jaoude, Department of Psychiatry, University of Toronto

The first four authors made equal contribution to the paper. All authors meet ICMJE authorship criteria. Correspondence to: Jon.Jureidini@adelaide.edu.au

Guarantor Jon Jureidini

Competing interests
From 2003 to 2011, Dr. Healy was an expert witness for plaintiffs in legal cases involving GlaxoSmithKline's drug paroxetine. He continues to be a witness for plaintiffs in actions involving other antidepressants with the same mechanism of action as paroxetine.

Dr. Jureidini was paid by Baum, Hedlund, Aristei & Goldman, Los Angeles, California to provide expert analysis and opinion about documents obtained from GlaxoSmithKline in a class action over study 329, and from Forrest in relation to a paediatric citalopram RCT. He is a member of Healthy Skepticism.

Dr. Raven is a member of Healthy Skepticism.

Drs Le Noury, Nardo, Tufanaru and Abi-Jaoude have nothing to declare.

A randomised, controlled trial of the efficacy and harms of paroxetine and imipramine in the treatment of adolescent major depression: Restoring Study 329

Abstract

Objectives: The primary objective was to compare the efficacy and safety of imipramine and paroxetine to placebo in the treatment of adolescents with unipolar major depression. Data from a randomised controlled trial (GSK's Study 329) published by Keller et al. in 2001 were reanalysed under the Restoring Invisible and Abandoned Trials (RIAT) initiative.

Design: Randomised placebo-controlled trial.

Setting: 12 academic psychiatry centres (10 US, 2 Canadian), from 1994 to 1998,

Participants: 275 adolescents (12 to 18 years old) with major depression at least 8 weeks in duration. Exclusion criteria included a range of comorbid psychiatric and medical disorders and suicidality.

Interventions: Participants were randomised to 8 weeks double-blind treatment with paroxetine (20–40 mg), imipramine (200–300 mg), or placebo.

Main outcome measures: The pre-specified primary efficacy variables were: change from baseline to the end of the acute treatment phase in total Hamilton Depression Scale (HAM-D) score; and the proportion of responders (HAM-D score ≤8 or ≥50% reduction in baseline HAM-D) at acute endpoint. Pre-specified secondary outcomes were (1) changes from baseline to endpoint in the following parameters: depression items in K-SAD-L; Clinical Global Impression; Autonomous Functioning Checklist; Self-Perception Profile; Sickness Impact Scale, (2) predictors of response, (3) number of patients who relapse during the maintenance phase.

Results: The responses to paroxetine and imipramine were not statistically or clinically significantly different from placebo for any pre-specified primary or secondary efficacy outcome. HAM-D scores decreased by 10.73, 8.95 and 9.08 points, respectively, for the paroxetine, imipramine and placebo groups ($p = 0.204$). Clinically significant increases in harms, including suicidal ideation and behaviour and other serious adverse events, were observed in the paroxetine group.

Conclusions: Paroxetine was neither well tolerated nor effective for major depression in adolescents. Imipramine, given in high doses, was also poorly tolerated and was not shown to be effective. This study has demonstrated that when there is access to primary data, trial conclusions will ordinarily be provisional rather than authoritative.

Trial registration: Registration number and name of trial register: SmithKline Beecham study 29060/329.

Funding of Study 329: SmithKline Beecham/GlaxoSmithKline. No funding was obtained to support this restoration.

Supplementary material / data can be found at [URL TBA]

A randomised, controlled trial of the efficacy and harms of paroxetine and imipramine in the treatment of adolescent major depression: Restoring Study 329

Background

In 2013, in the face of the selective reporting of outcomes of randomised controlled trials (RCTs), an international group of researchers called on funders and investigators of abandoned (unpublished) or misreported trials to publish undisclosed outcomes or correct misleading publications.[1] This initiative was dubbed 'restoring invisible and abandoned trials' (RIAT). The researchers identified many trials requiring restoration, and emailed the funders, asking them to signal their intention to publish the unpublished trials or publish corrected versions of misreported trials. Should funders and investigators fail to undertake to correct a trial that has been identified as unpublished or misreported, independent groups were encouraged to publish an accurate representation of the clinical trial based on the relevant regulatory information.

The current article represents a RIAT publication of Study 329. The original study was funded by SmithKline Beecham (SKB; subsequently GlaxoSmithKline, GSK) and led by Dr Martin Keller. We acknowledge the work of the original investigators. This double-blinded RCT to evaluate the efficacy and safety of paroxetine, imipramine and placebo for adolescents diagnosed with major depression was reported in the *Journal of the American Academy of Child and Adolescent Psychiatry* (*JAACAP)* in 2001 (hereafter 'Keller et al.'). [2] The RIAT researchers named Study 329 as an example of a misreported trial in need of restoration. Keller et al., which was largely ghostwritten,[3] claimed efficacy and safety for paroxetine at odds with the data.[4] This is problematic because the article has been influential in the literature supporting the use of antidepressants in adolescents.[5]

On 14 June 2013, the RIAT researchers asked GSK whether it had any intention to restore any of the trials it sponsored. GSK did not signal any intent to publish a corrected version of any of its trials. In later correspondence, GSK stated that it does 'not agree that the article is false, fraudulent or misleading', and asserted that Keller et al. 'accurately reflects the honestly-held views of the clinical investigator authors'.[6]

Study 329 was a multicenter eight-week double-blind RCT (acute phase), followed by a six-month continuation phase. SKB's stated primary objective was to compare the efficacy and safety of imipramine and paroxetine to placebo in the treatment of adolescents with unipolar major depression. Secondary objectives were to identify predictors of treatment outcomes across clinical subtypes; to provide information on the safety profile of paroxetine and imipramine when these agents were given for 'an extended period of time'; and to estimate the rate of relapse among imipramine, paroxetine and placebo responders who were maintained on treatment. Study enrolment took place between April 1994 and March 1997.

The first RIAT trial publication was a surgery trial that had only been partly published before.[7]. As far as we are aware, this is the first time that a previously published RCT has been reported in a published paper by a different team of authors.

Methods

We have reanalysed Study 329 according to the RIAT recommendations. To this end, we have used the Clinical Study Report (CSR; SKB's 'Final Clinical Report') available on the GSK website,[8] other publically available documents,[9] and the data access system SAS Solutions OnDemand,[10] on which GSK has posted some Study 329 documents (available only to users approved by GSK). Following negotiation,[11] GSK posted de-identified individual case report forms (CRFs) on that site. A table of sources of data consulted in preparing each part of this paper is available as Appendix 1.

Except where indicated, in accordance with RIAT recommendations, our methods are those set out in the 1994 Study 329 protocol,[12] as outlined in our RIAT Audit Record (RIATAR) (Appendix 1). In cases where the methodology published by Keller et al. diverged from the protocol, we followed the protocol. Where the protocol was not specific, we chose by consensus standard methods that best presented the data. The original 1993 protocol had minor amendments in 1994 and 1996. Furthermore, the CSR reported some procedures that varied from those specified in the protocol, and we have noted variations wherever they were considered significant.

Participants

275 adolescents between the ages of 12 and 18 years, meeting *DSM-IV* criteria[13] for a current episode of major depression of at least 8 weeks duration, were recruited for the study (the protocol specified *DSM-III-R* criteria, which are very similar). Table 1 lists the eligibility criteria.

Table 1. Study eligibility criteria.

Inclusion Criteria	Exclusion Criteria
Adolescents between ages of 12 and 18, meeting *DSM-III-R* criteria for major depression for at least 8 weeks; Child Global Assessment Scale severity score < 60; Hamilton Depression Scale (17-item) score ≥ 12; Medically healthy; IQ ≥ 80 (based on Peabody Picture Vocabulary Test).	Current or past *DSM-III-R* diagnosis of: bipolar disorder, schizoaffective disorder, anorexia nervosa, bulimia, alcohol or drug abuse/dependence, obsessive-compulsive disorder, autism/pervasive mental disorder, or organic psychiatric disorder; Current (within 12 months) *DSM-III-R* diagnosis of post-traumatic stress disorder; Adequate antidepressant trial within 6-months; Suicidal ideation with a definite plan, suicide attempt during current depressive episode, or history of suicide attempt by medication overdose; Medical illness which contraindicates the use of heterocyclic antidepressants; Current use of psychotropic medications (including anxiolytics, antipsychotics, mood stabilizers), or illicit drugs; Organic brain disease, epilepsy or mental retardation;

	Patients who are pregnant or lactating; Sexually active females not using reliable contraception; Use of an investigational drug within 30 days or within five half-lives of the investigation drug.

An undisclosed number of patients identified by telephone screening as potential participants were subsequently evaluated at the study site by a senior clinician (psychiatrist or psychologist). Multiple meetings and teleconferences were held by the sponsoring company with site study investigators to ensure standardization across sites. Patients and parents were interviewed separately using the Schedule for Affective Disorders and Schizophrenia for Adolescents - Lifetime Version (K-SADS-L). Following this initial assessment, the study informed consent form was signed by both patient and parent; there is no mention of a separate assent form in the protocol or in the clinical study report. A 7 to 10 day screening period was used to obtain past clinical records and to document that the depressive symptoms were stable. At the end of the screening period, only patients continuing to meet the inclusion criteria (DSM-III-R major depression and the HAM-D total score of 12 or greater) were randomized. There was no placebo lead-in phase.

The protocol called for 300 subjects based on the estimated 80% power required to detect a four-point difference between placebo and active drug groups, a difference deemed by the protocol to be clinically significant. In addition, the number of sites was increased from 6 centres to 12 (10 in the United States and 2 in Canada). The centres were affiliated with either a university or a hospital psychiatry department and had experience with adolescent patients. The investigators were selected for their interest in the study and their ability to recruit study patients.

The recruitment period ran from April 1994 until 15 March 1997, and the acute phase was completed on 7 May 1997. In a small number of patients, 30-day follow-up data in cases that went into the continuation phase were collected into 1998.

Interventions

Study medication was provided to patients in weekly blister packs. Patients were instructed to take the medication twice daily. There were 6 dosing levels. Over the first four weeks, all patients were titrated to level 4, corresponding to paroxetine 20 mg or imipramine 200 mg, regardless of response. Non-responders (those failing to reach responder criteria) could be titrated up to level 5 or 6 over the following four weeks. This corresponds to a maximum dose of paroxetine 60 mg and a maximum dose of imipramine of 300 mg.

Medication compliance was evaluated based on the number of capsules dispensed, taken, and returned. Non-compliance was defined as taking less than 80% or greater than 120% of the number of capsules expected to be returned at two consecutive visits, and resulted in withdrawal. Any patient missing two consecutive visits was also withdrawn from the study.

Patients were provided with 45-minute weekly sessions of supportive psychotherapy,[14] primarily for the purpose of assessing the treatment effects.

Outcomes
Patients were evaluated weekly during the 8 week duration of the acute treatment phase.

1. Efficacy Endpoints

Primary Efficacy Variables

The pre-specified primary efficacy variables were: change in total Hamilton Depression Scale (HAM-D)[15] score from the beginning of the treatment phase to the endpoint of the acute phase; and the proportion of *responders* at the end of the eight week acute treatment phase. *Responders* were defined as patients who had a 50% or greater reduction in the HAM-D or a HAM-D score equal to or less than 8. (Scores on the HAM-D can vary from 0 to 52.)

Secondary Efficacy Variables
The pre-specified secondary efficacy variables were:
a) Changes from baseline to endpoint in the following parameters:

- Depression items in K-SAD-L
- Clinical Global Impression (CGI)
- Autonomous Functioning Checklist[16] (listed in the protocol as Autonomic Function Checklist)
- Self-Perception Profile
- Sickness Impact Scale.

b) Predictors of response (endogenous subtypes, age, prior episodes, duration and severity of present episode, comorbidity with separate anxiety, attention deficit, and conduct disorder).
c) The number of patients who relapse during the maintenance phase (referred to in the CSR and in this paper as 'continuation phase').

However, both before and after breaking the blind, changes were made by the sponsors to the secondary outcomes as previously detailed.[4] We could not find any document that provided any scientific rationale for these post-hoc changes,[17] and the outcomes are therefore not reported in this paper.

Box 1: Challenges in carrying out RIAT

This is the first RIAT effort by an external team of authors, so there are no clear precedents or guides. **Challenges** included:

Potential or perceived bias
A RIAT report is not intended to be a critique of a previous publication. The point is rather to produce a thorough independent analysis of a trial that has remained unpublished or called into question. We acknowledge, however, that any RIAT team may be seen as having an intrinsic bias, in that questioning the earlier published conclusions is what brought some members of the team together. Consequently, we took all appropriate procedural steps to avoid such putative bias.

Correction for testing multiple variables
We had multiple sources of information: The protocol; the published paper; the documents posted on the GSK web site including the CSR and Individual Patient Data; and the raw primary data in the CRFs provided by GSK on a remote desk-top for this project. The protocol declared two primary and six secondary variables for the three treatment groups in two differing datasets [OC and LOCF]. The CSR contained statistical comparisons on 28 discrete variables using two comparisons [paroxetine vs placebo and imipramine vs placebo] in the two datasets [OC and LOCF]. The published paper listed eight variables with two statistical comparisons each in one dataset [LOCF]. But the original authors nowhere addressed the need for corrections for multiple variables - a standard requirement when there are multiple outcome measures. In the final analysis, there were no statistically or clinically significant findings, so corrections were not needed for this analysis.

Statistical testing
The protocol called for ANOVA testing [GLM] for continuous variables using a model that included the effects of SITE, TREATMENT, and SITE x TREATMENT interaction, with the latter dropped if $p \geq 0.10$. Logistical Regression [chi Square 2x3] was prescribed for categorical variables under the same model. Both methods begin with an omnibus statistic for the overall significance of the dataset, then progress to pairwise testing if and only if the omnibus statistic meets alpha [0.05]. Yet all statistical outcomes in the CSR and published paper were reported only as the pairwise values for only two of the three possible comparisons [paroxetine vs placebo and imipramine vs placebo] with no mention of the omnibus statistic. Therefore, we conducted the needed omnibus analyses, which are negative as shown. The pairwise values are available in the online Appendix 2 (table i).

Missing values
The protocol called for evaluation of the OC and LOCF datasets, with the latter being definitive. The LOCF method for correcting missing values was the standard at the time the study was conducted. It continues to be widely used, though newer models such as Multiple Imputation or Mixed Models are now frequently preferred. We chose to stick to the protocol and use the LOCF method rather than introduce a *post hoc* analytic tool.

Non-protocol specified outcome variables
There were four outcome variables in the CSR and in the published paper that were not specified in the protocol. These were the only outcome measures reported as significant. They were in no version of the protocol as amendments nor were they submitted to the Institutional Review Board. The CSR (section 3.9.1) states they were part of an 'analysis plan' developed some two months before the blind was broken. No such plan appears in the CSR and we have no contemporaneous documentation of that claim, despite having repeatedly requested it from GSK.

Conclusions:

After prolonged discussions, we decided that the best and most unbiased course of action was to analyse the efficacy data in the IPD based on the last guaranteed *a priori* version of SKB's own protocol [1994]. Although the protocol omitted a discussion of corrections which we would

have thought necessary, correction for multiple variables is designed to prevent false positives and there were no positives. We agreed with the statistical mandates of the protocol, but while we saw pairwise comparisons in the absence of overall significance as inappropriate, we recognize that this is not a universal opinion, so we included them in the online Appendix 2, table i.

Finally, although investigators can explore the data however they wish, additional outcome variables outside those in the protocol cannot be legitimately declared once the study is underway, except as 'exploratory variables' - appropriate for the discussion or as material for further study, but not for the main analysis. The *a priori* protocol and blinding are the bedrock of a randomized controlled trial - guaranteeing that there is not even the possibility of the HARK phenomenon ['hypothesis after results known']. While we can readily demonstrate that none of the reportedly 'positive' four non-protocol outcome variables stands up to scrutiny, the primary mandate of the RIAT enterprise is to reaffirm essential practices in RCTs, so we did not include these variables in our efficacy analysis.

2. Harm Endpoints

An adverse experience/event (AE) was defined in the protocol (p. 18) as:

> 'any noxious, pathologic or unintended change in anatomical, physiologic or metabolic functions as indicated by physical signs, symptoms and/or laboratory changes occurring in any phase of the clinical trial whether associated with drug or placebo and whether or not considered drug related.
>
> This includes an exacerbation of pre-existing conditions or events, intercurrent illnesses, drug interaction or the significant worsening of the disease under investigation that is not recorded elsewhere in the case report form under specific efficacy assessments.'

AEs were to be elicited by the investigator asking a non-leading question such as: 'Do you feel different in any way since starting the new treatment/the last assessment?'. Details of treatment emergent AEs, their severity, including any change in study drug administration, investigator attribution to study drug, any corrective therapy given, and outcome status were documented. Attribution or relationship to study drug was judged by the investigator to be 'unrelated', 'probably unrelated', 'possibly related', 'probably related' or 'related'.

Vitals signs and ECGs were obtained at weekly visits. Patients with potentially concerning cardiovascular measures either had their medication dose reduced or were withdrawn from the study. In addition, if the combined serum levels (obtained at weeks 4 and 8) of imipramine and desipramine exceeded 500 mcg/ml, the patient was to be withdrawn from the study.

Clinical laboratory tests, including clinical chemistry, hematology and urinalysis were carried out at the screening visit and at the end of week 8. Clinically significant laboratory abnormalities were to be included as adverse events.

Source of harms data

The harms data in this paper cover the acute phase, a taper period and an up to 30-day follow-up phase for those who discontinued because of adverse events. To ensure comparability with Keller et al, none of the tables contains data from the continuation phase.

AE data come from the CSR lodged on GSK's website,[18] primarily Appendix D. Appendix B provides details of concomitant medications. Additional information was available from the summary narratives in the body of the CSR for patients who had AEs that were designated as serious or led to withdrawal. (Of the eleven paroxetine patients with AEs designated as serious, nine discontinued because of AEs.) However, the large number of other patients discontinued because of AEs that were not regarded as serious, or discontinued for lack of efficacy or protocol violations (see Figure 1), did not generate patient narratives. The tables laid out in Appendix D of the CSR give the clinical descriptors used by the blind investigators along with Adverse Drug Events Coding System (ADECS) codes used to code these clinical descriptions, ratings of severity and ratings of relatedness.

It became clear when we examined the key clinical terms that there were a number of events that had been left uncoded into ADECS, and had not been tabulated. For instance, a number of patient narratives of serious AEs that led to discontinuation from the trial contained AEs that had not been coded or assembled within the tables of AEs.

Therefore we approached GSK for access to CRFs. GSK made available all 275 CRFs for patients entered into Study 329. However, the CRFs were only available through a remote desktop facility (SAS Solutions OnDemand Secure Portal)[9], which made it difficult and extremely time-consuming to inspect the records properly.[19] Effectively only one person could undertake the task, with backup for ambiguous cases. Accordingly we could not examine all CRFs. Instead we decided to focus on those 85 participants identified in CSR Appendix H who were withdrawn from the study, along with 8 further participants who were known from prior inspection of the CSRs to have become suicidal. 31 of the CRFs that were checked were from the paroxetine group, 40 from the imipramine group and 22 from placebo.

All CRFs were reviewed by JLN, who is trained in the use of the Medical Dictionary for Regulatory Activities (MedDRA®, MedDRA terminology is the international medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) www.meddra.org). The second reviewer (MN) is a clinician, untrained in this system. There was agreement between these two reviewers about reasons for discontinuation and side effect coding (no quantitative indicator of inter-rater agreement was used).

These 93 CRFs were scrutinised for all AEs occurring during the acute, taper and follow-up phases, and total AEs were compared with the AE totals reported in CSR Appendix D.

This review process gave rise to additional AEs. It also led to recoding of a number of the reasons for discontinuation. The new AEs and the reasons for changing discontinuation category are recorded in Tables ii, iii and x in Appendix 2 accompanying this paper.

Roughly 1000 pages were missing from the CRFs reviewed with no discernible pattern to missing information.

Coding of Adverse Events

All of the initial coding from the clinical descriptions in the CSR was done blind, as was coding from the CRFs. Only for six events from the eleven serious adverse event narratives was it not possible to be blind. This was 0.005% of events.

The original protocol for Study 329 makes no mention of how AEs from this trial would be coded. The CSR specifies that the AEs noted by clinical investigators in this trial were coded using the Adverse Drug Experience Coding System (ADECS) that was being used by SKB at the time. ADECS was derived from a coding system developed by the United States Food and Drug Administration (FDA), Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), but is not itself a recognized system.

We coded AEs using MedDRA, which has replaced COSTART for the FDA, because it is by far the most commonly used coding system today, and it is not possible to access ADECS. For coding purposes, we have taken the original terms used by the clinical investigators as transcribed from the original CRFs into the CSR, and applied MedDRA codes to these descriptions.

In general, MedDRA coding stays closer to the original clinician description of the event than ADECS does. For instance, MedDRA codes 'sore throat' as 'sore throat', but SKB, using ADECS, coded it as 'pharyngitis' (inflammation of the throat). Sore throats may arise because of pharyngitis, but when someone is taking SSRIs they may indicate a dystonic reaction in the oro-pharyngeal area.[20]

Classifying a problem as a 'respiratory system disorder' (inflammation) rather than as a 'dystonia' (a central nervous system disorder) can make a significant difference to the apparent AE profile of a drug.

In staying closer to the original description of events, MedDRA codes suicidal events as 'suicidal ideation' or 'suicidal events' rather than the ADECS option of 'emotional lability'; similarly, aggression is more clearly flagged as 'aggressive events' rather than 'hostility'.

The initial recoding was done blind, but it was not possible to be blind in relation to the 0.005% of additional events located in the serious AE and discontinuation narratives, because allocation status was written into the narrative of the events.

Box 2: Coding Challenges

Most recoding was straightforward. Patient 00039, who had a severe (but not serious) AE, was our most ambiguous case.

Within two weeks of starting the acute phase, this patient was reported as 'more tired' and 'more sick'. There was also an additional handwritten note, 'softness of speech', beside item 8 of the HAM-D, which was rated as 'Obvious retardation at interview'. These were not coded as AEs in CSR Appendix D.

During week 2, the patient was recorded under AEs as being 'more depressed' and having 'superficial scratches'. These were coded by SKB as 'depression' and 'trauma'. We recoded them as 'aggravated depression' and, initially, 'self harm/suicide attempt'.

However, self-harm and suicide attempt are different phenomena. It may or may not be possible to resolve whether self-harm or suicide attempt is the correct coding.

The patient discontinued treatment during the continuation phase. Had she been deemed to have discontinued because of an AE, there would have been a patient narrative that might have made it clearer which of these options was more likely; however, because she was deemed to have discontinued for lack of efficacy, there is no patient narrative.

At the week 6 visit, a number of AEs were noted – 'fatigue', 'more angry' (missing from Appendix D), 'more depressed', 'irritable mood', 'grimacing face' and 'blinking eyes' (the last two were coded together as myoclonus by SKB but were recoded separately by us).

On the basis of being more angry, depressed and irritable, along with an increase in HAM-D suicide item score from 1 or 2 at screening, baseline and the initial weeks of the study to 3 (suicide idea or gesture) in weeks 5 & 6, we opted for 'suicide attempt' as the correct coding for what SKB had coded as trauma at week 2 (see above).

At the final visit, notes were made in a section headed 'adverse experiences', describing the patient as having 'headaches – more severe than usual' and 'Worse general/overall feeling depressed; HAM-D score of 24'.

'Worsening Depression' was not recorded as an AE in Appendix D. The patient was noted as 'OUT OF STUDY' and designated as discontinuation for 'lack of efficacy'. We recoded this as 'Adverse Event (depression worsening)'. Had SKB coded this way, the patient would have required a patient narrative.

Analysis of harms data

In analysing the harms data we have explored the discrepancies in the number of events between CRFs and the CSR; we present all AEs rather than only those happening at a particular rate (as Keller et al. did); we group events into broader system-organ-class (SOC) groups – psychiatric, cardiovascular, gastrointestinal, respiratory and other; we break down events by severity, selecting AEs coded as severe, and utilising the listing in CSR Appendix G of patients who discontinued for any reason; we include an analysis of the effects of prior treatment, presenting the run-in phase profiles of medication taken by patients entering each of the three arms of the study, and comparing the list of AEs experienced by patients on concomitant medication (from Appendix B) versus those not on other medication; and we extract the events occurring during the taper and follow-up phase.

We have not undertaken statistical tests of harms data, as discussed below.

3. Patient withdrawal

A study patient could withdraw or be withdrawn prematurely for any of the following six reasons: 'Adverse experiences including intercurrent illness'; 'Insufficient therapeutic effect'; 'Deviation from protocol including non-compliance'; 'Loss to follow-up'; 'Termination by SB [SKB/GSK]'; 'Other (specify)'.

The CSR states that the primary reason for withdrawal was determined by the investigator. We have reviewed the codes given for discontinuation from the study, which are found in CSR Appendix G, and in a proportion of cases changed these.

Sample Size

The acute phase of the trial was initially based on a power analysis that indicated that a sample size of 100 patients per treatment group was required in order to have a statistical power of 80% for a two-tailed alpha level of 0.05 and an effect size of 0.40. This effect size entailed a difference of 4 in the HAM-D Total change from baseline scores at endpoint, specified in the protocol to be large enough to be clinically meaningful, considering a standard deviation (SD) of 10. No allowance was made in the power calculation for attrition (anticipated dropout rate) or non-compliance during the study.

Recruitment was slower than expected, and reportedly medication supplies (mainly placebo) ran short due to expiry. Therefore a midcourse evaluation of 189 patients was carried out, without breaking the blind, revealing less variability in HAM-D scores (SD 8) than anticipated. Therefore the recruitment target was reduced to 275 on the grounds that it would have no negative impact on the estimated 80% power required to detect a four-point difference between placebo and active drug groups.

Randomisation

A computer-generated randomization list of 360 numbers for the acute phase was generated and held by SKB. According to the CSR, treatments were balanced in blocks of 6 consecutive patients; however, there is an inconsistency in that in CSR Appendix A Randomisation Code details block sizes of both 6 and 8. Each investigator was allocated a block of consecutively numbered treatment packs, and patients were assigned treatment numbers in strict sequential order. Patients were randomised in a 1:1:1 ratio to treatment to paroxetine, imipramine, or placebo.

Blinding

Paroxetine was supplied as film-coated, capsule-shaped yellow (10 mg) and pink (20 mg) tablets. Imipramine (50 mg) was bought commercially and supplied as green film-coated round 50mg tablets. ‘Paroxetine placebos’ matched the paroxetine 20 mg tablets, and ‘imipramine placebos’ matched the imipramine tablets. All tablets were over-encapsulated in bluish-green capsules to preserve blinding.

The blind was to be broken only in the event of a serious AE that the investigator felt could not be adequately treated without knowing the identity of the study medication. The identity of the

study medication was not otherwise disclosed to the investigator or SKB staff associated with the study.

Statistical Methods

The primary population of interest was the intent-to-treat (ITT) population that included all patients who received at least one dose of study medication and had at least one post-baseline efficacy assessment. The demographic characteristics, description of the baseline depressive episode, additional psychiatric diagnoses, and personal history variables of the patients were summarized descriptively by treatment group.

The acute phase eight-week endpoint was of primary interest. Statistical conclusions concerning the efficacy of paroxetine and imipramine were made using data obtained from the last observation carried forward (LOCF, i.e. the last on-therapy assessment during the acute phase) and observed cases (OC) datasets.

We followed the methodology of the a priori 1994 study protocol. It did not provide explicit statistical hypotheses (null hypotheses and alternative hypotheses); nor were there justifications for the proposed statistical approaches or statistical assumptions underlying them.

One of the two primary efficacy variables, proportion of responders (response), and one secondary efficacy variable, proportion of patients relapsing, were treated as categorical variables. The second primary efficacy variable, change in total HAM-D score over the acute phase, and the remaining secondary efficacy variables were treated as continuous variables.

In accordance with the protocol, the continuous variables were analyzed using parametric analysis of variance (ANOVA) with effects in the model including treatment, investigator, and treatment by investigator interaction. Pairwise comparisons were not done if the omnibus (overall) ANOVA was not statistically significant (two-sided $p<0.05$), as specified by the protocol (we acknowledge differing opinions about this issue in the statistical literature [21] so we included them in the online Appendix 2 for completeness). The categorical variable was analyzed using logistic regression, with the same effects included. In either case, if the treatment by investigator interaction resulted in a two-sided p value >0.10, the interaction term was dropped from the model. All statistical tests were done using the Linear Model (LM) and General Linear Models (GLM) procedures of the R statistical package (version 2.15.2)[22] as provided by GSK.

For the relapse rate analyses, we included all responders (HAM-D ≤ 8 or ≥50% reduction in symptoms) meeting the original criteria for entry to the continuation phase of the study. Patients were considered to have relapsed if they no longer met the responder criteria (HAM-D ≤8 or ≥50% reduction in symptoms) or if they were withdrawn for 'Intentional Overdose'.

Results

The demographics of the groups are shown in Table 2, along with depression parameters, comorbidities, and baseline scores for the efficacy variables.

Table 2. Baseline characteristics

	Paroxetine n = 93	Imipramine n = 95	Placebo n = 87
Age (yr) [SD]	**14.8 [1.6]**	**14.9 [1.6]**	**15.1 [1.6]**
Sex M/F	**35/58**	**39/56**	**30/57**
Race %			
Caucasian	**82.8%**	**87.4%**	**80.5%**
African American	**5.4%**	**3.2%**	**6.5%**
Asian American	**1.1%**	**2.1%**	**2.3%**
Other	**10.8%**	**7.4%**	**10.3%**
Depression			
Episode duration (mo) [SD]	**14 [18]**	**13 [17]**	**13 [17]**
Age first episode (yr) [SD]	**13.1 [2.8]**	**13.7 [2.7]**	**13.5 [2.3]**
Prior episodes 0	**0%**	**2%**	**0%**
1	**81%**	**79%**	**77%**
2	**12%**	**14%**	**14%**
>3	**7%**	**6%**	**8%**
Comorbidity			
Any comorbid disorder §%	**50%**	**45%**	**41%**
Current Anxiety disorder §%	**26%**	**28%**	**19%**
ODD, CD, or ADHD §%	**25%**	**26%**	**20%**
Baseline Scores LSM [SEM]			
HAM-D	**18.93 [0.44]**	**18.12 [0.43]**	**18.98 [0.44]**
K-SADS-L	**28.31 [9.52]**	**27.53 [0.51]**	**28.31 [0.52]**
Autonomous Function	**93.35 [3.10]**	**96.96 [3.10]**	**94.16 [3.17]**
Self Perception Profile	**63.97 [2.22]**	**63.54 [2.19]**	**63.35 [2.28]**
Sickness Impact Profile	**32.35 [1.23]**	**30.82 [1.23]**	**32.88 [1.27]**

§ from the Screening K-SADS-L Structured Interview

Figure 1 summarises the allocations and discontinuations among the three treatment groups during the acute study period.

Insert Figure 1 here.

The flow chart covers the ITT population for the acute phase and the efficacy analysis. The paroxetine group was titrated to a dose of 20mg/day by week 4, with 55% moving to a higher dose (mean 28.0 mg/day, SD 8.4 mg) by week 8. The imipramine group was titrated to 200 mg/day by week 4, with 40% going higher (mean 205.8 mg/day, SD 63.9 mg) by week 8. 28 patients reached the highest permissible dose of 40 mg of paroxetine, and 20 patients were titrated to the maximum 300 mg of imipramine.

Efficacy

There were no discrepancies between any of our analyses and those contained in the CSR. Figure 2 illustrates the longitudinal values for the two primary efficacy variables: mean change from baseline in the HAM-D score; and the percent responding, defined as a decrease in HAM-D score by 50% or more from baseline or a final HAM-D score of 8 or below. The difference between paroxetine and placebo fell short of the pre-specified level of clinical significance (4 points) and neither primary outcome achieved statistical significance at any measured interval during the acute phase.

Insert Figure 2 here.

The analysis included both OC and LOCF datasets. The results at week 8 are shown in Table 3.

Table 3. OC and LOCF datasets for primary and secondary outcomes

Primary Efficacy Variables [8 Weeks]

		Paroxetine		Imipramine		Placebo		*p*
	Data	**LSMean [SEM]**	**n**	**LSMean [SEM]**	**n**	**LSMean [SEM]**	**n**	**ANOVA**
HAM-D Change	**OC**	**-12.18 [0.88]**	**67**	**-10.59 [0.97]**	**56**	**-10.51 [0.88]**	**66**	**0.255**
	LOCF	**-10.73 [0.81]**	**90**	**-8.95 [0.81]**	**94**	**-9.08 [0.83]**	**87**	**0.204**
		criteria met	**[+/-]**	**criteria met**	**[+/-]**	**criteria met**	**[+/-]**	X^2
HAM-D Response ≥50% drop or ≤8	**OC**	**80.60%**	**54/13**	**73.20%**	**41/15**	**65.2%**	**43/23**	**0.131**
	LOCF	**66.7%**	**60/30**	**58.5%**	**55/39**	**55.2%**	**48/39**	**0.269**

Secondary Efficacy Variables [8 Weeks]

		Paroxetine		Imipramine		Placebo		*p*
		LSMean [SEM]	**n**	**LSMean [SEM]**	**n**	**LSMean [SEM]**	**n**	**ANOVA**
K-SADS-L Change	**OC**	**-12.05 [0.91]**	**67**	**-10.70 [1.00]**	**56**	**-10.71 [0.92**	**65**	**0.459**
	LOCF	**-11.43 [0.84]**	**83**	**-9.47 [0.82]**	**88**	**-9.39 [0.83]**	**85**	**0.131**
CGI Mean Score	**OC**	**1.89 [0.15]**	**68**	**2.16 [0.17]**	**56**	**2.36 [0.16]**	**66**	**0.086**
	LOCF	**2.36 [0.16]**	**90**	**2.69[0.15]**	**94**	**2.72[0.16]**	**87**	**0.155**
Autonomous	**OC**	**14.35 [2.83]**	**58**	**13.34 [3.04]**	**52**	**9.29 [2.81]**	**60**	**0.325**

Function Check List Change	**LOCF**	**14.68 [2.80]**	**60**	**11.55 [2.92]**	**57**	**9.27 [2.76]**	**62**	**0.367**
Self Perception Profile	**OC**	**12.89 [2.31]**	**60**	**13.24 [2.46]**	**55**	**12.68 [2.30]**	**60**	**0.875**
Change	**LOCF**	**13.22 [2.33]**	**61**	**13.06 [2.41]**	**60**	**11.38 [2.27]**	**63**	**0.877**
Sickness Impact	**OC**	**-11.18 [1.57]**	**62**	**-13.51 [1.70]**	**55**	**-10.63 [1.57]**	**62**	**0.244**
Profile Change	**LOCF**	**-11.36 [1.55]**	**63**	**-12.98 [1.62]**	**60**	**-9.87 [1.51]**	**65**	**0.233**

LSMean - Least Square Means adjusted over the site covariate. (Using arithmetic means did not alter the findings.)
SEM – Standard Error of the Mean.
ANOVA – All Treatment [Omnibus] Analysis of Variance with Treatment and Site Effects in the model
X^2 - Logistical Regression with Treatment and Site Effects in the model
OC – Observed Cases
LOCF – Last Observation Carried Forward
Note - All p values uncorrected for multiple variable sampling

There was no statistical significance (considered at $p<0.05$) or clinical significance demonstrated for any of the pre-specified primary or secondary efficacy variables in either the OC or LOCF datasets, so pairwise analysis was considered unjustified.

Although the protocol listed predictors of response among the secondary efficacy variables, the absence of statistically or clinically significant differences among the three arms rendered this analysis void.

The protocol also listed the relapse rate in the continuation phase for responders as a secondary outcome variable. Our calculation differed from the CSR calculation because we included those whose HAM-D scores rose above the 'response' range and those who intentionally overdosed. In the continuation phase, the dropout rates were too high in all groups for any precise interpretation: paroxetine 33/51 [65%]; imipramine 25/39 [64%]; and placebo 21/34 [62%]. The recorded relapses were paroxetine 25/51 [49%]; imipramine 16/39 [41%]; and placebo 12/34 [35%]. Although the relapse rate was lower in the placebo group, the results were not statistically significant, p=0.440 [Chi-square 2x3].

Harms

Review of Clinical Records Forms

The review of 34% of CRFs produced the data shown in Table 4.

Table 4. AEs found in CRFs vs. AEs listed in Appendix D

	Paroxetine (n=31)	Imipramine* (n=40)	Placebo (n=22)
AEs found in CRFs	159	257	77

AEs found in Appendix D	136	240	67
% underestimate in relying only on Appendix D	14%	7%	13%

*In considering adverse effects from imipramine, it should be noted that doses (mean 205.8 mg) were high for adolescents. In the six comparator studies submitted by SKB as part of their 1991 Approval NDA for paroxetine in adults, the mean imipramine dose overall was 140mg, with a mean endpoint dose of 170mg.[23]

Recoding and Representation of Adverse Event Data

Table 5 presents AEs found in this study according to System-Organ-Class (SOC) recoded from the CSR Appendix D (RIAT MedDRA recoded), and additional AEs found in our reanalysis of 93 CRFs. A full listing of AEs can be found in table iii in Appendix 2 to this paper.

Table 5. Adverse events in CSR and 93 CRFs

	Paroxetine N=93		**Imipramine N=95**		**Placebo N=87**	
Type of Adverse Event	**CSR** RIAT MedDRA recoded	additional AEs found in 93 CRFs	**CSR** RIAT MedDRA recoded	additional AEs found in 93 CRFs	**CSR** RIAT MedDRA recoded	additional AEs found in 93 CRFs
Cardiovascular SOC*	45	0	131	5	32	0
Gastrointestinal SOC	112	4	147	4	79	2
Psychiatric SOC*	101	12	63	1	24	4
Respiratory SOC	42	0	22	1	39	1
All other SOCs	179	7	189	6	156	3
TOTAL	**479**	**23**	**552**	**17**	**330**	**10**

* In the Keller et al paper the AEs 'dizziness' and 'headache' were grouped with psychiatric AEs under the heading 'Nervous System'. In the CSR recoding and CRF review these AEs have been reported under 'Cardiovascular SOC' for dizziness and 'Other/General SOC' for headaches. See also Appendix 2, table iii

Behavioural adverse events are further broken down in Table 6.

Table 6. Behavioural adverse events (acute phase plus taper)

Psychiatric disorders	**Paroxetine N=93**	**Imipramine N=95**	**Placebo N=87**

	CSR RIAT MedDRA recoded	additional AEs found in 93 CRFs	**CSR** RIAT MedDRA recoded	additional AEs found in 93 CRFs	**CSR** RIAT MedDRA recoded	additional AEs found in 93 CRFs
Abnormal dreams	3	0	5	0	2	0
Depression worsening	5	2	3	0	2	1
Aggression/ anger	7	1	3	0	0	0
Agitation	0	1	1	0	0	0
Akathisia	18	0	12	0	8	0
Anxiety	2	0	0	0	1	1
Depersonalisation	0	0	1	0	1	0
Disinhibition	4	0	1	0	2	0
Hallucinations	1	0	1	0	0	0
Paranoia	1	0	0	0	0	0
Psychosis	1	1	0	0	0	0
Suicidal ideation	4	2*	3	0	1	1*
Suicide attempt	9	1*	3	1	0	0
Total AEs	**55**	**8**	**33**	**1**	**17**	**3**
Total patients	**35**		**23**		**12**	

* For the paroxetine group the total suicidal ideation/suicide attempt AEs were 16 from a total of 10 patients. For the placebo group the 2 suicidal ideation AEs were from 2 patients.

Severity Ratings

The CSR reported 11 serious AEs (defined as events that 'resulted in hospitalization, was associated with suicidal gestures, or was described by the treating physician as serious') in the paroxetine group, five in the imipramine group, and two in the placebo group. Designating an AE as serious hinged on the judgement of the clinical investigator. We are therefore not able to make comparable judgements of seriousness, but there are two other methods to approach the issue of severity of AEs. One is to look at those rated as severe rather than moderate or mild at the time of the event. The second is to look at rates of discontinuation due to AEs. Table 7 presents the data rated as severe by the original investigator. In this table, the events are only from the CSR, because new events detected in the review of 93 CRFs do not include severity ratings.

Table 7. Adverse events rated as 'severe' (acute phase plus taper)

System Organ Class	Paroxetine N=93	Imipramine	Placebo

(MedDRA)			N=95		N=87	
	Appendix D RIAT MedDRA recoded	**Severe AEs reported**	**Appendix D** RIAT MedDRA recoded	**Severe AEs reported**	**Appendix D** RIAT MedDRA recoded	**Severe AEs reported**
Cardiovascular disorders	45	1 (2.2%)	131	4 (3.1%)	32	0
Gastrointestinal	112	25 (22.3%)	147	20 (13.6%)	79	4 (5.1%)
Psychiatric disorders	101	32 (31.7%)	63	4 (6.3%)	24	5 (20.8%)
Respiratory & Thoracic disorders	42	2 (4.8%)	22	1 (4.5%)	39	4 (10.3%)
All other SOCs	179	10 (5.8%)	189	21 (11.2%)	156	12 (7.7%)
Total AEs	**479**	**70 (14.6%)**	**552**	**50 (9.1%)**	**330**	**25 (7.6%)**

Note the high number and proportion of severe psychiatric events in the paroxetine group. In contrast, few of the many cardiovascular events in the imipramine group were rated as severe.

Discontinuations

Table 8 presents the data on rates of discontinuation due to AEs and other causes. Note that we examined all discontinuation CRFs.

Table 8. Reasons for withdrawal during acute phase and taper

Reason for withdrawal		Paroxetine (n=93)*		Imipramine (n=95)		Placebo (n=87)	
		Appendix G	Appendix H	Appendix G	Appendix H	Appendix G	Appendix H
Adverse Event	Aggression	1	0	0	0	0	0
	Mania	1	2	0	0	0	0
	Overdose	1	1	0	0	0	0
	Depression worsening	0	1	0	0	0	1

	Agitation	0	1	0	0	0	0
	Suicidality	0	5*	0	2	0	1
	Hallucinations	0	0	0	1	0	0
	Conduct disorder	1	1	0	0	0	0
	Hospitalisation/surgery	1	0	1	1	0	0
	Fatigue	0	0	1	1	0	0
	Sedation	0	1	0	1	0	0
	Nausea/vomiting	0	1	2	5	0	1
	Rash/acne	0	0	2	3	1	1
	Cardiac	0	1	9	15	3	2
	Accidental injury	0	0	1	0	0	0
	Urinary	0	0	1	1	0	0
	Pregnancy	0	0	1	1	0	0
	Intercurrent illness**	6	0	12	0	2	0
	Total AE dropouts - n (%)	11 (11.8%)	14 (15.0%)	30 (31.5%)	31 (32.6%)	6 (6.9%)	6 (6.9%)
Protocol violation***	Non compliance with med	3	1	4	4	6	4
	By investigator	0	0	0	0	0	4
	Recreational drug use	0	0	1	1	1	1
	Total	3 (3.2%)	1 (1.1%)	5 (5.3%)	5 (5.3%)	7 (8.0%)	9 (10.3%)
Lost to Follow-up		5 (5.4%)	4 (4.3%)	1 (1.1%)	1 (1.1%)	1 (1.1%)	1 (1.1%)
Lack of efficacy		3 (3.2%)	3 (3.2%)	1 (1.1%)	0 (0%)	6 (6.9%)	4 (4.6%)
Withdrawn consent		4 (4.3%)	5 (5.4%)	1 (1.1%)	1 (1.1%)	1 (1.1%)	1 (1.1%)

Total dropout rate - n (%)	26 (28%)	27 (29%)	38 (40%)	38 (40%)	21 (24%)	21 (24%)

*Patient **329.002.00058** was found to have stopped meds 3 days prior to attempting suicide. Originally this had been classed as a 'continuation phase' drop out, but has now been moved to '30 day discontinuation' period. Reason for withdrawal was originally 'AE including intercurrent illness' but was changed to 'suicide attempt'.
**We replaced the term 'Adverse Events: Intercurrent Illness' with more specific AE terms.
***Four patients enrolled in the study violated the inclusion criterion. Two had cardiovascular problems, one had a C-GAS score greater than 60, and one was 'extremely' suicidal at screening. All four were randomised to placebo. It was unclear how to categorize their reasons for discontinuation; we chose 'protocol violations'.

All changes of coding for discontinuation are laid out in our Appendix 2 (Table x).

In a study that has a continuation phase, the assessment of AEs throws up a methodological difficulty not yet addressed by groups such as CONSORT. If a study only has an acute phase, then all AEs are counted for all patients on treatment as well as in any taper phase, and often for a 30-day follow-up period. When a study has a continuation phase, the taper and 30-day follow-up periods are displaced. To ensure comparable analysis of all participants, we have tallied the AEs across the acute phase and both taper and follow-up phases whether displaced or not. We have not been able to ascertain what SKB did in this regard.

Taking this approach in Study 329 revealed a conundrum. In addition to the 86 dropouts from the acute phase noted by SKB, there were 65 dropouts after week 8 ratings were completed. SKB regarded these patients as participants in the continuation phase, although none of them took a continuation phase pill or had a continuation phase rating. The coding for discontinuation was particularly ambiguous for this group.

The majority of patients stopped at this point were designated by SKB as lack of efficacy (see Table 9). Investigators in four centres reported lack of efficacy as a reason for stopping six placebo patients even though the HAM-D score was in the responder range and as low as 2 or 3 points in some instances.

In some cases there were clear protocol violations or factors such as the unavailability of further medication (placebo in particular). We have recategorized the lack of efficacy dropouts based on factors such as AEs and HAM-D scores.

Our analysis of reasons for withdrawal at the end of the acute phase is shown in table 9.

Table 9. Reasons for withdrawal from Study 329 – patients discontinued at the end of the Acute Phase (n=65)

Reason for withdrawal	**Paroxetine group (acute completers n=67)**	**Imipramine group (acute completers n= 56)**	**Placebo group (acute completers n=66**

		SKB/GSK coded, **App G**	**RIAT proposed***	SKB/GSK coded, **App G**	**RIAT proposed***	SKB/GSK coded, **App G**	**RIAT proposed***
Adverse event	Aggression/paranoia	1	1	0	0	0	0
	Mania	0	1	0	0	0	0
	Overdose	1	0	0	0	0	0
	Depression worsening	0	1	0	0	0	0
	Homicidality	0	0	1	1	0	0
	Suicidality	0	1	0	0	0	0
	Rash	1	1	0	0	0	0
	Cardiac	0	0	1	2	0	0
	Dry mouth	0	0	0	1	0	0
	TOTAL AE drop outs **N (%)**	**3**	**5**	**2**	**4**	**0**	**0**
Protocol violation	Non compliance with study meds	1	1	2	2	0	0
	Recreational drug use	0	0	0	0	1	1
	PV by Investigator	0	1	0	2	0	3
	TOTAL PV drop outs **N (%)**	**1**	**2**	**2**	**4**	**1**	**4**
Lost to follow Up		**0**	**2**	**0**	**0**	**0**	**0**
Lack of efficacy		**9**	**5**	**12**	**8**	**23**	**17**
Withdrawn consent		**1**	**1**	**0**	**0**	**4**	**5**
Other	Misc (HAM-D responder)	0	1	0	1	0	6
	General surgery	1	0	0	0	0	0
	No study meds available	1	0	0	0	3	0
	ADHD symptoms	0	0	1	0	0	0
	Moved out of state	0	0	0	0	1	0

	TOTAL 'other' drop outs **N (%)**	**2**	**1**	**1**	**1**	**4**	**6**
TOTAL DISCONTINUED AT WEEK 8		**16**	**16**	**17**	**17**	**32**	**32**

*Following a review of the codes given for reasons for withdrawal from the study that were found in the CSR (Appendix G), along with a review of patient narratives and CRFs where applicable, we proposed changes to these reasons for withdrawal in a proportion of those discontinued.

Withdrawal Effects

The protocol for Study 329 called for a taper phase for all subjects and in addition a 30-day follow up period for all subjects who discontinued because of adverse events. The data in the CSR Appendix D make it possible to identify adverse events happening in the taper and follow-up periods.

The data are presented in Table 10.

Table 10. Adverse events from taper phase

System Organ Class (MedDRA)	**Paroxetine N=19**		**Imipramine N=32**		**Placebo N=9**	
	AEs reported (RIAT MedDRA recoded**)**	**AEs reported as severe**	**AEs reported (**RIAT MedDRA recoded**)**	**AEs reported as severe**	**AEs reported (**RIAT MedDRA recoded**)**	**AEs reported as severe**
Cardiovascular disorders	4	0	7	0	0	0
Gastrointestinal disorders	9	4	18	4	4	0
Psychiatric disorders	15	7	2	0	1	1
Respiratory & thoracic disorders	3	0	1	0	0	0
All other SOCs	16	1	20	3	5	0
Total AEs	**47**	**12**	**48**	**9**	**10**	**1**

The Effect of Other Medications

In Table 11 we present data on the effects of other medications on the AEs recorded. It is clear that those taking other medications had more AEs than those who were not. This effect is

slightly more marked in the placebo group, and as such works to the apparent benefit of the active drug treatments in minimizing any excess of effects over placebo.

Table 11. Use of other medications in the month prior to enrolment, and incidence of AEs

	Paroxetine (n=93)		Imipramine (n=95)		Placebo (n=87)	
	Other medications	No other medications	Other medications	No other medications	Other medications	No other medications
% patients	26% (n=24)	74% (n=69)	33% (n=31)	67% (n=64)	30% (n=26)	70% (n=61)
Psychiatric AEs subgroup* (acute + taper)	15	38	13	21	6	11
Total AEs (acute + taper)	155	298	215	325	137	190

* PSYCH AEs included in this subgroup include: Abnormal dreams, aggravated depression, agitation, akathisia, anxiety, depersonalisation, disinhibition, hallucinations, paranoia, psychosis, suicidal ideation/gesture/attempt.

Discussion

We have reported Study 329 according to the original protocol and analysed the efficacy data accordingly. Appendix 1 shows the sources of information used in preparing this paper, which should aid other researchers who wish to access the data, either to check our analysis or to interrogate it in other ways. We draw minimal conclusions regarding efficacy and harms, inviting others to offer their own analysis.

The RIAT approach revealed different outcomes from those reported in the CSR and Keller et al. Re-examination of the data, including a review of 34% of the cases, revealed no significant discrepancies in the primary efficacy data. The marked difference in the reporting of efficacy outcomes was predominantly a product of our analysis keeping faith with the protocol methodology and its designation of primary and secondary outcome variables.

The authors/sponsors departed from their study protocol in the CSR itself by performing pairwise comparisons of two of the three groups when the omnibus ANOVA showed no significance in either the continuous or dichotomous variables. They also reported four other variables as significant that had been unmentioned in the protocol or its amendments, without any acknowledgment that these measures were introduced post hoc. This contravened provision II of Appendix B Administrative Matters, according to which any changes to the study protocol were required to be filed as amendments/modifications.

With regard to AEs, there were large and clinically meaningful differences between the data as analysed by us and those reported in Keller et al. These differences arise both from inadequate entry of data from CRFs to summary data sheets in the CSR, and the analysis and reporting of these data sheets in Keller et al. Keller et al reported 265 adverse events with paroxetine, while we identified 479 from our analysis of the CSR, and found a further 23 that had been missed from the 93 CRFs that we reviewed. For all AEs combined, Keller et al. reported a paroxetine burden of AEs 1.25 times that of the placebo burden, compared with 1.5 times in the CSR figures.

One reason why the Keller et al. figures are lower than ours is because Keller et al. only presented data for AEs reported for 5% of patients or more. The CSR and CRF figures also differ substantially from other figures quoted in Keller et al, because we did not code 'dizziness' and 'headache' under Nervous System, since the former is more likely to be attributable to 'cardiovascular' while headaches most commonly stem from muscles and blood vessels to the scalp.

In Keller et al, the paroxetine rate of psychiatric AEs (Table 12) was 1.8 times the placebo rate, while in the CSR figures it is 4 times, making the differences between placebo and paroxetine more salient in the primary datasets than in Keller et al. There was also a major difference between the frequency of suicidal thinking and events reported by Keller et al, and the frequency documented in the CSR. Our CRF review adds even more cases.

Table 12. Comparison of Psychiatric SOC and suicidality using different safety methodologies

	Keller et al.		RIAT MedDRA recoded		additional AEs found in 93 CRFs	
	Paroxetine	Placebo	Paroxetine	Placebo	Paroxetine	Placebo
Psychiatric SOC	115	65	101	24	12	4
Suicidal ideation/gesture	≤5*	≤2*	4	1	2	1
Suicide attempt	0	0	9	0	1	0
Total suicidality	≤5*	≤2*	13	1	3	1

* Classified under 'emotional liability (e.g., suicidal ideation/gestures)'

Our finding is consistent with other findings, including a recent study that examined 142 studies of six psychotropic drugs for which journal articles and clinical trial summaries were both available.[24, 25] Most deaths (94/151, 62%) and suicides (8/15, 53%) cited in trial summaries were not reported in journal articles. Only one of nine suicides in olanzapine trials was reported in published papers.

With regard to dropouts, Keller et al. stated that 69% of patients completed the acute phase. It would be wrong to assume that this meant that 69% continued. In fact only 45% went on to the continuation phase.

Our reanalysis of study 329 revealed significant variations in the way AEs can be reported, demonstrating several ways in which the analysis and presentation of safety data can influence the apparent safety of a drug (see Box 3).

Box 3. Potential confounders of accurate reporting of harms

1. Use of an idiosyncratic coding system

The term 'emotional lability', as used in SKB's ADECS, masks discrepancies in suicidal behaviour between paroxetine and placebo.

2. Failure to transcribe all AEs from the clinical record to the side effect database

Our review of CRFs disclosed significant under-recording of AEs.

3. Filtering data on AEs through statistical techniques

For instance, Keller et al. (and GSK in subsequent correspondence) ignored unfavourable harms data on the grounds that the difference between paroxetine and placebo was not statistically significant. In our opinion, statistically significant or not, all relevant primary and secondary outcomes, and harms outcomes, should be explicitly reported. Testing for statistical significance is most appropriately undertaken for the primary outcome measures. We have not undertaken statistical tests for harms, since we know of no valid way of interpreting them. To get away from a dichotomous (statistically significant/non significant) presentation of evidence, we opted to present all original and recoded evidence to allow readers their own interpretation. The data presented in Appendix 2 and related worksheets lodged at www.xxx will, however, readily permit other approaches to data analysis for those interested, and we welcome other analyses.

4. Restriction of reporting to events that occurred above a given frequency in any one group

In the Keller et al. paper, reporting only AEs that occurred in more than 5% of patients obscured the harms burden. In contrast, we report all AEs that have been recorded. These are available in Table v in Appendix 2 that accompanies this paper.

5. Coding an event under different headings for different patients (dilution)

The effect of reporting only AEs that have a frequency of more than 5% is compounded when, for instance, agitation may be coded under agitation, anxiety, nervousness, hyperkinesis and emotional lability; thus, a problem occurring at a rate of >10% could vanish by being coded under different subheadings such that none of these reach a threshold rate of 5%.

Aside from making all the data available so that others can scrutinize it, one way to compensate for this possibility is to present all the data in broader SOC groups. MedDRA offers the following higher levels: psychiatric; cardiovascular; gastrointestinal; respiratory; and other. In Appendix 2, table v, the data coded here under 'Other' is broken down under the additional MedDRA SOC headings - general, nervous system, metabolic, musculo-skeletal, endocrine, eye, renal,

'immune system, blood and lymphatic disorders, skin, infectious, reproductive system, ear, injuries, surgical, and pregnancy.

6. Grouping of AEs

Even when presented in broader system groups, grouping common and benign symptoms with more important ones can mask safety issues. For example, in the Keller paper, common AEs such as dizziness and headaches are grouped with psychiatric AEs in the 'nervous system' SOC heading. Since these AEs are frequent across treatment arms, this grouping has the effect of diluting the difference in psychiatric side effects between paroxetine, imipramine and placebo.

We have reported dizziness under 'cardiovascular' events and headache under 'other'. There may be better categorisations; our grouping is provisional rather than strategic. In Appendix 2, table v, we have listed all events coded under each SOC heading and we invite others to further explore these issues, including alternative higher level categorisation of these AEs.

7. Rating Severity

In addition to coding AEs, investigators rate them for severity. If no attempt is made to take severity into account, readers may get the impression that there was an equal AE burden in each arm, when in fact all events in one arm might be severe and enduring while those in the other might be mild and transient.

One way to manage this is to look specifically at those patients who drop out of the study because of AEs. Another method is to select those AEs coded as severe for each drug group while omitting those coded as mild or moderate. We used both approaches.

8. Relatedness coding

Judgements by investigators as to whether an AE is related to the drug can lead to discounting the importance of an effect. We have included these judgements in the worksheets lodged at www.xxx [TBA] but have not analysed them, because it became clear that the blind had been broken in several cases before relatedness was adjudicated by the original investigators, and because some judgements were implausible. For instance, it is documented in the CSR (p 279) that an investigator, knowing the patient was on placebo, declared that a suicidal event was 'definitely related to treatment', on the grounds that 'the worsening of depression and suicidal thought were life threatening and definitely related to study medication [known to be placebo] in that there was a lack of effect'. Notably, of the 11 patients with serious AEs on paroxetine (compared to two on placebo) reported in the Keller paper, only one 'was considered by the treating investigator to be related to paroxetine treatment', thus dismissing the clinically significant difference between the paroxetine and placebo groups for serious AEs.

9. Masking effects of concomitant medication

In almost all trials, patients will be on concomitant medications. The AEs from these other medications will tend to obscure differences between active drug treatment and placebo. This may be a very significant factor in trials of treatments such as statins, where patients are often on multiple medications.

Accordingly we also compared the list of AEs in those on concomitant medication versus those not on other medication. There are other medications instituted in the course of the study that we have not analysed, but the data are available in our Appendix 2 and worksheets lodged at www.xxx, and in Appendix B from the CSR. There are a number of other angles in the submitted data that could be further explored, such as the effects of withdrawal of concomitant medication on AE profiles as the spreadsheets submitted offer the day of onset of AEs and the dates of starting or stopping any concomitant medication. Another option to explore is the possibility of any prescribing cascades triggered by AEs related to study medication.

10 The Effects of Medication Withdrawal

The protocol included a taper phase lasting 7-17 days that investigators were encouraged to adhere to even in patients who were discontinued because of adverse events. The original paper did not analyse these data separately. We have done. They reveal evidence consistent with dependence on and withdrawal from paroxetine.

This RIAT exercise proved to be demanding of resources. We have logged (www.xxx [TBA]) over 130,000 words of email correspondence amongst the team over a year. Gaining access to the CRFs required extensive correspondence with GSK.[10] Although GSK ultimately provided CRFs, the mode of access was excessively time-consuming. It required of the order of one thousand hours to examine only a third of the CRFs. Less restricted access to the CRFs would have significantly reduced the burden.

Our analysis indicates that while CSRs are useful, and in this case all that was needed to reanalyse efficacy, analysis of adverse events requires access to individual patient level data in the form of CRFs.

Since we have been breaking new ground, we do not always have precedents to call on in analysis and reporting, and we are open to future collaborations to do things differently. We invite readers to contact us for clarification of any ambiguities through a public Q&A forum at www.xxx.com [TBA], where we will provide an initial response within two working days to any queries about our data or analysis, with further follow-up as required.

Conclusion

Study 329 showed no advantage of paroxetine or imipramine over placebo in adolescent depressive symptomatology on any of the pre-specified parameters. There were clinically significant increases in AEs in the paroxetine and imipramine arms, including serious, severe, and suicide related AEs.

As with most scientific papers, Keller et al. conveys an impression that 'the data has spoken'. This authoritative stance is only possible in the absence of access to the data. When the data become accessible to others, it becomes clear that scientific authorship is provisional rather than authoritative.

Box 4. Strengths and limitations of this study

Study 329 was a randomised controlled trial with a reasonable sample size.

The RIAT analysis included a review of 34% of CRFs conducted by two investigators, using MedDRA (by far the most commonly used coding system today) to check AE data. The analysis generated a useful taxonomy of potential confounders of accurate reporting of AEs.

This study has significant limitations. There was evidence of protocol violations, including some cases of blind-breaking. Some AEs were miscoded, raising the possibility that some other data might be unreliable. Time and resources prevented access to all CRFs because of the difficulties in using the portal for accessing the study data and because significant data were missing.

The trial duration was only eight weeks. Participants had relatively chronic depression (mean duration more than one year), which would limit the generalizability of the results, particularly to primary care, because many cases of adolescent depression have shorter durations.[26] Generalizability to primary care would also be limited by the fact that participants were recruited via tertiary settings.

Trial Registration: Registration number and name of trial register: SmithKline Beecham study 29060/329.

Trial Protocol: SmithKline Beecham study 29060/329, Final Clinical Report (Acute Phase), Appendix A, Protocol, from p. 531.[12]

Trial Funding: SmithKline Beecham study.

Funding of the RIAT re-analysis: No funding received.

Data Analysis Protocol for RIAT re-analysis: Submitted to GSK on 28 October 2013. Approved by GSK on 4 December 2013.

We thank Tom Jefferson and Leemon McHenry for comments on various drafts.

Appendices/Supplementary material

1. RIATAR audit record, showing sources of data
2. Adverse event appendices

References

1. Doshi P, Dickersin K, Healy D, Vedula SS, & Jefferson T. Restoring invisible and abandoned trials: a call for people to publish the findings. BMJ. 2013;346:f2865–f2865.

2. Keller MB, Ryan ND, Strober M, Klein RG, Kutcher SP, Birmaher B, et al. Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. J Am Acad Child Adolesc Psychiatry. 2001;40(7):762–772.

3. McHenry L, Jureidini J. Industry-sponsored ghostwriting in clinical trial reporting: A case study. Account Res. 2008;15:152–167.

4. Jureidini J, McHenry L, Mansfield P. Clinical trials and drug promotion: Selective reporting of study 329. Int J Risk Saf Med. 2008;20:73–81.

5. Jureidini J, McHenry L. Conflicted medical journals and the failure of trust. Account Res. 2011;18:45-54.

6. Kraus JE., letter to Dr. Jon Jureidini, [Internet]. 2013 [Accessed 19 April 2014]. Available from: http://www.bmj.com/content/suppl/2013/11/12/bmj.f6754.DC1/doshinov16.ww1_default.pdf

7. Treasure T, Monson K, Fiorentino F, Russell C. The CEA Second-Look Trial: a randomised controlled trial of carcinoembryonic antigen prompted reoperation for recurrent colorectal cancer. BMJ Open. 2014 May 13;4(5):e004385.

8. SmithKline Beecham, A multi-center, double-blind, placebo controlled study of paroxetine and imipramine in adolescents with unipolar major depression –acute phase, Final clinical report. [Internet]. [Accessed 20 August 2014]. Available from: http://www.gsk.com/media/389566/depression_329_full.pdf

9. Healthy Skepticism International News, Paxil Study 329: Paroxetine vs Imipramine vs Placebo in Adolescents, [Internet]. 2010 [Accessed 19 April 2014]. Available from: http://www.healthyskepticism.org/global/news/int/hsin2010-01

10. Secure Access Management. [Internet]. [Accessed 29 July 2014]. Available from: https://www.ondemand.sas.com/sam/?sid=1393012805&rid=DqqHX0rCqAWZ7TJlCIPijRTscQ.

11. Correspondence between Jureidini and GSK, Rapid Responses to Putting GlaxoSmithKline to the test over paroxetine. BMJ 2013;347:f6754 [Internet]. [Accessed 29 July 2014]. Available from: http://www.bmj.com/content/347/bmj.f6754/rapid-responses

12. SmithKline Beecham, A multi-center, double-blind, placebo controlled study of paroxetine and imipramine in adolescents with unipolar major depression 1993/amended 1996. [Internet]. [Accessed 20 August 2014]. Available from: http://www.gsk.com/media/360485/329-AppA.PDF

13. Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition–Revised (DSM-III-R). Washington, DC: American Psychiatric Association; 1987.

14. Fawcett, J., Epstein, P., Fiester, S. J., Elkin, I., & Autry, J. H. Clinical management--imipramine/placebo administration manual. NIMH Treatment of Depression Collaborative Research Program. Psychopharmacol Bull. 1987;23(2):309–324.

15. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol. 1967; 6:278-296.

16. Sigafoos AD, Feinstein CB, Damond M, Reiss D The measurement of behavioral autonomy in adolescence: the Autonomous Functioning Checklist. Adolesc Psychiatry. 1988;15:432-62.

17. SKB. Draft Minutes: 4/22/97 Teleconference. Paroxetine Study 329 Efficacy Analysis. [Internet]. [Accessed 29 July 2014]. Available from: http://www.healthyskepticism.org/files/docs/gsk/paroxetine/study329/970422teleconference.pdf

18. GlaxoSmithKline, Paroxetine - paediatric and adolescent patients. [Internet]. [Accessed 20 August 2014]. Available from: http://www.gsk.com/en-gb/media/resource-centre/paroxetine/paroxetine-paediatric-and-adolescent-patients/

19. Jureidini JN, Nardo JM. Inadequacy of remote desktop interface for independent reanalysis of data from drug trials. BMJ. 2014 Jul 9;349:g4353.

20. Fitzgerald K, Healy D. Dystonias and dyskinesias of the jaw associated with the use of SSRIs. Human Psychopharmacology. 1995;10:215-20.

21. Kline RB. Beyond Significance Testing. Statistics Reform in the Behavioral Sciences. Second Edition. Washington, DC: American Psychological Association, 2013

22. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. 2013 [Accessed 16 August 2014].

Available from: http://www.R-project.org/

23. Brecher M. Review and evaluation of clinical data. Original NDA 20 – 031. Paroxetine (Aropax). Efficacy review. SmithKline Beecham pharmaceuticals; 1991 Jun. 3.

24. Hughes S, Cohen, D, Jaggi R. Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: a cross-sectional study. BMJ Open. 2014;4:e005535.

25. Maund E, Tendal B, Hróbjartsson A, Jørgensen, KJ, Lundh A, Schroll J, Gøtzsche PC. Benefits and harms in clinical trials of duloxetine for treatment of major depressive disorder: comparison of clinical study reports, trial registries, and publications. BMJ. 2014;348:g3510

26. Lewinsohn PM, Clarke GN, Seeley JR, Rohde P. Major depression in community adolescents: age at onset, episode duration, and time to recurrence. J Am Acad Child Adolesc Psychiatry. 1994 Jul-Aug;33(6):809-18.